Mustafa Temiz

Ph.D. Thesis

AGU 2024

DESIGN AND DEVELOPMENT OF MACHINE LEARNING MODELS FOR DISEASE PREDICTION AND BIOMARKERS DETECTION

Ph.D. THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF ABDULLAH GUL UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Ph. D.

By

Mustafa Temiz

June 2024

DESIGN AND DEVELOPMENT OF MACHINE LEARNING MODELS FOR DISEASE PREDICTION AND BIOMARKERS DETECTION

Ph.D. THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF ABDULLAH GUL UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Ph. D.

By

Mustafa Temiz

June 2024

SCIENTIFIC ETHICS COMPLIANCE

I hereby declare that all information in this document has been obtained in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name-Surname: Mustafa Temiz
Signature :

REGULATORY COMPLIANCE

Ph.D. thesis titled "Design and Development of Machine Learning Models for Disease Prediction and Biomarkers Detection" has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

Prepared By	Advisor	Co-Advisor
Mustafa Temiz	Assoc. Prof. Burcu Bakır Güngör	Prof. Dr. Malik Yousef

Head of the Electrical and Computer Engineering Program

Asst. Prof. Samet GÜLER

Signature

ACCEPTANCE AND APPROVAL

Ph.D. thesis titled "Design and Development of Machine Learning Models for Disease Prediction and Biomarkers Detection" and prepared by Mustafa Temiz has been accepted by the jury in the Electrical and Computer Engineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

26 /06 / 2024

(Thesis Defense Exam Date)

JURY:

Advisor : Assoc. Prof. Burcu BAKIR GÜNGÖR

Member : Assoc. Prof. Mete ÇELİK

Member : Assoc. Prof. Rifat KURBAN

Member : Asst. Prof. Bilge Kağan DEDETÜRK

Member : Asst. Prof. Gülay YALÇIN ALKAN

APPROVAL:

The acceptance of this Ph.D. thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Engineering & Science, Executive Board dated ….. /….. / …….. and numbered ……..……. .

…….. /…….. / ……..

(Date)

Graduate School Dean
Prof. Dr. İrfan ALAN

ABSTRACT

DESIGN AND DEVELOPMENT OF MACHINE LEARNING MODELS FOR DISEASE PREDICTION AND BIOMARKERS DETECTION

Mustafa Temiz
Ph.D. in Electrical and Computer Engineering
Advisor: Assoc. Prof. Burcu Bakır Güngör
Co-advisor: Prof. Dr. Malik Yousef
June 2024

In medical science, the prediction of diseases and the identification of biomarkers play an important role in the diagnosis and treatment of various health conditions. The recent proliferation of data mining techniques has accelerated the development of disease prediction systems. In particular, machine learning methods are an effective way to analyze medical data and identify patterns to predict the likelihood of the disease development. Machine learning methods also help to identify biomarkers. Recently, the increasing incidence and mortality rates of inflammatory bowel disease, colorectal cancer and type 2 diabetes have drawn researchers' attention to these research areas. The aim of this thesis is to reduce the number of features and improve the prediction performance of machine learning based on complex biological datasets with a large number of disease-related features, as well as to identify potential biomarkers. In this thesis, three different studies are presented. The first study predicts eleven different cancer subgroups using miRNA data and biological domain knowledge and identifies potential biomarkers for these diseases. The second study predicts three different diseases using metagenomic data and biological domain knowledge and identifies potential biomarkers. The third study uses metagenomic data related to colorectal cancer to conduct global and population-based comprehensive experiments with traditional feature selection methods to identify potential biomarkers. This thesis presents a promising avenue for early disease detection, facilitating expedited treatment protocols, improving human survival rates, and potentially alleviating economic burdens within these critical research domains.

Keywords: Disease prediction, Machine Learning, Identify Biomarkers, Feature Selection, Colorectal Cancer, Type 2 Diabetes, Inflammatory Bowel Disease

ÖZET

HASTALIK TAHMİNİ VE BİYOBELİRTEÇLERİN TESPİTİ İÇİN MAKİNE ÖĞRENİM MODELLERİNİN TASARIMI VE GELİŞTİRİLMESİ

Mustafa Temiz
Elektrik ve Bilgisayar Mühendisliği Anabilim Dalı Doktora
Tez Danışmanı: Doç. Dr. Burcu Bakır Güngör
İkinci Tez Danışmanı: Prof. Dr. Malik Yousef
Haziran 2024

Tıp biliminde, hastalıkların tahmini ve biyobelirteçlerin tanımlanması, çeşitli sağlık koşullarının teşhis ve tedavisinde önemli bir rol oynamaktadır. Veri madenciliği tekniklerinin son zamanlarda yaygınlaşması, hastalık tahmin sistemlerinin gelişimini hızlandırmıştır. Özellikle makine öğrenim yöntemleri, tıbbi verilerin analizinde ve hastalığın ortaya çıkma olasılığını tahmin etmeye yönelik kalıpların belirlenmesinde etkili bir yöntemdir. Makine öğrenim yöntemleri, biyobelirteçlerin tanımlanmasına da yardımcı olmaktadır. Son zamanlarda inflamatuar bağırsak hastalığı, kolorektal kanser ve tip 2 diyabet hastalıkları ile karşılaşma sıklığının artması ve artan ölüm oranları araştırmacıların dikkatini bu araştırma alanlarına çekmektedir. Bu tezin amacı, hastalık ile ilişkili karmaşık ve çok sayıda özellik içeren biyolojik veri setlerinden yola çıkarak özelliklerin sayısını azaltmak ve makine öğrenmesi tahmin performansını artırmaktır ve ayrıca potansiyel biyobelirteçleri tanımlamaktır. Bu tezde üç farklı çalışma tanıtılmaktadır. İlk çalışma miRNA verileri ve biyolojik alan bilgisi kullanılarak on bir farklı kanser alt grubu tahmin edilmekte ve bu hastalıklar için olası biyomarkörler belirlenmektedir. İkinci çalışma da metagenomik veriler ve biyolojik alan bilgisi kullanılarak üç farklı hastalık tahmin edilmekte ve olası biyomarkörler belirlenmektedir. Üçüncü çalışma kolorektal kanser ile ilişkili metagenomik verileri kullanarak geleneksel özellik seçim yöntemleri ile küresel ve popülasyonlara bağlı kapsamlı deneyler gerçekleştirilmekte ve olası biyomarkörler belirlenmektedir. Bu tez, erken hastalık tespiti için umut verici bir yol sunmakta, hızlandırılmış tedavi protokollerine olanak tanımakta, insan sağkalım oranlarını artırmakta ve bu kritik araştırma alanlarında potansiyel olarak ekonomik yükleri azaltmaktadır.

Anahtar kelimeler: Hastalık tespiti, Makine Öğrenmesi, Biyomarkör Belirleme, Özellik Seçimi, Kolorektal Kanser, Tip2 Diyabet, İnflamatuar Bağırsak Hastalığı

Acknowledgements

I would like to thank my advisor, Assoc. Prof. Burcu Bakır Güngör, who supported me in every aspect during my doctoral education, guided me at every stage of my thesis work, and never spared her trust and support while carrying out my studies, and whom I see as an idol for my success.

I would like to special thanks to my Co- advisor, Prof. Dr. Malik Yousef, for his patience, faith and trust in me, and valuable ideas and suggestions. It has been a great honor for me to work with him.

I would like to thank Assoc. Prof. Mete Çelik, Assist. Prof. Bekir Hakan Aksebzeci and Assoc. Prof. Rifat Kurban for helping me with their knowledge and taking their valuable time to follow my development and advise me.

Finally, I would like to gratefully thank my family (my father, my mother, and my sister) who have patiently waited for me and supported me at every moment of my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AUC	Area Under the Curve
BLCA	Bladder Urothelial Carcinoma
CMIM	Conditional Mutual Information Maximization
CRC	Colorectal Cancer
FCBF	Fast Correlation Based Filter
FS	Feature Selection
GBDT	Gradient Boosting Decision Tree
G-S-M	Grouping, Scoring, and Modeling
HMDD	Human MicroRNA Disease Database
IBD	Inflammatory Bowel Disease
IG	Information Gain
KNN	k-Nearest Neighbor
LUSC	Lung Squamous Cell Carcinoma
LOOCV	Leave-one-out-cross validation
LR	Logistic Regression
MicroRNA	miRNA
MCCV	Monte Carlo cross-validation
MRMR	Maximum Likelihood and Minimum Redundancy
NGS	Next-Generation Sequencing
OTU	Operational Taxonomic Units
RF	Random Forest
SKB	Select K Best
STAD	Stomach Adenocarcinoma
SVM	Support Vector Machines
TCGA	The Cancer Genome Atlas
THCA	Thyroid Carcinoma
T2D	Type 2 Diabetes
WHO	World Health Organization

To my family

Chapter 1

Introduction

1.1 Prediction of Disease using Machine Learning Methods

Globally, health problems such as diabetes, cancer, and inflammatory bowel disease (IBD) are increasing rapidly in both developed and developing countries. The increasing prevalence of these diseases has been leading to a dramatic rise in human mortality rates. In 2019, almost 4.9 million cases of IBD were reported, highlighting the impact of IBD on global healthcare systems [1]. Type 2 Diabetes (T2D) is responsible for 6.8% of deaths worldwide [2]. In 2022, more than 20 million new cases of cancer were reported worldwide, with almost half of these cases leading to death, accounting for 9.7 million deaths annually. By 2050, the average number of cancer cases is expected to increase by almost 77%, placing a further burden on healthcare systems, individuals and communities [3]. The identification of biomarkers associated with diseases for early diagnosis and the development of early treatment procedures by utilizing these biomarkers have the potential to reduce the number of deaths caused by these diseases and shed light on the treatment procedures and measures to be taken. The development of machine learning-based predictive systems and biomarker identification systems can facilitate the effective diagnosis and treatment of these diseases. These systems can be used to monitor symptoms, predict disease progression, and optimize treatment strategies. In this way, medical staff can manage patients' conditions more effectively and improve treatment outcomes. In this respect, the use of machine learning based prediction systems can play an important role in supporting healthcare systems in the fight against these life-threatening diseases [4].

Machine learning (ML) algorithms represent the process of automatically extracting information from data obtained by statistical methods. These algorithms have proven to

be very effective in a number of application areas such as medicine and biology. Machine learning methods play a crucial role in the field of medical disease diagnosis, and are extensively utilized in bioinformatics [5]. Supervised learning and unsupervised learning are the two basic types of machine learning that are commonly defined. The main difference between these two learning strategies is the presence or absence of a label in the data. In today's bioinformatics research, machine learning algorithms including Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN) are successfully used for various tasks such as classification, clustering, regression or dimensionality reduction [6]. To create an efficient and fast prediction (or classification) model, effective feature selection techniques or biological domain knowledge should be used to reduce the complexity of the feature space. In this thesis, an effective machine learning and feature selection algorithm with superior performance is developed using fewer features by reducing the number of features for complex datasets associated with diseases. This model predicts diseases for unknown data using known disease data, and biological structures associated with these diseases are identified as possible biomarkers.

In the first (miRdisNET) and second study (microBiomeGSM), grouping (clustering) is performed with the help of biological domain knowledge using the G-S-M approach. Thanks to this grouping, the number of features is reduced, the classification performance of the machine learning is evaluated, and possible biomarkers are identified. In the third study (CCPRED), classification is performed using the features obtained by union and the features selected by the feature selection methods are intersected, and possible biomarkers selected by this model are identified. Reducing the number of features primarily reduces the computational cost of the analysis. Additionally, the identification of possible biomarkers by using only relevant features increases the effectiveness of the developed method.

1.2 Feature Selection and Machine Learning

Machine learning and feature selection play a crucial role in various fields, including data analytics, prediction models, and classification tasks. By utilizing machine learning algorithms and carefully selecting relevant features, researchers, and practitioners can improve the accuracy of their models, gain insights from complex datasets, and reduce data gathering and processing costs. High-dimensional biological

datasets often have many features that are redundant or superfluous. In this case, the performance of the classifier can be affected by the presence of repeated and irrelevant features. To solve the problem of redundancy, feature selection (FS) algorithms are often used to determine the appropriate subset of features from high-dimensional datasets. Feature subset selection, or simply feature selection, is the process of selecting features for the target data. By removing noisy and redundant data from high-dimensional datasets, FS improves the clarity of the output. It also reduces the complexity of the classifier and minimizes the problem of overfitting, making the results easier to interpret and understand. The main goal of feature selection is to remove redundant data in order to improve the accuracy of a classifier.

One of the most important preprocessing techniques for machine learning problems is feature selection. However, it can be difficult to choose from the growing range of feature selection techniques. Numerous reviews have looked at the advantages and disadvantages of different feature selection techniques. Research has addressed important issues such as noise in the target class, noise in the input features and a much higher number of samples than features [7]. Feature selection methods are used in a variety of ways in the bioinformatics, such as in the studies by (Remeseiro and Bolon-Canedo, 2019) [8]. In the field of machine learning, feature selection methods are categorized into three main types: filtering, wrapper, and embedded, based on their interaction with data or a learning algorithm. In addition, ensemble and hybrid approaches are proposed in the literature, where different methods are appropriately combined or integrated to identify the most effective subset of features. In addition, recent research has explored integrative methods that utilize existing domain knowledge, particularly in bioinformatics and computational biology, where this knowledge is central to interpreting results and improving performance. In summary, machine learning algorithms combined with feature selection techniques are powerful tools in extracting relevant insights and improving the accuracy of models. Machine learning and feature selection are essential components of data analysis and modeling.

In this thesis, grouping-based feature selection using biological domain knowledge is applied in the first and second studies, while conventional feature selection methods are used in the third study. In this way, the performance of machine learning is investigated in large-scale experiments with meaningful features obtained using feature selection methods, and potential biomarkers are identified.

1.3 Literature Review of Disease Prediction

For miRdisNET, there are review and research studies using machine learning methods with miRNA and disease associations. In recent years, several computational methods, especially the ones using machine learning algorithms, have been proposed for predicting associations between miRNA and disease [9]. Chen et al. proposed a novel computational method called RKNNMDA for predicting related miRNAs for several diseases [10]. For prediction, they use potential miRNA-disease associations by combining with disease similarity networks, miRNA similarity networks, and known disease-miRNA associations. They first used the KNN algorithm and the SVM ranking algorithm to obtain the k-nearest neighbors for both miRNAs and diseases. Secondly, they ranked the k-nearest neighbors according to their similarity scores to the central miRNA/disease. Finally, they obtained a ranking of all miRNA-disease associations with weighted voting. In experiments using the leave-one-out-cross validation (LOOCV) technique, they obtained an AUC of 0.8221 [10]. Yao et al. proposed a structural model for inferring miRNA-disease association using random forest algorithm. Their method called IRFMDA achieved AUC of 0.9363, 0.8728, 0.9398 by using 5-fold cross-validation, local leave-one-out cross-validation and global leave-one-out cross-validation, respectively [11]. Liu et al. presented a method (SMALF) for miRNA-disease association prediction [12]. This method learns latent miRNA and disease features using a stacked autoencoder from the original association matrix between miRNA and disease. Using the XGBoost algorithm and cross-validation technique, they reported performance of 0.95 AUC [12]. Ding et al. utilizes semantic similarity of diseases, functional similarity of miRNAs and the miRNA-disease associations to rank disease-miRNA association pairs. They used the KNN algorithm and the LOOCV technique for classification. Their procedure called IIMCMP reached an AUC of 0.9016 [13]. Zhou et al. proposed a novel model in which they extract features using the Gradient Boosting Decision Tree (GBDT) [14]. For classification, they used the logistic regression (LR) algorithm, and they achieved an AUC of 0.9274 with 5-fold cross-validation [14]. To predict the association of miRNA-disease, Liu et al. presented a computational model called DFELMDA [15]. They created a dataset by combining the disease similarity network, the miRNA similarity network and the verified disease-miRNA associations. They represent this high-dimensional dataset in smaller dimensions by using the Deep Auto-Encoder for each disease-miRNA association. For classification, they used a deep random forest algorithm.

In experiments with 5 and 10-fold cross-validation, the best models obtained an AUC of 0.9552 and 0.9560, respectively [15].

Following the research efforts on the impacts of microRNAs on different biological processes, various studies have shown that mutations affecting the function of microRNA may play an important role in human diseases. Recently, microRNAs have been found to have a significant effect on various human diseases. Additionally, developmental studies increasingly focus on the use of microRNAs for the diagnosis and treatment of human diseases [16]. microRNAs clinically demonstrate an important relationship between the innate and adaptive immune systems; and deficiencies or excesses of miRNA cause many important diseases. For example, Jiang et al. demonstrated the relationships between microRNAs and diseases in miR2Disease, revealing the pathogenic role of microRNA deregulation in various conditions, including cardiovascular disease, cancer, and metabolic disorders [17]. In related cancer research, it has been found that abnormalities of miRNA in cells also cause healthy cells to transform into malignant cells [18] [19]. In addition, several studies have demonstrated the properties of miRNAs as tumor suppressor genes [20]. Huang et al. demonstrated that CD44 is suppressed and leads to breast cancer due to the upregulation of miR-520c and miR-373 [21]. Most of these existing approaches present the identified miRNAs on human complex diseases, and the performance of the machine learning methods using similarity networks (disease-disease similarity network, miRNA-miRNA similarity network, miRNA-disease similarity network). However, many of these methods lack sufficient information on data partitioning, CV technique and data preprocessing steps, which could negatively impact performance results and limit the repeatability of results. Additionally, the existing studies do not present a detailed performance evaluation. This thesis presents a novel approach named miRdisNET that helps us to discover microRNA biomarkers that are associated with diseases utilizing biological knowledge-based Machine Learning (ML). Compared with traditional ML approaches, biological knowledge-based ML approaches exploit known relations between biological entities; and incorporate that information into the ML algorithm. Incorporating biomedical knowledge into machine learning models can reveal patterns in noisy data and aid model interpretation. Along this line, in this thesis we have incorporated the knowledge of known miRNA-disease associations as biological information and developed an ML method to solve the classification problem of predicting patients vs. healthy controls using epigenomic data (miRNA expression profiles). Within our ML approach, the most informative miRNAs are suggested as

potential miRNA biomarkers of disease under investigation. In this way, promising miRNA-disease relationships are estimated by extracting meaningful insights from known disease-miRNA relationships (biological knowledge) and by using machine learning methods.

Table 1.1 Performance results of the existing studies for miRdisNET.

Study	Year	Disease	Dataset	Classification Method	Number of miRNA and Disease	Cross-Validation	AUC (%)
[10]	2017	Cancer	HMDD	KNN	495 miRNAs 383 diseases 5430 associations	Leave-one-out cross validation (LOOCV)	0.82
[11]	2019	Cancer	HMDD	RF	495 miRNAs 383 diseases 5430 associations	Local LOOCV, 5-fold cross validation, Global LOOCV	0.93
[12]	2021	Cancer	HMDD	XGBoost	495 miRNAs 383 diseases 5430 associations	5-fold cross validation	0.95
[13]	2019	Cancer	HMDD	KNN	495 miRNAs 383 diseases 5430 associations	LOOCV	0.90
[14]	2020	Cancer	HMDD	LR	495 miRNAs 383 diseases 5430 associations	5-fold cross-validation	0.92
[15]	2022	Cancer	HMDD	RF	495 miRNAs 383 diseases 5430 associations	5-fold cross validation and 10-fold cross validation	0.95

For microBiomeGSM, in the literature, there are numerous articles investigating microbiomes associated with three specific diseases: Colorectal Cancer (CRC), Type 2 Diabetes (T2D) and Inflammatory Bowel Disease (IBD). In particular, several studies aiming to uncover microbiomes related to T2D are summarized ([22], [23], [24], [25]). Microbiomes associated with CRC are reviewed [26], [27], [28], [29]. Several studies reviews the microbiomes associated with IBD [30], [31], [32], [33], [34], [35]. More specifically, Deschênes et al. [36] employed machine learning techniques to predict diseases by representing microbiomes using gene-based representations and taxonomic profiles. Through the creation of taxonomic profiles from shotgun metagenomic data, they identified significant taxa using their proposed methodology. They conducted experiments for five different diseases, namely obesity, T2D, CRC, liver cirrhosis, and IBD. For both CRC and IBD disease, the datasets used in [36] are the same datasets used by the proposed approach in the present study. In their study, they assessed the performance of nine distinct classifiers, including random forest, decision tree, two support vector machines with a linear kernel, random set coverage machine (rSCM), two logistic regressions, SVM with a radial basis function kernel (SVMrbf), and an ensemble algorithm derived from set coverage machine (SCM). For each dataset, they applied

embedded feature selection techniques, such as random forest and ranking features based on resulting models, followed by machine learning model application. They reported improved classification performance for certain diseases by employing taxonomic profiling. The most effective results in taxonomic profiling were achieved using the random forest algorithm for liver cirrhosis, yielding an AUC of 0.88. Their study demonstrated the effective use of converting microbiome data into taxonomic representation data for disease prediction. They reported that Lachnospiraceae microbiome is found to be associated with T2D and it can be considered as a biomarker for this disease. Sharma et al. [37] predicted disease states using machine learning methods by examining related Operational Taxonomic Units (OTUs) at the same phylum taxonomic level, exploiting the connections among OTUs at this taxonomic rank. Their investigation focused on the relationship between disease and the microbiome, utilizing shotgun datasets for two distinct diseases, T2D and Cirrhosis. The dataset they chose for T2D analysis is the same as the dataset used by our proposed tool. They applied their proposed method, which they called "TaxoNN," to a dataset with 174 cases and 170 controls for T2D [38] and a dataset with 118 cases and 114 controls for cirrhosis [39]. TaxoNN is a Deep Learning based multi-layered approach to group OTU information based on phylum clusters. It trains clusters containing OTUs that share the same phylum separately using Convolutional Neural Networks (CNNs). It combines features from each cluster to enhance prediction accuracy via an ensemble learning technique. Their proposed method was evaluated using six different classifiers, including Random Forest, Gaussian Bayes Classifier, Naive Bayes, Ridge Regression, Lasso Regression, and Support Vector Machines. The TaxoNN method yielded the highest result, achieving an AUC of 92% for cirrhosis and 75% for T2D. Moreover, TaxoNN identified microbiomes at the level of three dominant phyla (Firmicutes, Proteobacteria, and Actinobacteria) for both diseases, highlighting their impact on the diseases. Giliberti et al. [40] investigated the influence of the relative abundance of microbial taxa on host phenotype classification using human metagenomes. They employed machine learning methods to construct species-level taxonomic profiles and accurately detected the presence of microbial taxa. In their evaluation scheme, they encompassed a total of 4,128 samples from 25 shotgun metagenomic datasets. Among the datasets used in their study, T2D dataset is same with the dataset used in this study. They also explored the effect on disease prediction using relative abundance values at three different taxonomic levels: genus, family, and order. Employing the Random Forest classification algorithm on species level dataset, they

achieved the best performance for IBD dataset, across other datasets containing seven distinct disease categories (atherosclerotic cardiovascular disease, Alzheimer's disease, Behçet's disease, colorectal cancer, irritable bowel disease, type 1 diabetes, and type 2 diabetes). They identified statistically significant microbiomes for the diseases they identified. Among these microbiomes for these cases, the most significant result was obtained for Clostridium and this microbiome was followed by Streptococcus and Ruthenibacterium. Pasolli et al. [41] investigated the utility of microbiomes in disease prediction using metagenomic datasets for five different diseases: liver cirrhosis, CRC, IBD, obesity, and T2D. Among the datasets used in this study, T2D dataset is also utilized within this study. They conducted species-level prediction using microbiome profiles at the species level derived from metagenomic data. Their analysis encompassed a total of 2,424 shotgun metagenomic data samples from eight distinct studies. Employing cross-validation techniques, they compared classification outcomes using two widely employed classifiers in metagenomic data analysis, Random Forest, and Support Vector Machine. In addition to these classifiers, they also evaluated the effectiveness of elastic network, neural network, and multiple regression methods. In addition to predicting diseases using microbiome data, they highlighted prominent microbiomes related to these diseases. Notably, they identified the Peptostreptococcus microbiome for colorectal cancer, the Streptococcus microbiome for T2D, and the Lachnospiraceae microbiome for IBD as influential microbiomes in disease prediction. Collectively, these studies advance our understanding for the potential role of the microbiome in these diseases using a variety of approaches and analyses.

Table 1.2 Performance results of the existing studies for microBiomeGSM.

Study	Year	Disease	Classification Method	Number of Features	Cross-Validation	AUC (%)
[36]	2023	T2D, CRC, IBD,..	Nine different classifier (RF, DT, SVM,…)	-	10-fold cross-validation	0.68 (T2D) 0.78 (IBD)
[37]	2020	Cirrhosis and T2D	CNN	184 (Cirrhosis) and 208 (T2D)	10-fold cross validation	0.75 (T2D) 0.92 (Cirr)
[40]	2022	T2D, CRC, IBD,..	RF, SVM, Enet, Lasso	-	10-fold cross validation	0.75 (T2D) 0.92 (CRC) 0.99 (IBD)
[41]	2016	CRC, IBD, T2D,…	SVM, RF, Enet, Lasso, LR	503 (CRC) 443 (IBD) 572 (T2D)	10-fold cross validation	0.87 (CRC) 0.89 (IBD) 0.74 (T2D)

For CCPRED, recently, a growing number of studies [42] [43] [44] [45] [46] [47] [48] [49] [50] have attempted to establish a link between the gut microbiota and CRC. These studies have provided compelling evidence for significant and notable changes in

the microbiota of individuals suffering from CRC [51] [52]. While these studies have underscored the central role of the gut microbiota in the pathophysiology of CRC, it is worth noting that the field is still in its early stages before it comes a fully established science. Different classification criteria and methods used in previous research studies have made the identification of key species involved in the development and progression of CRC somewhat challenging [53] [54]. Given the complexity of metagenomic studies, the application of machine learning techniques has become increasingly important in this newly emerging field, offering the possibility of answering a wide range of questions. In this context, the idea of finding taxonomic biomarkers for diseases by correlating the microbiome with disease states through taxonomically informed feature selection has been developed [55]. In a study conducted by Bose et al. [56], the researchers used data from different populations spanning four different regions—Argentina, Chile, Vietnam, and India. They used taxonomic microbiome data analyzed at the genus level and focused primarily on examining the microbiomes that exerted significant influence in the Indian population. Their findings highlighted the prominent role of the Prevotella in CRC in the Indian population. In developing their machine learning model, Bose et al. obtained a high AUC value of 86% using the Random Forest classifier. This AUC value indicates strong performance of their model in discriminating between positive and negative cases and demonstrates its effectiveness in predicting CRC based on the analyzed microbiome data. A high AUC value indicates that the model has a good balance between sensitivity and specificity, which is critical for accurate disease prediction. In addition, Bose et al. underscored the existence of a global CRC microbiome and pointed out that certain observations from their study are consistent with results from other studies in various regions. This emphasizes the notion that there may be common characteristics and microbial influences in colorectal cancer development and progression that transcend geographic boundaries and are important on a global scale. Zhen et al. [57] conducted an in-depth study involving CRC patients and controls from diverse populations. In their comprehensive review article, they meticulously examined the results from an extensive pool of 700 different studies. Through this exhaustive analysis, the due importance of Fusobacterium nucleatum in the context of CRC was brought to light. This highlights the dramatic role of this particular microorganism in the landscape of CRC research and its potential relevance as a biomarker or target for further investigation and therapeutic intervention. Yu et al. [58] conducted a thorough population survey on CRC that included diverse populations from various countries. In their comprehensive review of 5696

different studies, they gained important insights into the microbiome associated with CRC and identified numerous influential factors that contribute to the development and progression of this disease. Their comprehensive research sheds light on the complex interplay of factors and microbial communities involved in CRC and contributes significantly to the understanding of this disease.

Table 1.3 Performance results of the existing studies for CCPRED.

Study	Year	Dataset	Classification Method	Cross-Validation	AUC (%)
[46]	2021	Asian Populations	Naïve Bayes	-	0.88
[48]	2022	Japan, China, USA, Germany, France and Austria	Lasso	5-fold cross-validation	0.84
[56]	2023	Vietnam, India, Chile, Argentina	RF	-	0.86
[57]	Review paper				
[58]	Review paper				

1.4 Motivation of This Thesis

The aim of this thesis is to first build a robust classification model using machine learning algorithms for disease prediction and analyze the performance of this model. Then, different approaches are proposed to highlight effective biomarkers for diseases. In the first of the three different studies presented in this thesis (miRdisNET), disease prediction and identification of potential biomarkers are performed using miRNA data. Biological domain knowledge is used for disease prediction and classification is performed using the G-S-M approach. The potential biomarkers identified in this thesis are supported by literature studies. The literature contains miRNAs that have already been shown to be associated with diseases by experiments, which underlines the power of the proposed approach. Potential miRNA biomarkers identified as associated with related diseases are expected to shed light on future studies for the diagnosis and treatment of these diseases. The miRNAs not mentioned in the literature will serve as a guide for future studies.

In the second (microBiomeGSM) and third study (CCPRED), a powerful classification model was created using metagenomic data to predict diseases and identify potential biomarkers. The microBiomeGSM tool is based on the G-S-M approach and uses biological domain knowledge for grouping. Based on the microbiomes associated with the identified diseases, a disease prediction is performed, and possible biomarkers

are identified. The third study (CCPRED) uses the intersection and union features of the features selected by the feature selection algorithms and investigates the success of these features in classification performance. Different sub-approaches are applied to approaches with high classification performance and effective features are highlighted and identified as potential biomarkers.

1.5 Limitation of This Thesis

While this thesis endeavors to contribute to the field of disease prediction and biomarker identification using machine learning approaches, it is essential to acknowledge several limitations inherent in the methodology and scope of the research. I would like to mention some limitations of this thesis. One of the primary challenges encountered during the research process was the availability and quality of data. The complexity of biological systems poses inherent challenges in accurately modeling disease processes and identifying biomarkers. While machine learning techniques offer powerful tools for analysis, they may oversimplify or fail to capture the intricate interactions within biological pathways, leading to potential biases or inaccuracies in predictions. Feature selection and machine learning methods were selected as the most used methods. The integration of the current methods into the proposed models will be investigated. The applicability of the developed models and identified biomarkers may be limited to the specific datasets and diseases examined in this study. Extrapolating findings to broader populations or disease contexts requires careful consideration of underlying biological mechanisms and external validation in independent cohorts.

Chapter 2

Materials and Methods

2.1 Dataset and Data Preprocessing

2.1.1 Dataset of miRdisNET

We used the Human microRNA Disease Database (HMDD) v3.2 (https://www.cuilab.cn/hmdd) for obtaining disease-miRNA associations. We downloaded the entire database including 1206 miRNAs, 894 diseases, and 18732 experimentally verified miRNA-disease associations. We have extracted the relevant sets of miRNAs related to each disease. A few examples of miRNA-disease associations are shown in Table 2.1. Table 2.1 presents sample disease groups, i.e., Acute Brucellosis, Alopecia, Cataract, Carcinoma Embryonal and Pancreatic Diseases. For example, Group 1 is represented by Alopecia, and Group 2 is represented by Acute Brucellosis disease. Group 1 has 10 associated miRNAs (hsa-miR-106b, hsa-miR-125b-1, hsa-miR-125b-2, hsa-miR-221, hsa-miR-410, hsa-miR-203, hsa-miR-575, hsa-miR-602, hsa-miR-106a, hsa-miR-125b) based on HMDD database. On the other hand, Group 2 includes only two associated miRNAs (hsa-miR-126, hsa-miR-4753) according to HMDD. This indicates that the association between these two miRNAs and Acute Brucellosis is experimentally verified, based on HMDD.

miRCancer database [59], which contains miRNA-cancer associations is used to evaluate and validate the prediction lists of our miRdisNET tool. miRCancer includes 876 different miRNA-disease associations between 236 miRNAs and 79 human cancers with more than 26 thousand published articles in PubMed. miRCancer provides a web interface for the study of miRNA-cancer associations. The results obtained by miRCancer are validated in PubMed and in miRBase.

Table 2.1 An example grouping procedure based on miRNA-disease relationships using HMDD

Disease	miRNA
Alopecia	hsa-mir-106b, hsa-mir-125b-1, hsa-mir-125b-2, hsa-mir-221, hsa-mir-410, hsa-mir-203, hsa-mir-575, hsa-mir-602, hsa-mir-106a, hsa-mir-125b
Acute Brucellosis	hsa-mir-126, hsa-mir-4753
Cataract	hsa-mir-184, hsa-mir-125b, hsa-mir-589, hsa-mir-326, hsa-mir-675, hsa-mir-34a, hsa-mir-15a
Carcinoma, Embryonal	hsa-mir-372, hsa-mir-373, hsa-mir-29c, hsa-mir-19, hsa-mir-29c, hsa-mir-134, hsa-mir-140, hsa-mir-302b, hsa-mir-27, hsa-mir-34a, hsa-mir-601
Pancreatic Diseases	hsa-let-7b, hsa-mir-495

Table 2.2 Details of the TCGA datasets used in miRdisNET.

TCGA cancer types	Normal	Tumor	Pubmed ID
Breast Invasive Carcinoma (BRCA)	87	760	PMID: 31878981
Stomach Adenocarcinoma (STAD)	35	370	PMID: 25079317
Kidney Chromophobe (KICH)	25	66	PMID: 25155756
Uterine Corpus Endometrial Carcinoma (UCEC)	23	174	PMID: 23636398
Kidney Renal Papillary Cell Carcinoma (KIRP)	32	291	PMID: 28780132
Lung Adenocarcinoma (LUAD)	20	449	PMID: 25079552
Bladder Urothelial Carcinoma (BLCA)	19	405	PMID: 24476821
Prostate Adenocarcinoma (PRAD)	52	494	PMID: 26544944
Kidney Renal Clear Cell Carcinoma (KIRC)	71	255	PMID: 23792563
Papillary Thyroid Carcinoma (THCA)	59	512	PMID: 25417114
Lung Squamous Cell Carcinoma (LUSC)	38	342	PMID: 22960745

The Cancer Genome Atlas (TCGA) project provides comprehensive data to obtain the expression profiles of several different miRNAs in cancer samples. To test our miRdisNET tool, we downloaded 11 cancer miRNA expression profiles from the TCGA portal (https://portal.gdc.cancer.gov/). The datasets contained paired data (tumor samples and matched normal samples) from HiSeq platform, where miRNA was selected only if 50% of the samples had normalized expression value > 1. All of the expression profiles were normalized to RPM (Reads per Million). Further details of the processing steps can be found in [60]). The detailed cancer names, sample sizes, and PubMed accession numbers are presented in Table 2.2.

2.1.2 Dataset of microBiomeGSM

The data used in this study are obtained from the NCBI Sequence Read Archive (SRA045646, SRA050230) provided by [38] for T2D; accession number PRJNA398089 in the SRA for the Integrative Human Microbiome Project for IBDMDB [61]. IBD dataset is obtained from the MetaHit project [62] (ERA000116). The CRC metagenomic dataset containing 1262 samples was created by [61]. Microbiome sequencing data is classified into disease states based on the metadata associated with them. To ensure data quality, we applied quality filtering to meet the standards outlined in the Human Microbiome Project Consortium SOP (2012), as referenced in [63]. This procedure allowed us to categorize the raw sequencing data according to relevant disease states, enabling our subsequent analyses. The microbiome samples were associated with the microbial species of origin (taxa) using the MetaPhlAn tool [64], and the relative abundance composition for each taxon was generated accordingly. These taxa and their relative abundances serve as features or variables in our machine learning approaches. MetaPhlAn first assigns reads to microbial clusters using clade-specific genes for assignment. It then presents the relative abundance of microbial taxa based on these readings. In this study, the assignment to microbial species of origin (taxa) was determined for each DNA sequence using the MetaPhlAn tool. The relative abundance value is normalized by dividing the number of reads for each taxonomic level by the total number of reads for only one sample. In this way, the taxonomic abundance values are expressed as real numbers in the range [0,1] with a sum of 1 for each sample. Samples with less than one million total reads were not included in our study. For each sample, we determined the diversity of disease-relevant microbiomes, where diversity represents the presence and relative abundance of microorganisms [65].

Table 2.3 The list of datasets used to test the model. Number of samples who have positive class label are shown in the second column. The number of features/Species is shown in the third column. Number of groups created at Order, Family and Genus taxonomic levels are listed at the 4-6th columns respectively.

#	Dataset	# of Samples	# of positives	# of features (Species)	# of Groups (Genus)	# of Groups (Family)	# of Groups (Order)
1	CRC	1262	600	912	261	100	49
2	IBDMDB	1638	1209	579	187	77	43
3	IBD	382	148	1456	448	177	84
4	T2D	290	155	1456	448	177	84

The four microbiome datasets used to evaluate the microBiomeGSM tool are listed in Table 2.3. The table presents the number of samples in each dataset and the number of samples that are labeled as positive. Positive samples refer to patients, while negative samples refer to controls. Each dataset contains the abundance values of the species, which we consider as features. We have considered 3 taxonomic levels for creating the groups, i.e., genus, family, and order. For each dataset, the number of extracted groups is listed in the corresponding column, while '-' denotes missing information.

2.1.3 Dataset of CCPRED

Beghini et al. (2021) compiled a total of 1262 metagenomic samples (662 controls and 600 CRC patients) at different molecular levels (species, enzymes and pathways) from nine different datasets [61]. In their study, the raw microbiome DNA of each sample was downloaded from the respective project site and MetaPhlAn [64] and HUMAnN [61] were used to calculate the relative abundance values of all subgroups of each dataset. To ensure data quality, they applied quality filtering to meet the standards outlined in the Human Microbiome Project Consortium SOP (2012), as referenced in [63].

Table 2.4. Distribution of data by populations and the numbers of samples in each population

Name of population	# of Samples	# of Healthy Samples	# of CRC samples
Austria (AUT)	107	46	61
China (CHN)	128	75	53
Germany (DEU)	125	60	65
France (FRA)	114	53	61
Indian (IND)	60	30	30
Italy (ITA)	106	57	49
Japan (JP)	80	40	40
Japan (JPN)	438	187	251
United State of America (USA)	104	52	52

The CRC-associated metagenomics dataset used in the present study includes the relative abundance values of 917 different species, 2895 different enzymes, and 551 different pathways calculated for 1262 samples from nine different datasets. The

distribution of data by population and the number of the samples in each population is shown in Table 2.4.

2.2 Feature Selection Algorithms

There is where that the number of observations used for the training data is higher than the number of observations used for the test data. This situation is undesirable if studies are to produce more effective results, and researchers are proposing various methods of resolution, particularly feature selection methods. Although the process of feature selection in disease prediction problems based on metagenome data has not been well studied, the literature suggests that this process may be as important as the choice of a classification method [32]. The process of feature selection in metagenome-based disease prediction could help us learn more about disease development mechanisms. Therefore, further research in this direction is warranted. In metagenomics studies, in order to reduce the number of taxa , i.e., to select informative species (features), min Redundancy Max Relevance (mRMR) [66], Lasso [67], Elastic Net [68], and the iterative sure select algorithm [69] have been used extensively. Another feature selection method, called Fizzy, addresses the challenge of using classification techniques to identify important functional elements for downstream analysis [64]. Oudah and Henschel presented an alternative taxonomy-based method for feature selection [70]. Bakir-Gungor et al., (2021) applied CMIM [71], FCBF [72], mRMR [66], and Select K best (SKB) [73] to type 2 diabetes-associated metagenomics datasets and obtained powerful performance metrics [74]. Jabeer et al. also proposed a robust classification method for evaluating colorectal cancer associated metagenomic datasets using a combination of feature selection methods and machine learning methods [75]. Bakir-Gungor et al., (2022) also proposed a powerful method for IBD classification with fewer features by combining feature selection methods and machine learning methods [7]. While these feature selection approaches have produced effective results in a variety of fields, they have only recently been applied to microbiome-based disease prediction problems.

2.2.1 Minimum Redundancy Maximum Relevance (mRMR)

mRMR is a feature selection method that favors features that are not strongly correlated with each other and have a strong relation to the output (class). The mRMR criterion is an instrument for evaluating the importance of a feature. With this strategy,

features can be categorized according to how relevant they are to the goal and how redundant they are. A feature with a lower value indicates that maximum relevance and minimum redundancy have been better balanced [76]. The main goal of this method is to identify the most valuable features in the dataset and ignore the others. The mRMR approach treats each feature independently and uses the mutual information to measure the similarity between the features [77].

2.2.2 Conditional Mutual Information Maximization (CMIM)

CMIM is a very fast and efficient method for selecting multivariate filter features, which was developed by Fleuret [71] and is derived from Conditional Mutual Information (CMI). CMIM uses CMI to determine relevance and redundancy. Based on a preselected feature set, it selects features that maximize the mutual information with the class prediction. This criterion determines that the selected features are not meaningful on their own, as they do not provide additional information about the class prediction, although they are different from the previously selected features. This approach offers a balanced choice between relevance and redundancy. By comparing each new feature with the previously selected features, the CMIM approach attempts to achieve a balanced compromise between the individual influence and the independence of the features [78].

2.2.3 Extreme Gradient Boosting (XGBoost)

Chen and Guestrin (2016) presented Extreme Gradient Boosting (XGBoost), an effective and scalable machine learning technique [79]. This model was created by fusing multiple Categorization and Regression Trees (CART) with Gradient Boosting Decision Trees (GBDT). The basic idea is to train a new model with the residuals of the old model so that the new model can correct the errors of the old model. To increase the accuracy of the model, XGBoost adds a regularization term to the objective function, which regulates the complexity of the model to prevent overfitting and approximates the loss function using the quadratic Taylor expansion. This method achieves better generalization capabilities [80].

2.2.4 Information Gain (IG)

Information gain (IG) is a measure used in feature selection by evaluating the contribution of each variable to the target variable. The information gain is calculated for

the independent variables and then the features are ranked in descending order according to their individual information gain. A cut-off point is chosen and all features above this cut-off point are included in the machine learning algorithms. In this way, the most important features are identified and included in the modeling process [81].

2.2.5 Select K best (SKB)

Select K Best (SKB) [73] is a feature selection algorithm (FS) and is used to select the k attributes with the highest scoring attributes. The scores are calculated using a test that evaluates the relationship between each attribute and the target. The most common of these tests are the Chi-square test, the F-test and the ANOVA F-test. In addition, the recursive feature elimination (RFE) method can also be used within the SKB algorithm. This method attempts to select the best feature by sequentially removing features [82].

2.2.6 Fast Correlation-Based Filter (FCBF)

FCBF (Fast Correlation-Based Filter) was developed by [83]. It is a multivariate feature selection procedure that takes into account the quality of features to determine which subset of features is the best. The method starts with an entire set of features and evaluates the dependency between the features using symmetric uncertainty. Symmetric uncertainty (SU) is an information-theoretic metric used to quantify the dependencies between features. It is a normalized combination of entropy and conditional entropy values. Compared to other subset selection techniques, FCBF, a correlation-based method for selecting subsets of features, is faster [72].

2.3 Machine Learning Algorithms

The modeling of systems that generate predictions by drawing conclusions from data using statistical and mathematical operations is known as machine learning. In this process, a model is developed, and a result is obtained by analyzing data. In order to generate precise forecasts, the model continuously refines itself on the basis of data and learns from it. Machine learning algorithms do not follow a predetermined set of rules but learn from data sets to make judgments. Numerous areas, including data mining, image processing, robotics, bioinformatics, and natural language processing, can benefit from the application of these algorithms.

Recently, there has been increasing interest in machine learning techniques for predicting disease through the visualization of microbiomes using gene-based representations and taxonomic profiles. In addition, several computational methods have been proposed in recent years to predict associations between miRNA and disease, in particular using machine learning algorithms. Accurate prediction of diseases and potential biomarkers is crucial before time-consuming, costly and difficult manufacturing processes are undertaken. Accordingly, several computational approaches have been proposed to predict diseases and identify promising biomarkers without expensive wet-lab testing.

2.3.1 Random Forest (RF)

The Random Forest (RF) algorithm is a supervised machine learning technique that is often used in classification and regression tasks. This algorithm is based on the principle of ensemble learning, where multiple decision trees are combined on different subsets of input data to improve prediction accuracy. Random Forest is considered a fundamental concept in machine learning and its effectiveness, and problem-solving capabilities increase in direct proportion to the number of trees it contains [84].

Random Forest parameters can help to improve the predictive power of the model or simplify the training process of the model. Many parameters are optimized to improve the performance of the algorithm. The parameter "max_features" specifies the maximum number of features used in each tree. This parameter controls how much diversity the random forest can generate in each decision tree. The parameter "n_estimators" refers to the number of trees that should be generated before the maximum matching of predictions occurs. This parameter determines the total number of trees in the random forest, and in general, the more trees, the better the performance. The parameter "max_depth" specifies the length of the longest path from the root node of a tree to the leaf node. This parameter controls how deep each tree is and can help avoid overfitting issues. Default values are used for all parameters provided by the scikit-learn library. This provides a starting point that users can easily use instead of specifying parameters.

2.3.2 Support Vector Machine (SVM)

For both regression and classification tasks, the Support Vector Machine (SVM) is a powerful machine learning method with strong theoretical foundations and excellent

generalization capabilities. Statistical learning theory (SLT) is the source of SVM, which is based on the idea of structural risk minimization (SRM). Its many properties include sparsity, the kernel technique and a singular global solution. Numerous pattern categorization tasks, including bioinformatics, face recognition, text categorization, handwritten character and digit recognition, and cancer detection, have been successfully solved using the SVM technique [85].

The two most important hyperparameters used by the Support Vector Machine (SVM) are the penalty C for the classifier and the gamma parameter, which controls the radius of influence. The parameter C determines the penalty for training misclassifications in the classifier. If C is very large, a high penalty is applied to misclassifications in training so that the margin is minimized. A low C leads to a lower penalty and a larger margin. The radius of influence of a training point is determined by the gamma parameter. A large similarity radius, which is indicated by low gamma values, can lead to more points being grouped. For points with high gamma values to belong in the same group (or class), they must be relatively close to each other. The Scikit-learn library determines the C and Gamma parameters using default parameter values. This is an option that the user can use as a starting point before setting their own parameters.

2.3.3 Decision Tree (DT)

One of the most fascinating applications of machine learning in the field of medicine are decision trees (DT). They provide a clearer picture of the contribution of each variable to the prediction than other algorithms. This makes it possible to understand which features are important for analyzing clinical data and to make an informed choice. Selecting the features that is at the root node and from which the branches recursively emanate is the first step in building a DT. Each instance available at a node is progressively "distributed" to the branches, which are formed based on the standards established at each stage. The development of the branches ends when all instances have the same classification at any point in time. If a further division of the data is not possible or if the creation of new branches does not improve the categorization, the branch evolution also ends. The requirement to create the smallest possible tree influences the selection of nodes to be included in the tree as well as the root node. Consequently, at each stage, the feature that produces the "purest" node—, i.e. the one with the largest number of occurrences belonging to the same classification, is selected. The entropy of the class distribution can be used to evaluate the information gain, which increases with

the average purity of the subsets created by splitting the samples and is usually used to determine the nodes to be selected [86].

2.3.4 LogitBoost

LogitBoost is an iterative machine learning algorithm that combines multiple weak classifiers to improve classification performance. For binary classification problems in particular, Logit Boost trains weak classifiers, usually decision trees, and aims to combine them into a strong classifier. In doing so, each weak classifier is weighted based on log-odds ratios when classifying instances, and the subsequent weak classifiers focus on correcting previous errors. In this way, the logit boost algorithm improves classification performance by creating a strong classifier that is created by combining weak classifiers [87].

2.3.5 AdaBoost

The boosting technique creates a strong learner by combining several weak learners. With this method, the predictors are trained step by step. First, a weak learner is used to train the training set. After the training phase, the incorrectly predicted examples are of great importance for this algorithm. In the first iteration, more emphasis is placed on the mislearned data, and these examples are considered again in the next training phase to obtain a more accurate model. In this way, a new learner is added in each iteration and a stronger learner is created by focusing on the errors [88].

Chapter 3

mirDisNET

3.1 Motivation

miRdisNET detects microRNAs that are associated with the disease based on the Grouping, Scoring, and Modeling (G-S-M) approach. We first construct specific disease groups containing the related miRNAs. Secondly, each group is scored by the tool to assign a score of its importance in the two-class classification task. We implemented internal Monte-Carlo stratified cross-validation to evaluate the computational prediction performance of miRdisNET. We also evaluate miRdisNET from a biological point of view. To this end, miRNAs that are predicted by miRdisNET as associated with a specific disease is comparatively evaluated with biological databases.

3.2 Proposed Model

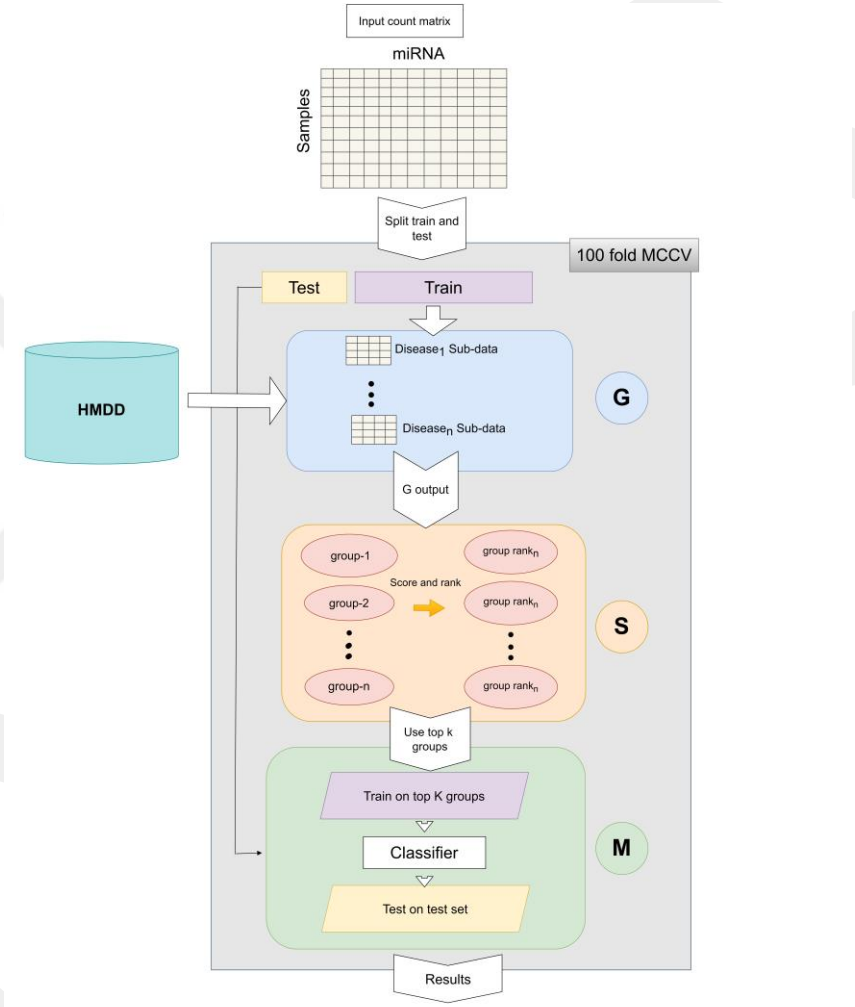
The general workflow of miRdisNET is illustrated in Figure 3.1. Based on the idea in the G-S-M approach, in this study the groups of miRdisNET are extracted from prior biological knowledge about the miRNAs that are associated with a specific disease (G component). A group is a disease, and its members are the miRNAs that are associated with this disease. Hence, from now on we refer to a set of miRNAs that are associated with a disease as the specific disease group. The aim of the miRdisNET is to score (S Component) the groups/diseases to detect the top significant groups to be used for training the classifier (M component).

As illustrated in Figure 3.1, the miRdisNET is based on three main components:

1. G Component: Creates the groups and its associated two-class subdatasets
2. S Component: Computes a score of each group (two-class subdataset) which measures to what extent it is differentially expressed.

3. M Component: Uses the miRNAs expression values from the top ranked groups to train the model. We have used the Random Forest classifier as the machine learning algorithm.

Let D represent the miRNA expression data set. D is split into $D_{\_train}$ and $D_{\_test}$. The $D_{\_train}$ is used for three different processes: (i) assigning an importance score for ranking, (ii) training the random forest classifier, (iii) building the model. However, $D_{\_test}$ is only used to evaluate the performance of the tool.



Performance Metric

# Groups	# miRNA	Acc	Sens	Spec
10	345	73.20	81.67	77.32
....	....	....	....	....

Top Disease groups and significant miRNAs

Disease group	MiRNA list	Score
Sarcoma, Kaposi	hsa-mir-101, hsa-mir-123.....	94
....	....	....

Figure 3.1 The general approach of miRdisNET. It consists of three components: G component generates sub-datasets for each specific disease group; S component performs scoring and then ranking of the specific disease groups; M component creates the classifier, trains and evaluates the performance of miRdisNET.

3.2.1 Component G (Grouping)

Figure 3.2 illustrates the flow of the grouping component G. The G component receives two inputs. The two-class miRNA expression dataset D, where the columns are the miRNAs, and the rows are the samples. The labels of the samples are indicated in the column "class" where the value 'pos' indicates the sample is obtained from a cancer patient and 'neg' indicates from healthy / normal sample. The R table is the groups. The name of the group is the disease name while the set is a set of miRNA names that are associated with the specific disease.

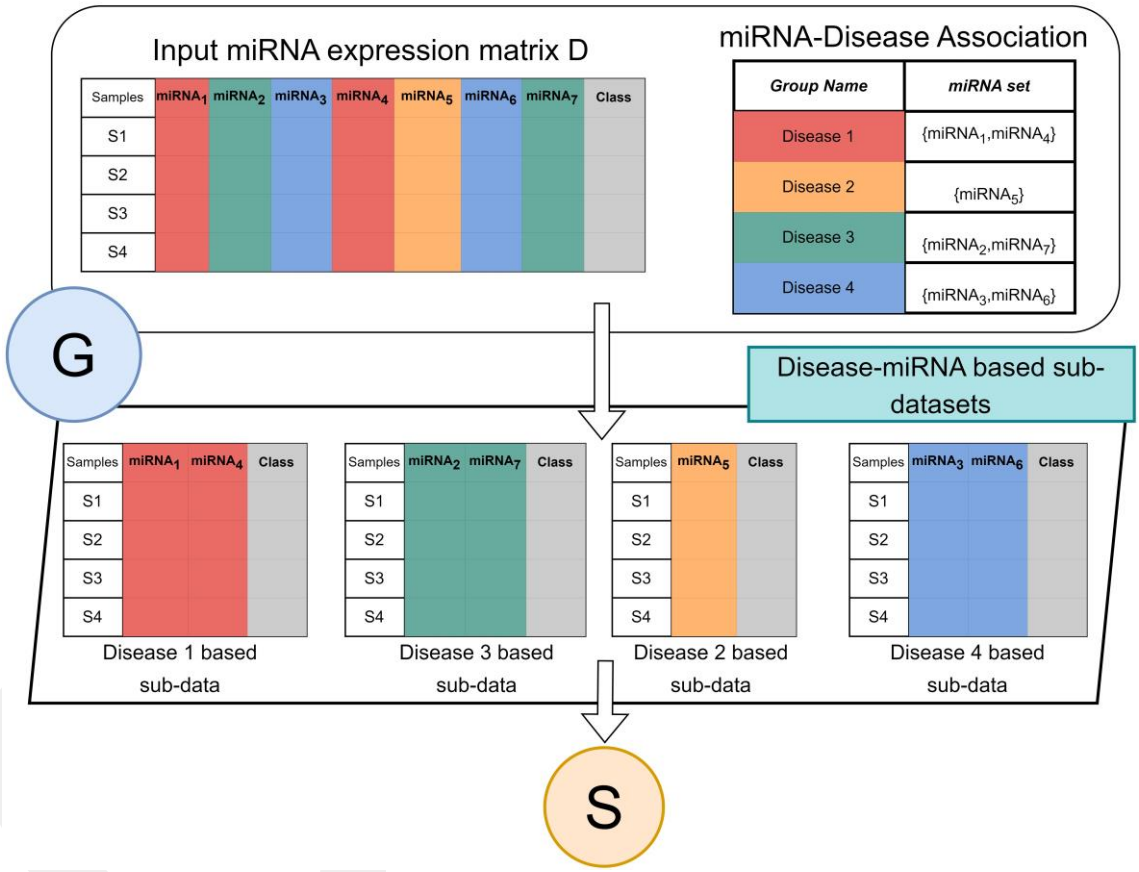


Figure 3.2 Architecture of G component for miRdisNET. An example showing how to construct disease sub-datasets based on miRNAs associated with a disease.

Component G creates for each group a two-class subdataset that extracts the miRNA columns from the data D with its class labels. Thus, each group is represented as a two-class sub dataset that will serve as an input to the S component for performing the scoring and ranking.

There are a total of 894 groups which correspond to unique diseases. Figure 3.3 represents the distribution of each disease group in terms of its size (the size of the respective miRNAs related to the disease). About 75% of the disease groups have 20

miRNAs which are associated with them, while a few groups have greater than 100 miRNA which are associated with the disease group.

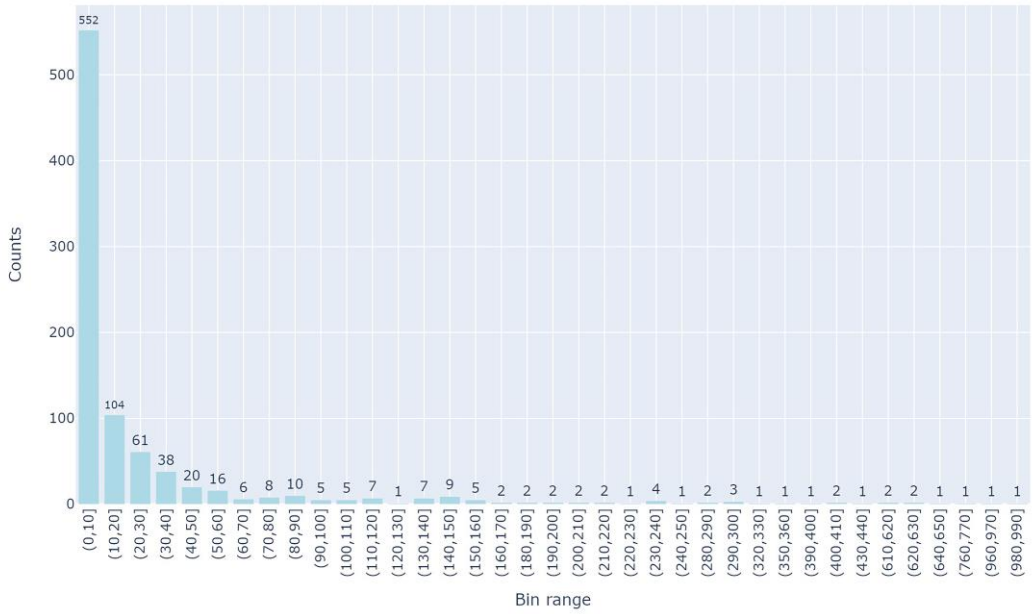


Figure 3.3 The distributions of miRNA in each of the groups. *Y*-axis is the number of miRNAs in a group and *X*-axis represents the group size which is binned in 10 intervals.

3.2.2 Component S (Scoring)

The second component is the scoring step, where a score is generated for each disease to assign an importance score to each disease group containing miRNAs associated with that disease, as shown in Figure 3.4. In this component S, the Random Forest algorithm is used for model training. In component S, machine learning model with the Monte Carlo cross-validation (MCVV) is used to assign an importance score for each disease found in each sub-dataset. In MCCV, the dataset is randomly divided into two groups: 70% of all known interactions as a training set, and 30% for the testing set. In order to solve the sample imbalance problem, an equal distribution among class labels (pos, neg) is achieved by applying the stratified sampling method. We repeated this approach five times to avoid overriding and to provide balance in the training and test datasets.

The main purpose of scoring is to generate the predictive value obtained by testing the class labels (pos, neg) of miRNAs associated with specific diseases. There are various performance evaluation metrics such as Recall, Accuracy, F1 score, Precession. This component focuses on mean classification accuracy as a performance evaluation metric

for assigning an importance score to diseases and ranking them according to that importance.

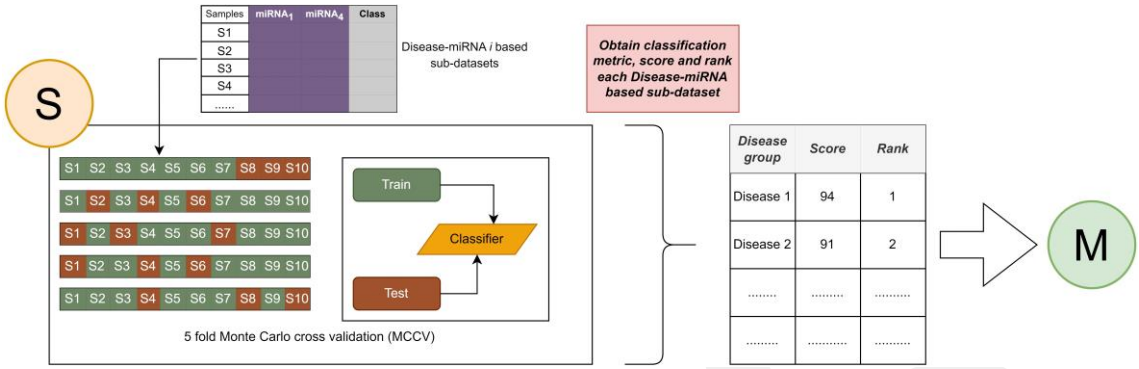


Figure 3.4 Assign an importance score to the associated disease and apply the ranking process.

Importance scores are assigned to diseases based on miRNA expressions from TCGA and relationships between disease and miRNA from the HMDD v3.2 dataset. Table 3.1 shows a sample output obtained after the scoring step for the BLCA dataset.

Table 3.1 An example output of the component S for the BLCA dataset. The first column represents the name of the disease, the second column is the mean accuracy, and the third column is the ranking based on the second column.

Disease	Score as Accuracy	Rank
Graft-versus-host disease	0.9636	1
Human immunodeficiency virus infection	0.9636	1
Hypertrophy	0.9636	1
Kaposi sarcoma	0.9636	1
Bladder carcinoma	0.9454	2
Acute promyelocytic leukemia	0.9454	2
Ischemia-reperfusion injury	0.9454	2
Oral squamous cell carcinoma	0.9272	3

3.2.3 Component M (Modeling)

The third component is represented by M, which contains two major processes: (i) train classifier (usually use random forest classifier), and (ii) create model. The main aim of this component is to evaluate the cumulative performance of the model and train the classifier to reveal the top-ranked miRNAs in an accumulated order. In each iteration and for each top-ranked group, component M randomly selects the training set for training and uses the remaining dataset as the test dataset to test this trained model.

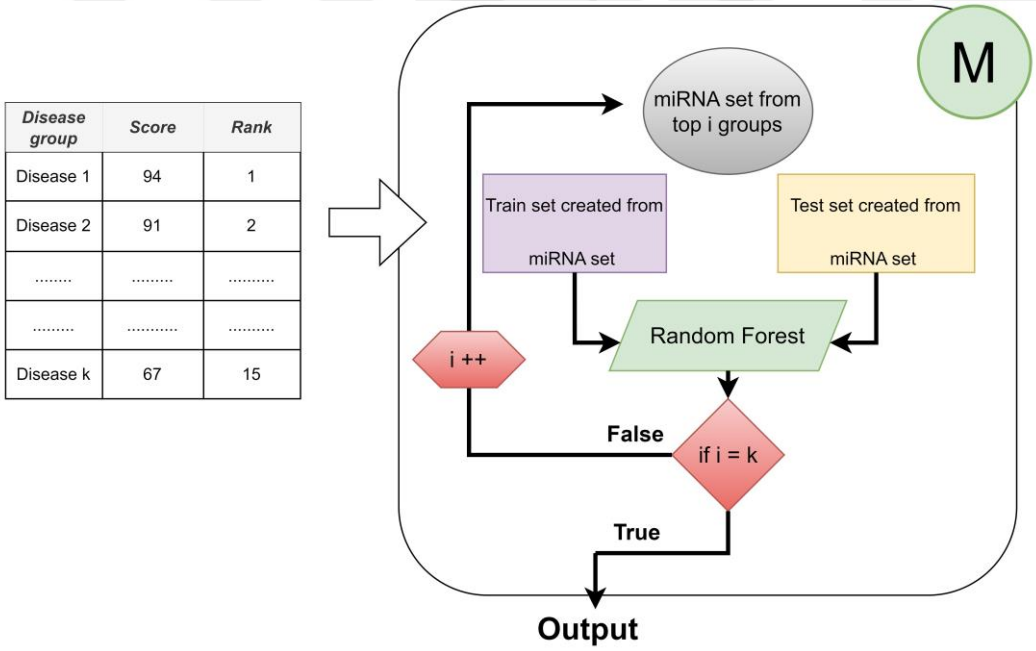
Component M contributes to the research with its three inherent processes as following:

• First iteration, building a machine learning model (Random Forest): only using the miRNA expression values of the top-scoring disease, where top-scoring disease is determined after applying the component S.

• Second iteration, accumulated groups: it combines the miRNA expressions belonging to the highest scoring disease and the miRNA expressions belonging to the second top-scoring disease. By this way, new sub data is created to train and test the model. This accumulative approach is repeated for top 3, top 4,…,top t groups, where t is the number of all disease groups.

• Component M is completed after all diseases have been processed in this manner.

By following this approach, we can find the best feature set that presents the best performance in terms of combinations of diseases, i.e., the top 1 scored disease, top 2 scored disease, until top 10 scored disease. Architecture of the component M is shown in Figure 3.5.



Disease group	Score	Rank
Disease 1	94	1
Disease 2	91	2
.....	.....	.....
.....	.....	.....
Disease k	67	15

# Groups	# miRNA	Acc	Sens	Spec	AUC
10	345	73.20	81.67	77.32	85.34
....	....	....	....	....	....

Disease group	MiRNA list	Score
Disease 1	$miRNA_1$, $miRNA_2$....	94
....	....	....

Figure 3.5 The architecture of Component M: Providing the best performance with the best feature set based on disease combinations.

3.3 Implementation of miRdisNET

miRdisNET tool have been implemented on the open-source KNIME platform. This platform can be used for a wide variety of data types and operations. Figure 3.6 illustrates how the workflow is implemented in KNIME. The user can set the parameters such as the number of iterations, rank function and number of iterations for MCCV. The user needs to select the miRNA dataset. The filter nodes remove any rows with missing values.

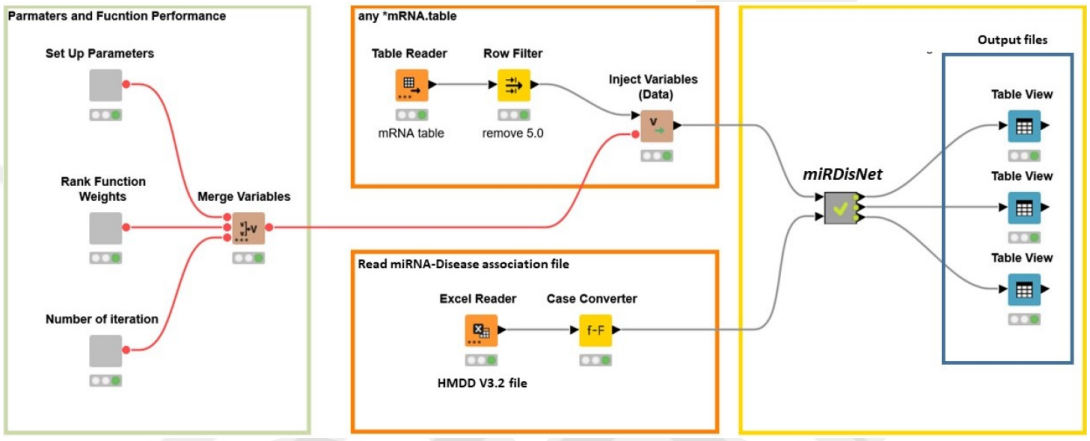


Figure 3.6 miRdisNET workflow in KNIME.

3.4 Model Performance Evaluation

To evaluate the predictive performance of miRdisNET, the input dataset was split into 90% for training, and 10% for testing. In this study, the class label of the dataset has an unequal distribution. In other words, the number of cases and controls is not equal. For this reason, we applied the under-sampling method for the unevenly (imbalanced) distributed dataset. This method reduces the size of the majority class, leaving all samples in the minority class, and solves the problem of the imbalanced dataset. We performed 100-fold Monte Carlo cross-validation (MCCV) for model training. MCCV has a repeatable structure due to its low variance, which makes it more effective than traditional cross-validation methods for miRdisNET. In MCCV, the data is randomly selected to train the model, and the remaining data issued as a test dataset. To obtain the criteria for performance evaluation, average values of 100-fold MCCV are calculated.

Various statistical methods are also used to comprehensively evaluate the performance of the Random Forest model such as Sensitivity, Specificity, and Accuracy. Area Under the Curve (AUC) is also used as one of the performance evaluation criteria

of classifiers. In this study, the performance of miRdisNET is evaluated according to the AUC measures.

In each iteration, we obtain lists of disease groups and miRNAs associated with those disease groups. Therefore, a prioritization approach is applied to assign importance scores to entities in both the disease and miRNA lists. For this purpose, we incorporated the algorithm called RobustRankAggreg [89], which is presented as an R package, to the miRdisNET workflow. The RobustRankAggreg method assigns a p-value to each entity (miRNA or disease) in the lists, indicating how well that entity ranks.

3.5 Results

3.5.1 Comparison with existing models

To evaluate the performance of miRdisNET in discovering potential miRNA–disease associations, miRdisNET is compared with several advanced methods such as RKNNMDA, HGIMDA, ABMDA. RKNNMDA uses disease similarity networks, miRNA similarity networks, gaussian interaction profile kernel similarity, and miRNA-disease relationships to identify potential associations between miRNA and disease. This tool implements the ranking-based KNN method by combining similarity matrices and disease-miRNA associations. They used the disease-miRNA associations obtained from the HMDD dataset in their study. They obtained an AUC of 0.8221 with the leave-one-out cross validation method. HGIMDA, a computational model is developed by integrating disease semantic similarity, miRNA functional similarity, gaussian interaction profile kernel similarity and verified miRNA-disease associations. They also used 5430 disease-miRNA associations obtained from the HMDD dataset in their study. This tool implemented global and local leave-one-out cross validation method and obtained an AUC of 0.8781 and 0.8077, respectively. ABMDA tool makes use of adaptive boosting for predicting the relationship between disease and miRNA. This tool performs random sampling based on k-means clustering to balance positive and negative samples. This tool integrates HMDD disease-miRNA association information and similarity matrices and obtains AUC of 0.9170 and 0.8220 by global and local leave-one-out cross validation, respectively. The AUC score of MCCV, achieved by miRdisNET by using the accumulated miRNA groups is shown in Figure 3.7. We evaluate the performance of miRdisNET using different cancer data samples presented in Table 2.2. The proposed

method shows the most important group with performance evaluation criteria using a machine learning method. As shown In Figure 3.7, the proposed method called miRdisNET has nearly an average AUC of 97% for all 11 TCGA datasets. The best results were obtained on average %99, %99, %99, %99 from KIRC, KICH, UCEC and KIRP, respectively.

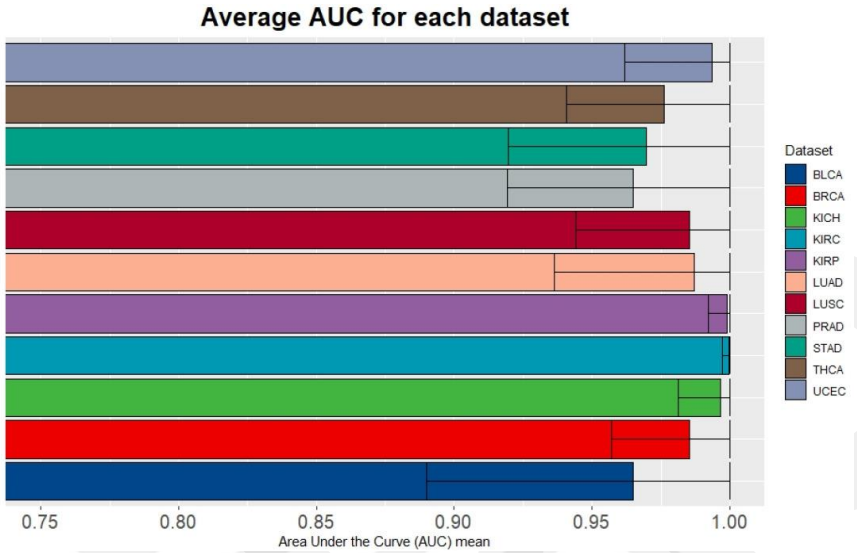


Figure 3.7 Average of AUC over the top 10 significant groups for all the 11 TCGA datasets.

The reasoning behind the higher AUC score of miRdisNET compared with other algorithms may be based on the following properties of the G-S-M approach:

i) miRdisNET considers relevant miRNAs for the grouping component.

ii) miRdisNET uses effective classifiers for the scoring component and highlights effective structures.

iii) For the modeling component, important disease groups are treated cumulatively with effective classifiers and classification techniques.

Therefore, with the developed classification techniques, the miRdisNET tool is applied to structures that are important for the disease and higher performance metrics as compared with other algorithms, are obtained.

One of the methods to evaluate the model performance is to compare the performances of miRdisNET models as a function of k parameters. k parameters are the number of groups (disease) in miRdisNET. Table 3.2 shows the performance obtained with 100-fold MCCV for the aggregated top-ranked 10 groups for the BLCA dataset. For group 1, we obtained a 95% AUC using an average of 6.76 miRNAs. For group 2, performance metrics for the top-ranked 2 groups are shown., combining the miRNAs from the first top-ranked group and those from the second top-ranked group. We obtained

a 96% AUC using an average of 10.36 miRNAs. In this way, miRdisNET provides cumulative performance results for the top 10 groups.

Table 3.2 A sample average table of 100-fold MCCV performances from miRdisNET for the top 10 ranked groups for the BLCA dataset cumulatively.

#Groups	#miRNAs	Accuracy	Sensitivity	Specificity	AUC
10	18.41	0.92	0.92	0.93	0.97
9	17.99	0.93	0.92	0.93	0.97
8	17.44	0.92	0.90	0.93	0.97
7	16.74	0.92	0.90	0.93	0.97
6	16.22	0.91	0.89	0.93	0.97
5	15.29	0.91	0.89	0.93	0.97
4	14.1	0.91	0.88	0.93	0.96
3	12.76	0.91	0.88	0.92	0.96
2	10.36	0.90	0.87	0.92	0.96
1	6.76	0.90	0.85	0.92	0.95

miRdisNET provides a list of miRNAs to which it has assigned an importance score (p-value) for disease groups using its RobustRankAggreg tool. Each disease group is assigned an importance score, while miRNAs associated with the disease group are assigned the same score as the group. A part of the reported miRNA list associated with disease groups obtained with the RobustRankAggreg tool is shown in Figure 3.8.

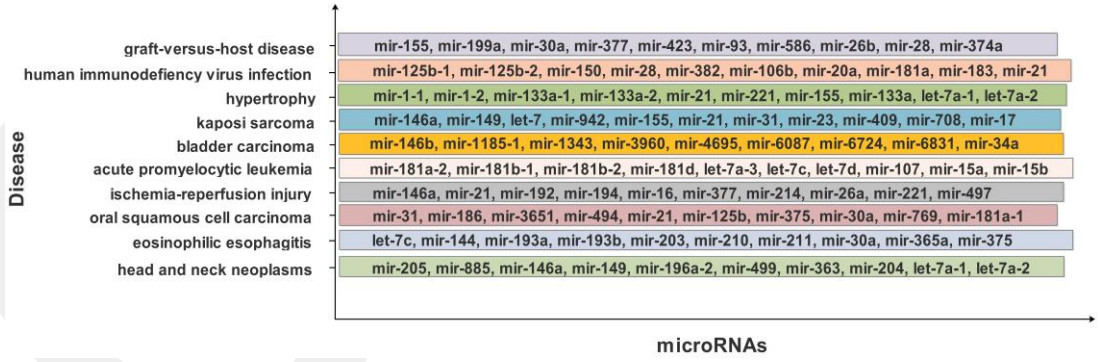


Figure 3.8 The top ranked 10 groups by RobustRankAggreg for the dataset BLCA. The name of the disease/group is shown at the y-axis and the bars denote the set of associated miRNAs.

miRdisNET assigns importance scores to miRNAs for the disease under investigation. These top ranking miRNAs can be potential biomarkers for disease under study. Table 3.3 displays the top six identified miRNAs for the BLCA dataset, and the scores of each miRNA where the score indicates the significance of the miRNA for bladder cancer. Due to inherent nature of cancers, some miRNAs are commonly identified as important miRNAs for different cancer types. For example, miRdisNET identified hsa-

let-7c, hsa-mir-128 and hsa-mir-107 as the top three significant miRNAs in BLCA dataset. The top three related miRNAs to UCEC dataset are found as hsa-let-7c, hsa-mir-128 and hsa-mir-107. The top three related miRNAs to THCA dataset are hsa-let-7c, hsa-mir-451 and hsa-mir-128. The top three related miRNAs to STAD dataset are hsa-mir-320a, hsa-mir-1 and hsa-mir-107. hsa-let-7c and hsa-mir-128 are commonly identified miRNAs for BLCA, UCEC, THCA cancer types. On the other hand, hsa-mir-320a, hsa-mir-1 are uniquely identified for STAD (Stomach Adenocarcinoma).

Table 3.3 An example of the first six ranking groups with an accuracy of miRNA groups in BLCA and an example of the first six ranking groups with accuracy of disease groups in BLCA.

Rank	miRNA	Score/Accuracy		Rank	Disease	Score/Accuracy
1	hsa-let-7c	0.96		1	graft-versus-host disease	0.96
1	hsa-mir-128	0.96		1	human immunodeficiency virus infection	0.96
1	hsa-mir-107	0.96		1	hypertrophy	0.96
2	hsa-let-7c	0.95		2	carcinoma, bladder	0.95
2	hsa-mir-429	0.95		2	leukemia, promyelocytic, acute	0.95
2	hsa-mir-320a	0.95		2	ischemia-reperfusion injury	0.95
3	hsa-let-7c	0.93		3	squamous cell carcinoma, oral	0.93
3	hsa-mir-429	0.93		4	eosinophilic esophagitis	0.93
3	hsa-mir-210	0.93		4	head and neck neoplasms	0.93
4	hsa-let-7c	0.93		4	kidney injury	0.93
4	hsa-mir-210	0.93		5	kidney neoplasms	0.91
4	hsa-mir-375	0.93		6	carcinoma, renal cell	0.91
5	hsa-mir-210	0.91		6	carcinoma, renal cell, chromophobe	0.91
6	osteosarcoma	0.91		6	osteosarcoma	0.91
6	hsa-mir-451a	0.91				
6	hsa-let-7c	0.91				

Similarly, miRdisNET assigns importance scores to disease groups. Table 3.3 shows the identified top six disease groups for BLCA dataset, and the scores of each disease where the score indicates the level of association of the identified disease group with the disease under study. For example, the top three related diseases to BLCA are Graft Versus Host Disease, Human immunodeficiency virus infection and Hypertrophy. The top three related diseases to BRCA are lung adenocarcinoma, glioblastoma and melanoma. The top three related diseases to KICH are hepatocellular carcinoma, cervical

neoplasms and lung neoplasms. The top three related diseases to KIRP are colon carcinoma, breast neoplasms and colorectal carcinoma. The top three related diseases to LUAD are endometrial adenocarcinoma, acute myocardial infarction, and acute kidney failure.

3.6 Discussion

3.6.1 Biological interpretation of results

In this section, we assess the relevance of our findings from a biological point of view. We evaluate and validate the miRNA-disease associations determined by miRdisNET using an independent database and previous studies in literature.

3.6.2 Validation of miRdisNET's findings on disease-miRNA association

Another output of miRdisNET is the list of significant miRNA groups predicted to be associated with disease groups. These miRNAs are ranked according to the p-value determined by the RobustRankAggreg method. Significant disease-miRNA groups obtained after applying miRdisNET were compared with other independent external datasets and with miRNA-disease relationships found in the literature. We utilized widely used miRNA-disease association databases (HMDD and miRCancer [59]) and some articles to comprehensively evaluate the results from a biological perspective. There are biological databases that report the functions of miRNAs and develop predictions based on experimental results or computational predictions. Although there are several databases that contain predicted associations between microRNAs and cancers using computational methods, there are only a few experimental results. However, the predictions obtained in studies evaluating miRNA function need to be verified experimentally. Even though numerous experiments have been performed to study the expression of microRNA in cancer cells, the results of the experiments are not consistent in the literature. miRCancer is a database that contains verified miRNA data based on PubMed. There are seven unique miRNAs (miR-133a, miR-218, miR-588, miR-218, miR-372, miR-448 and miR-223) in miRCancer related to LUSC. For LUSC patients, Yang et al. [90] reported the significance of 9 miRNAs (miR-30d, miR-185, miR-30a, miR-193a-3p, miR-125a, miR-101, let-7i, miR-126, and miR-15a) by using real-time

polymerase chain reaction (qRT-PCR) in their studies. In another study, Petkova et al. [91] validated 10 miRNAs (miR-144-3p, miR-4689-3p, miR-7-5p, miR-744-3p, miR-650, miR-375, miR-140-3p, miR-195-5p, miR-95-5p and miR-21-3p) related to LUSC.

We have evaluated the biological relevance of the top-10 disease-miRNA associations for LUSC dataset that were identified using miRdisNET. Appendix Table A1 presents the validated miRNA and disease groups, based on the above-mentioned external databases and support from literature. In Appendix Table A1, we show how many of the miRNAs obtained by miRdisNET are included in external databases or in scientific literature. For example, for LUSC dataset, 39 miRNAs associated with "aortic stenosis" disease were detected using the miRdisNET method. When the obtained miRNAs were compared with the literature, 5 miRNAs (hsa-miR-30a, hsa-miR-133a, hsa-miR-193a, hsamiR-21, hsa-miR-195) were previously reported as associated with LUSC.

Chapter 4

microBiomeGSM

4.1 Motivation

We present a novel approach, microBiomeGSM, to detect disease-associated taxonomic biomarkers by developing an efficient machine learning model based on the Grouping, Scoring and Modeling (G-S-M) approach. We have analyzed taxonomically transformed microbiome sequencing datasets with our proposed machine learning method. In this way, we aim to reveal the impact of the identified taxonomic biomarkers on specific diseases. To this end, our study contributes to the diagnosis and treatment of the disease under investigation. The proposed approach is applied on metagenomic datasets associated with 4 different datasets; and the taxonomic groups that have an impact on disease under study are identified. In the data preprocessing step, the MetaPhlAn tool developed by [64] is used to extract taxonomic data from microbiome sequencing data. In the first component (grouping component) of microBiomeGSM, the species identified in a sample are grouped according to the level of taxa known to be associated with them. In the second component (scoring component) of microBiomeGSM, importance scores are assigned to taxon groups using inherent machine learning techniques. The score is a predictor of how well a sample can be classified based on the abundance values of the species included in that taxon group. In the final (modeling) component of microBiomeGSM, three different outputs are generated. The first output is the performance metrics of the developed machine learning model. The second output is the list of important taxa groups associated with the disease under study, and these taxonomic features can be considered as biomarkers. The third output is the species associated with the taxa groups. Performance evaluation of microBiomeGSM is assessed separately for each disease, and for 3 different taxonomic levels (genus, family, order). Feature selection algorithms are applied to the same dataset in order to comparatively evaluate the performance of microBiomeGSM. The biological

relevance of the identified taxon groups at genus, family, order levels for different diseases is discussed with reference to existing knowledge in the literature.

4.2 microBiomeGSM

Utilizing the G-S-M approach, microBiomeGSM performs a search to identify the most important taxonomic groups in disease-associated metagenomic datasets. The relative abundance values of the species within the group can be checked for each sample; and the generated model decides whether the sample has the disease or not. By focusing on a specific taxonomic level, we can use the G component to find the most significant group for the disease under study. This approach provides the advantage of focusing on either the macroscopic or microscopic view of the most important group to distinguish between healthy samples and patient samples. An overview of the steps performed in microBiomeGSM is presented in Figure 4.1.

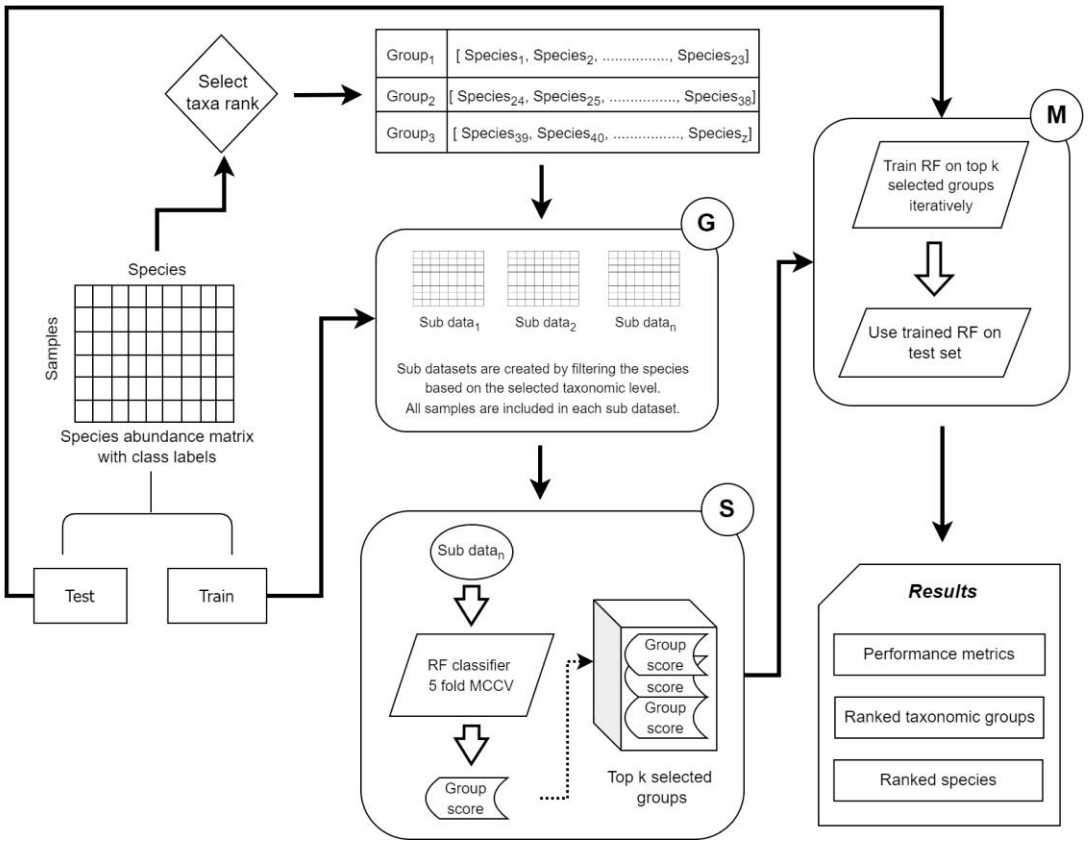


Figure 4.1 G-S-M approach in microBiomeGSM. MCCV denotes Monte Carlo Cross-Validation

Let X be the two-class dataset consisting of the species in the columns, and samples in the rows including the class labels (1 denoting the disease state and 0 denoting the

healthy state). To understand the approach in detail, let us assume that the taxonomic level is selected as "genus" for the "Select taxa rank" step in Figure 4.1. The input X_{abd} (abundance matrix) is first split into a training set (X_{train}) and a test set (X_{test}) with a ratio of 80:20 based on the class labels. Denote by S the feature space of all species in X_{abd} and by U_{genus} all unique genera for S. Grp{} denotes the selection function of each U_{genus} in S, grouping all species on the basis of similar genuses. Grp{U_{genus}^{i} for S} represents each genus in S, with all the species grouped by genus. For example, if we take Alistipes as one of the genus in U_{genus}, we get the following when we apply the Grp function.

Grp{U_{genus}^{i}}, where i = Alistipes and $\in$ S

Grp{Alistipes}= {alistipes_finegoldi, alistipes_indistinctus, alistipes_inops, alistipes_shahii}

Similarly, this approach is applied to all genuses that are present in X_{abd}, and a list of genus groups is created, as shown in Figure 4.1 after the select taxa rank step. This is repeated for the three taxonomic levels identified.

When Figure 4.1 is examined, firstly, in the grouping component G, for all the groups of genus, we partition X_{train} into sub data denoted as sub_d_x. Following the earlier example of Alistipes, this group yields sub_$d_{alistipes}$ which is created from X_{train}. The sub_$d_{alistipes}$ contains the labels of the samples, but the feature space is restricted only to species within the Alistipes genus. This is applied to all different genera created in the prior step, so we have multiple subsets of data with a feature space specified by genus. Secondly, in the scoring step S, the generated sub_d is trained on a Random Forest classifier with 5-fold cross-validation with randomized stratified shuffling. Each sub_d is given a score equal to the mean of the accuracy over all foldings based on the prediction of the labels. Each sub_d is scored and then sorted based on the score. The top k groups with the highest score are used for the subsequent step. The value chosen for k is 10, but other values for k have been tested. Following the example of selecting genus as the taxonomic level, the top 10 genus groups that show strong discriminative ability are used to build the classification model. Thirdly, in the modeling component, the species from the top 10 genus groups are used to train a Random Forest model with 100-fold Monte Carlo Cross-Validation (MCCV). The top ranking set of species corresponding to the top ranked group is trained on X_{train} and then tested on X_{test}. Then, the second set of species corresponding to the second highest scoring group is aggregated with the top scoring set of species; and then used to train and test the model. This process is repeated until all species in the top 10 ranked genus groups are aggregated; and used to train and test the

classifier. This whole process is repeated 100 times, stratifying the initial X_{abd} and randomly splitting it into X_{train} and X_{test} without replacement. The classification performance metrics are determined as the average of the metrics obtained in 100 folds. Similarly, the top ranked groups and the top ranked species are retained for each run.

4.3 Application of feature selection and classifiers using metagenomic data

In metagenomics research, it is observed that in studies using taxonomic features, the number of observations used for training data is higher than the number of observations used for testing data. This situation is undesirable if studies are to produce more effective results, and researchers are proposing various methods of resolution, particularly feature selection methods. Although the process of feature selection in disease prediction problems based on metagenome data has not been well studied, the literature suggests that this process may be as important as the choice of a classification method [32]. The process of feature selection in metagenome-based disease prediction could help us learn more about disease development mechanisms. Therefore, further research in this direction is warranted. In metagenomics studies, in order to reduce the number of taxa , i.e., to select informative species (features), min Redundancy Max Relevance (mRMR) [66], Lasso [67], Elastic Net [68], and the iterative sure select algorithm [69] have been used extensively. Another feature selection method, called Fizzy, addresses the challenge of using classification techniques to identify important functional elements for downstream analysis [64]. Oudah and Henschel presented an alternative taxonomy-based method for feature selection [70]. Bakir-Gungor et al., (2021) applied CMIM [71], FCBF [72], mRMR [66], and Select K best (SKB) [73] to type 2 diabetes-associated metagenomics datasets and obtained powerful performance metrics [74]. Jabeer et al. also proposed a robust classification method for evaluating colorectal cancer associated metagenomic datasets using a combination of feature selection methods and machine learning methods [75]. Bakir-Gungor et al., (2022) also proposed a powerful method for IBD classification with fewer features by combining feature selection methods and machine learning methods [7]. While these feature selection approaches have produced effective results in a variety of fields, they have only recently been applied to microbiome-based disease prediction problems.

In this study, we have comparatively evaluated microBiomeGSM with different classifiers and with different feature selection methods. As the feature selection methods, we have utilized Select K best (SKB), Fast Correlation Based Filter (FCBF), Extreme Gradient Boosting (XGBoost), Min Redundancy Max Relevance (mRMR), Information Gain (IG), and Conditional Mutual Information Maximization (CMIM). Wang and Liu (2020) compare the performance of classifiers with traditional methods and ensemble methods for disease prediction based on human microbiome data. They use Elastic Network and SVM as traditional methods and Random Forest and Extreme Gradient Boosting (XGBoost) as ensemble methods. In their study, they find that the XGBoost algorithm shows superior performance compared to other algorithms [92]. In another study, Marcos-Zambrano et al. (2021) conducted an important review paper to reveal the links between the microbiome and diseases. In this study, which included information on the performance of machine learning methods, they found that the Support Vector Machines (SVM), Random Forest (RF), k-Nearest Neighbors (k-NN), and Logical Regression (LR) algorithms were widely used. They concluded that when selecting a machine learning algorithm, several factors should be considered such as the set of observations, the set of features, the type of data, and the quality of the data. They suggest using several different methods, comparing them, and choosing the one that provides the best performance value [33].

4.4 Implementation of microBiomeGSM

The microBiomeGSM tool utilizes the pre-existing biological knowledge of the assignment of the species into different taxonomic levels, such as genus, family, and order. Experiments with the microBiomeGSM tool were conducted on the open-source KNIME platform [93]. This platform can handle a wide range of data types and operations. The user can configure the number of iterations, the rank function, and the number of iterations for MCCV. All rows with missing values are removed within the workflow.

4.5 microBiomeGSM Model Performance Evaluation

Accuracy, F1 score, sensitivity, specificity, and AUC were used to evaluate the predictive performance of the proposed models. AUC score is a common measure for performance evaluation and a reliable metric for evaluating balanced datasets. Other metrics such as F1 score, sensitivity, specificity, and accuracy, were used to evaluate the

performance of the created models because the dataset for this study has an uneven distribution of classes. When a balance between precision and recall is desired and there is an uneven distribution of classes, the F1 score is a good option among the performance metrics (many true negatives). Several classifiers report the probability values for their predictions, which can also be considered as confidence values for the prediction. The AUC often uses this information to figure out how often incorrect predictions occur at different confidence levels. In real life, test results from positive and negative examples overlap. AUC illustrates how the threshold or cut-off value for identifying positive examples affects the relationship between recall and precision. In this study, all of the above-mentioned metrics were calculated as the mean of 100 times MCCV. After each iteration, we obtain lists of significant taxonomic groups and species associated with these taxa groups for a given disease. To assign scores to the entities in the taxonomic groups list and in the species lists, a prioritization approach is used. For this purpose, we integrated the RobustRankAggreg algorithm [89] and microBiomeGSM. RobustRankAggreg algorithm is available as an R package. Each entity (taxonomic group or species) in the lists is given a p-value by the RobustRankAggreg technique, indicating how highly ranked that entity. Using the RobustRankAggreg tool, microBiomeGSM outputs a list of species to which it has assigned a significance value (p-value) for a specific taxonomic group. Each taxa group is assigned a significance value and the species associated with that group are assigned the same value.

4.5 Results

The main objective of this study is to identify the microbial communities that are associated with specific diseases. In order to facilitate disease diagnosis, using metagenomic data we develop an efficient classification model based on taxonomic levels. In this section we present our findings for four different datasets. Here we also present comparative evaluation results against other existing methods.

4.5.1 Comparing varying group size for microBiomeGSM

One approach to evaluate model performance in the context of microBiomeGSM is to compare model performance between different values of the parameter k. k represents the number of groups (taxa) used in microBiomeGSM models. This approach can help researchers determine the optimal value of k that balances model complexity and

predictive power, ultimately leading to more effective and interpretable models in microbiome-related research. It provides insight into how the inclusion or exclusion of specific taxa affects the overall performance of microBiomeGSM models.

Appendix Table A2 shows the performance metrics obtained with 100-fold MCCV for the aggregated top 10 groups for four different datasets compared at three different taxonomic levels (genus, family, order) for grouping. For the IBDMDB dataset, microBiomeGSM achieved an AUC of 93% using the top 1 group at the family level. Performance metrics are shown for the top 2 groups via combining species from the first and second highest scoring groups. We obtained an AUC of 97% when the top 2 groups are combined at the family taxonomic level for the IBDMDB dataset. In this way, microBiomeGSM provides cumulative performance results for the top 10 highest scoring groups. For the IBDMDB dataset, the highest performance metric (an AUC of 98%) is obtained using the species from the top 10 groups at the order taxonomic level. For the IBD dataset, the highest performance metric (an AUC of 93%) is obtained using the species from the top 9 groups at the order taxonomic level. For the T2D dataset, the highest performance metric (an AUC of %74) is obtained using the species from the top 9 groups at the order taxonomic level. For the CRC dataset, the highest performance metric (an AUC of %83) is obtained using the species from the top 10 groups at the family taxonomic level. While examining other performance metrics (such as accuracy, sensitivity, specificity in Appendix Table A2), it is noteworthy that satisfactory results are obtained with microBiomeGSM for each taxonomic level, especially for the IBDMDB dataset. The high sensitivity values that are reported for the CRC, IBDMDB, and IBD datasets display the success of the microBiomeGSM tool in terms of detecting the patient samples. In the CRC, IBDMDB, and IBD datasets, the strikingly high specificity values indicate that the microBiomeGSM tool correctly identifies the negative samples (i.e., individuals who do not have the disease). However, in the T2D dataset, the specificity rate appears to be relatively low compared to the other datasets. Nevertheless, the ability to detect negative samples remains at a reasonable level.

In addition, Figure 4.2 and Figure 4.3 show the sensitivity and specificity values obtained with the microBiomeGSM tool for all datasets. Figure 4.2 shows the sensitivity values obtained using the microBiomeGSM tool across all datasets. One can notice from Figure 4.2 (A) that for the CRC data set the highest sensitivity value (73%) is obtained for the order taxon level using 10 cumulative groups. In particular, the sensitivity values calculated for the IBDMDB dataset were quite impressive, especially in group 1 and

group 6, both at the family taxon level, reaching 99% sensitivity value, as shown in Figure 4.2 (B). Figure 4.2 (C) shows another impressive set of results for the IBD data set. In Figure 2 (C), we observe high values for sensitivity, in particular 87% sensitivity at the taxon level in group 1. As shown in Figure 4.2 (D), the highest sensitivity value for the T2D data set is 69%. This result is obtained for the genus taxon level using 10 cumulative groups. A sensitivity value of 69% is also obtained for the family taxon level using 4 cumulative groups.

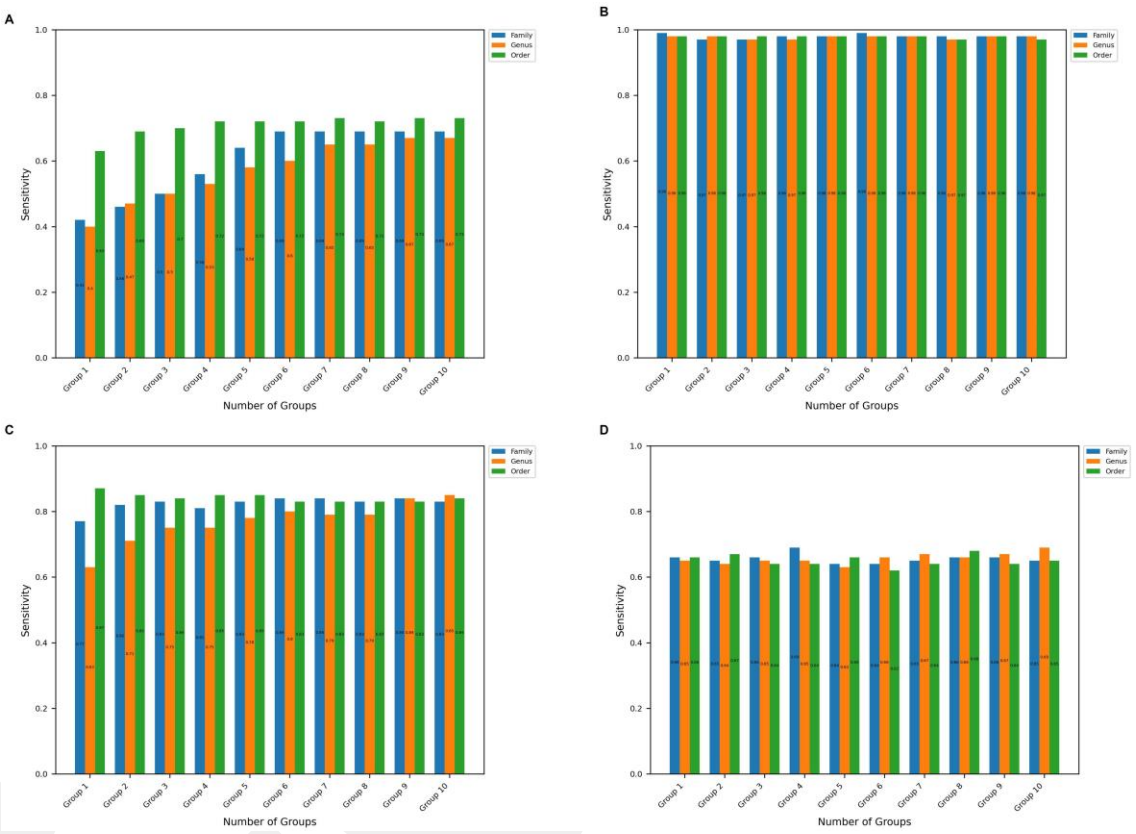


Figure 4.2 Sensitivity values obtained at the family, order, and genus taxon levels for the top 10 significant groups across all 4 datasets. (A–D) Represents the results obtained in CRC, IBDMDB, IBD, T2D datasets, respectively.

Figure 4.3 shows the specificity values obtained using the microBiomeGSM tool for all datasets. As shown in Figure 4.3 (A), the specificity value obtained for the CRC dataset is remarkable, reaching an impressive specificity value of 94% at the family taxon level for 1 group. Figure 4.3 (B) depicts that the highest specificity value obtained for the IBDMDB dataset is 93% for 1 group at the order taxon level. As displayed in Figure 4.3 (C), the highest specificity value obtained for the IBD dataset is 85% for the 4 cumulative groups at the order taxon level. The same result is also obtained at the order taxon level for the 5 cumulative groups. One can notice in Figure 4.3 (D) that the highest specificity

value that is obtained for the T2D dataset is 71% for the 6 cumulative groups at the order taxon level.

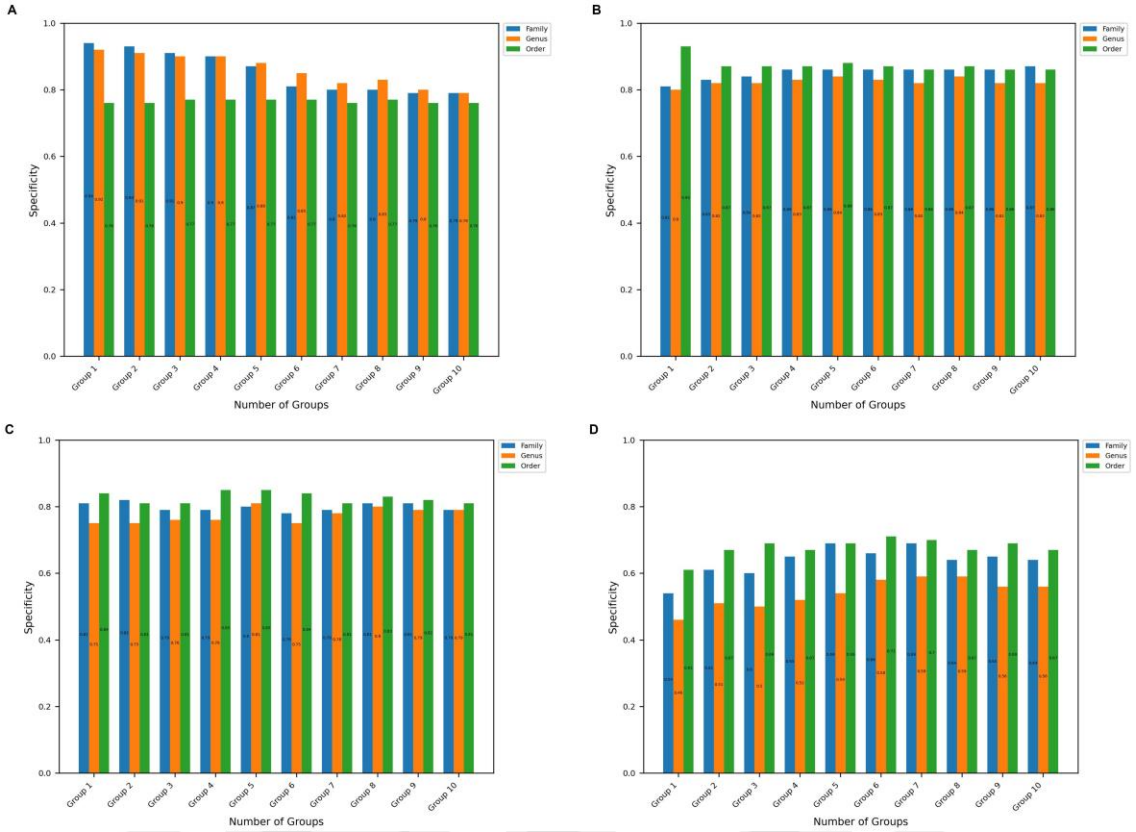


Figure 4.3. Specificity values at the order, genus, and family taxon level for the top 10 significant groups for all 4 disease datasets. (A–D) Represents the results obtained in CRC, IBDMDB, IBD, T2D datasets, respectively.

The number of significant groups used to train the model could affect the performance of microBiomeGSM. Table 4.1 shows the influence of the number of groups and the number of species at family, genus and order levels on four datasets. Table 4.1 presents the performance of the top 10 cumulative groups and top 1 group for each taxonomic level on different tested datasets. For the IBDMDB dataset, for the family taxonomic level, one can observe that the AUC increases by 5% when we consider the top 10 significant groups cumulatively, while we increase the number of species from 34 to 205. On the same dataset, an increase of 8% in AUC score is observed at the Genus taxonomic level via increasing the number of species from 34 to 119. For the same dataset, a decrease of 1% is observed at the Order taxonomic level. Order taxonomic level using the top group that includes 98 species achieves the highest AUC success rate of 98% for the IBDMDB dataset.

Table 4.1 The effect of the number of groups that are generated at different taxonomic levels on performance metrics for all datasets. Sen is the sensitivity, Spe is the specificity, AUC is the Area Under the Curve.

CRC								
Taxonomic hierarchy	**# of Groups**	**Average # of Species**	**Accuracy**	**Sen**	**Spe**	**F measure**	**AUC**	**Precision**
Family	10	239.11	0.78	0.84	0.72	0.79	0.84	0.76
Family	1	16.17	0.67	0.91	0.43	0.73	0.69	0.62
Genus	10	102.06	0.77	0.82	0.71	0.78	0.84	0.76
Genus	1	7.16	0.66	0.88	0.44	0.72	0.71	0.63
Order	10	604	0.82	0.86	0.78	0.82	**0.87**	0.81
Order	1	154.25	0.76	0.82	0.71	0.78	0.81	0.76
IBDMDB								
Taxonomic hierarchy	**# of Groups**	**Average # of Species**	**Accuracy**	**Sen**	**Spe**	**F measure**	**AUC**	**Precision**
Family	10	205.76	0.95	0.98	0.87	0.95	0.97	0.93
Family	1	34	0.93	0.98	0.81	0.94	0.93	0.9
Genus	10	119.4	0.92	0.98	0.82	0.95	0.97	0.93
Genus	1	34	0.92	0.98	0.80	0.94	0.91	0.91
Order	10	341.22	0.93	0.97	0.86	0.95	**0.98**	0.93
Order	1	98	0.96	0.98	0.93	0.97	0.98	0.95
IBD								
Taxonomic hierarchy	**# of Groups**	**Average # of Species**	**Accuracy**	**Sen**	**Spe**	**F measure**	**AUC**	**Precision**
Family	10	260.24	0.82	0.85	0.79	0.82	**0.91**	0.81
Family	1	51.59	0.78	0.78	0.78	0.78	0.86	0.78
Genus	10	121.78	0.81	0.83	0.79	0.81	0.88	0.8
Genus	1	12.26	0.7	0.67	0.74	0.69	0.78	0.73
Order	10	608.27	0.82	0.82	0.81	0.82	0.9	0.82
Order	1	174.86	0.81	0.82	0.8	0.81	0.9	0.81
T2D								
Taxonomic hierarchy	**# of Groups**	**Average # of Species**	**Accuracy**	**Sen**	**Spe**	**F measure**	**AUC**	**Precision**
Family	10	321.16	0.65	0.68	0.63	0.66	0.71	0.65
Family	1	39.86	0.59	0.71	0.47	0.63	0.63	0.58
Genus	10	129.8	0.64	0.64	0.64	0.64	0.69	0.65
Genus	1	15.94	0.56	0.62	0.49	0.58	0.58	0.55
Order	10	596.99	0.65	0.65	0.64	0.64	**0.72**	0.65
Order	1	138.28	0.59	0.67	0.52	0.62	0.64	0.59

Similarly, family taxonomic level using the top 10 combined groups achieves 97% AUC on the IBDMDB dataset, but these 10 combined groups include a much higher number of species (205 species). For the IBD dataset, the highest AUC value of 91% was obtained using the microBiomeGSM tool. This value at the family taxonomic level was obtained by cumulatively combining 10 groups, using an average of 260.4 species. For the T2D dataset, the highest AUC value of 72% was obtained using the microBiomeGSM tool. This value, obtained at the order taxanomic level, was obtained by combining 10 groups cumulatively. For 1 group, an average of 138.28 species are used at the taxonomic level, while for 10 groups, an average of 596.99 species are used. For the CRC dataset, the highest AUC value of 87% was obtained using the microBiomeGSM tool. This value at the order taxanomic level was obtained by cumulatively combining 10 groups, using an average of 604 species.

microBiomeGSM reports important groups of features that are detected at different taxonomic levels for the disease under study. Table 4.2 lists the top 10 important groups that are identified by microBiomeGSM for three different taxonomic levels on four different datasets. The identified features are ranked by their importance scores from high to low. The feature with the highest importance value is the strongest candidate to be announced as potential taxonomic biomarker for the disease under investigation.

The microBiomeGSM tool lists a number of associated species for each identified group. The species included in the top 5 significant groups are listed in Appendix Table A3, A4, A5 for family, order, and genus taxonomic levels, respectively for four different datasets.

For the IBDMDB dataset, the changes in the AUC score when the number of groups is increased from 1 to 10 are shown in Appendix Figure A1. For the IBDMDB dataset, a high AUC score is obtained at the order taxonomic level. When the number of groups was increased, the AUC score decreased relatively, and no significant change was observed after 5 groups. At the genus and family taxonomic levels, there is a significant increase in the AUC score until 5 groups are combined and no significant change after 5 groups.

Table 4.2. Top 10 groups identified by microBiomeGSM for different taxonomic levels, applied on all microbiome datasets.

Rank	Family	Order	Genus
CRC			
1	PEPTOSTREPTOCOCCACEAE	CLOSTRIDIALES	PARVIMONAS
2	PEPTONIPHILACEAE	TISSIERELLALES	PEPTOSTREPTOCOCCUS
3	FUSOBACTERIACEAE	BACTEROIDALES	FUSOBACTERIUM
4	BACILLALES UNCLASSIFIED	FUSOBACTERIALES	GEMELLA
5	VEILLONELLACEAE	BACILLALES	DIALISTER
6	LACHNOSPIRACEAE	VEILLONELLALES	LACHNOCLOSTRIDIUM
7	ERYSIPELOTRICHACEAE	ERYSIPELOTRICHALES	PREVOTELLA
8	RUMINOCOCCACEAE	LACTOBACILLALES	STREPTOCOCCUS
9	PREVOTELLACEAE	ACTINOMYCETALES	PORPHYROMONAS
10	STREPTOCOCCACEAE	DESULFOVIBRIONALES	SOLOBACTERIUM
IBDMDB			
1	BACTEROIDACEAE	BACTEROIDALES	BACTEROIDES
2	LACHNOSPIRACEAE	CLOSTRIDIALES	ALISTIPES
3	RUMINOCOCCACEAE	FIRMICUTES UNCLASSIFIED	EUBACTERIUM
4	RIKENELLACEAE	VEILLONELLALES	ROSEBURIA
5	FIRMICUTES UNCLASSIFIED	BURKHOLDERIALES	FIRMICUTES UNCLASSIFIED
6	TANNERELLACEAE	METHANOMASSILIICOCCALES	PARABACTEROIDES
7	EUBACTERIACEAE	DESULFOVIBRIONALES	RUMINOCOCCUS
8	CLOSTRIDIACEAE	ERYSIPELOTRICHALES	COPROCOCCUS
9	VEILLONELLACEAE	BIFIDOBACTERIALES	BLAUTIA
10	ODORIBACTERACEAE	EGGERTHELLALES	CLOSTRIDIUM
IBD			
1	LACHNOSPIRACEAE	CLOSTRIDIALES	BLAUTIA
2	BIFIDOBACTERIACEAE	CORIOBACTERIALES	BIFIDOBACTERIUM
3	CORIOBACTERIACEAE	BIFIDOBACTERIALES	EUBACTERIUM
4	RUMINOCOCCACEAE	ERYSIPELOTRICHALES	DOREA
5	ERYSIPELOTRICHACEAE	BACTEROIDALES	COLLINSELLA
6	CLOSTRIDIALES FAMILY XIII INCERTAE SEDIS	LACTOBACILLALES	PEPTOSTREPTOCOCCUS
7	EUBACTERIACEAE	SELENOMONADALES	COPROCOCCUS
8	PEPTOSTREPTOCOCCACEAE	VERRUCOMICROBIALES	ERYSIPELOTRICHACEAE NONAME
9	CARNOBACTERIACEAE	CANDIDATUS SACCHARIBACTERIA NONAME	LACHNOSPIRACEAE NONAME
10	CLOSTRIDIACEAE	BACILLALES	BACTEROIDES
T2D			
1	LACHNOSPIRACEAE	CLOSTRIDIALES	EUBACTERIUM
2	BIFIDOBACTERIACEAE	BIFIDOBACTERIALES	BIFIDOBACTERIUM
3	RUMINOCOCCACEAE	CORIOBACTERIALES	BLAUTIA
4	EUBACTERIACEAE	BACTEROIDALES	DOREA
5	CORIOBACTERIACEAE	LACTOBACILLALES	LACHNOSPIRACEAE NONAME
6	CLOSTRIDIALES FAMILY XIII INCERTAE SEDIS	ERYSIPELOTRICHALES	RUMINOCOCCUS
7	ERYSIPELOTRICHACEAE	SELENOMONADALES	COPROCOCCUS
8	PEPTOSTREPTOCOCCACEAE	VERRUCOMICROBIALES	PEPTOSTREPTOCOCCUS
9	CARNOBACTERIACEAE	METHANOBACTERIALES	ERYSIPELOTRICHACEAE_NONAME
10	BACTEROIDACEAE	BACILLALES	GRANULICATELLA

4.5.2 Comparing against traditional machine learning methods

Our Grouping-Scoring-Modeling (G-S-M) approach emerges as a paradigm shift from traditional feature selection methods. Instead of pinpointing individual informative features, the GSM methodology groups these features. These groups are then scored, and a classification model is built using these top-ranking feature conglomerates. The versatility of the GSM method, as detailed in our prior work (Yousef, Kumar and Bakir-Gungor, 2021), lies in its adaptability. Groups can be created either by computational/statistical methods or by using domain-specific knowledge. In order to use the GSM strategy for a given dataset, a deep domain expertise is required to skillfully define these groups, which makes each application different. The modifications required to tailor the G-S-M approach to the unique needs of microbiome research highlight the adaptability of the G-S-M method and the novelty of our current study.

We have comparatively evaluated the performance of microBiomeGSM against 4 different classifiers and 6 different feature selection methods using the same datasets. All algorithms are run with default parameters. The developed approach and feature selection methods were executed multiple times, and the results were averaged and shared. Table 4.3 shows the performance of the different feature selection algorithms and different classifiers on the same disease associated microbiome datasets. In these experiments, the number of features was set to 100. The best result for the IBDMDB dataset is obtained by using the XGBoost feature selection algorithm in combination with the Random Forest classification algorithm with 98% AUC. For the CRC dataset, the best result is obtained by using the XGBoost feature selection algorithm in combination with the Random Forest classification algorithm with an AUC of 85%. For the IBD dataset, the best result is obtained using the Random Forest classification algorithm with 92% AUC and the SKB feature selection algorithm. For the T2D dataset, the best result is obtained by using the XGBoost feature selection algorithm in combination with the Random Forest classification algorithm with 70% AUC. We would like to note that the primary objective of microBiomeGSM is not to compete with other feature selection methods (FS). Even if microBiomeGSM's performance is on par with or slightly less favorable than other FS methods, its fundamental contribution lies in identifying the most informative microbiomes. These microbiomes play a pivotal role in aiding researchers in gaining a deeper understanding of the biological underpinnings of the disease under investigation. In essence, microBiomeGSM's value lies in its ability to contribute to the advancement

of biological knowledge, rather than merely outperforming other feature selection techniques.

Table 4.3 Area under the curve (AUC) results obtained using 100 features for different feature selection methods and classifiers for all datasets.

CRC						
Model	**SKB**	**IG**	**XGB**	**FCBF**	**MRMR**	**CMIM**
Adaboost	0.75 ± 0.02	0.71 ± 0.05	0.78 ± 0.04	0.71 ± 0.05	0.63 ± 0.06	0.77 ± 0.04
DT	0.67 ± 0.04	0.64 ± 0.04	0.69 ± 0.04	0.63 ± 0.06	0.61 ± 0.04	0.65 ± 0.05
Logitboost	0.76 ± 0.04	0.72 ± 0.05	0.78 ± 0.06	0.70 ± 0.04	0.64 ± 0.06	0.76 ± 0.05
RF	0.82 ± 0.03	0.79 ± 0.04	**0.85 ± 0.03**	0.77 ± 0.05	0.74 ± 0.04	0.80 ± 0.03
IBDMDB						
Model	**SKB**	**IG**	**XGB**	**FCBF**	**MRMR**	**CMIM**
Adaboost	0.89 ± 0.04	0.90 ± 0.03	0.89 ± 0.06	0.49 ± 0.08	0.51 ± 0.08	0.51 ± 0.08
DT	0.83 ± 0.03	0.82 ± 0.04	0.84 ± 0.03	0.46 ± 0.07	0.50 ± 0.07	0.50 ± 0.06
Logitboost	0.89 ± 0.04	0.91 ± 0.03	0.86 ± 0.06	0.50 ± 0.06	0.51 ± 0.08	0.49 ± 0.08
RF	0.96 ± 0.01	**0.96 ± 0.01**	**0.98 ± 0.01**	0.46 ± 0.1	0.54 ± 0.08	0.52 ± 0.07
IBD						
Model	**SKB**	**IG**	**XGB**	**FCBF**	**MRMR**	**CMIM**
Adaboost	0,90 ± 0.07	0,89 ± 0.03	0,91 ± 0.03	0,51 ± 0.06	0,51 ± 0.03	0,66 ± 0.08
DT	0,78 ± 0.08	0,70 ± 0.08	0,73 ± 0.07	0,53 ± 0.08	0,51 ± 0.04	0,56 ± 0.09
Logitboost	0,90 ± 0.04	0,90 ± 0.05	0,92 ± 0.05	0,55 ± 0.1	0,53 ± 0.05	0,59 ± 0.1
RF	**0,92± 0.03**	0,88 ± 0.06	0,91 ± 0.04	0,53 ± 0.09	0,55 ± 0.07	0,63 ± 0.11
T2D						
Model	**SKB**	**IG**	**XGB**	**FCBF**	**MRMR**	**CMIM**
Adaboost	0,56 ± 0.12	0,60 ± 0.05	0,64 ± 0.07	0,50 ± 0.10	0,5 ± 0.01	0,50 ± 0.12
DT	0,52 ± 0.08	0,52 ± 0.08	0,53 ± 0.05	0,41 ± 0.10	0,51 ± 0.02	0,49 ± 0.10
Logitboost	0,55 ± 0.10	0,58 ± 0.09	0,62 ± 0.10	0,48 ± 0.08	0,50 ± 0.01	0,51 ± 0.11
RF	0,62 ± 0.11	0,62 ± 0.07	**0,70 ± 0.06**	0,49 ± 0.08	0,51 ± 0.03	0,54 ± 0.10

Table 4.4 shows the performance metrics of microBiomeGSM for each taxonomic level for four different datasets. The # of species column shows the number of species (features/variables) used to train and test the model. Since the number of species changes in each iteration of MCCV, we also report the standard deviation. Performance metrics are reported as the average of 100 iterations with the corresponding standard deviation. For the CRC dataset, among different classifiers the RF algorithm has the highest performance for all calculated metrics including the accuracy, sensitivity, specificity, precision, and AUC metric. The AdaBoost, LogitBoost and DT models show lower performance compared to the RF model. The performance metrics of these three algorithms are similar but not as high as RF model. At the order taxonomic level, the mean values of the performance metrics are stable, and the standard deviations are low.

This indicates that the order level is a more appropriate choice for CRC classification. Comparing the RF model and the microBiomeGSM model, similar performance metrics are obtained for the CRC dataset, but it is worth mentioning that the number of features used in the proposed tool is lower. In other words, for the CRC dataset the microBiomeGSM model can accurately classify using fewer taxonomic features. For the IBDMDB dataset, among different classifiers the RF algorithm has the highest accuracy, sensitivity, specificity, precision, and AUC values. In particular, RF model achieved very high sensitivity and AUC values. For the IBDMDB dataset, the microBiomeGSM tool achieves an AUC of 98% for the order taxon level, the same performance metrics as obtained by the RF classification algorithm. However, the microBiomeGSM tool uses 341 features for the order taxon level, while the RF model uses 579 features. For IBD dataset, the RF algorithm generates the highest performance on several metrics, including accuracy, sensitivity, specificity, precision, and AUC. It performs particularly well on sensitivity and AUC. In our analysis, microBiomeGSM achieved an impressive AUC value of 91% at the family taxon level. Equally remarkable is the similar performance of the RF classification algorithm (an AUC of 92%) for the same task. However, it is important to highlight an important difference between these two approaches. For IBD dataset the RF classification algorithm achieved an AUC of 92% by using a much larger set of features (1456 features) for the classification task. For the same dataset, the microBiomeGSM tool also showed remarkable performance (an AUC value of 91%). In stark contrast, microBiomeGSM achieved nearly equivalent AUC performance while using a much smaller set of features, only 260 features. This divergence in feature usage highlights the effectiveness and potential advantages of the microBiomeGSM tool in extracting meaningful information from microbiome data while optimizing computational resources. For T2D dataset, the RF classification algorithm outperforms other classification algorithms on several performance metrics including accuracy, sensitivity, specificity, precision and AUC. microBiomeGSM achieved an AUC value of 72% at the order taxon level. Interestingly, a similar level of performance is observed using the RF classification algorithm, which achieves an AUC value of 75%. However, it is important to note that the underlying mechanisms of these two methods are very different.

Table 4.4 Evaluation metrics obtained with microBiomeGSM on four datasets for different taxonomic levels, compared with traditional classifiers using all features.

CRC						
Model	**# of Species**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
AdaBoost	912	0.72 ± 0.06	0.79 ± 0.09	0.66 ± 0.17	0.7 ± 0.09	0.78 ± 0.04
DT	912	0.68 ± 0.09	0.75 ± 0.12	0.62 ± 0.26	0.66 ± 0.09	0.7 ± 0.04
LogitBoost	912	0.73 ± 0.06	0.78 ± 0.09	0.68 ± 0.18	0.71 ± 0.09	0.78 ± 0.04
RF	912	0.78 ± 0.05	0.82 ± 0.08	0.75 ± 0.14	0.76 ± 0.09	0.86 ± 0.03
microBiomeGSM: Family	292.88 ± 16.09	0.74 ± 0.65	0.7 ± 0.39	0.77 ± 0.91	0.75 ± 0.83	0.81 ± 0.67
microBiomeGSM: Genus	161.21 ± 5.17	0.74 ± 0.67	0.69 ± 0.41	0.79 ± 0.92	0.76 ± 0.84	0.8 ± 0.68
microBiomeGSM: Order	607.5 ± 188.32	0.73 ± 0.69	0.72 ± 0.66	0.75 ± 0.73	0.74 ± 0.71	0.81 ± 0.77
IBDMDB						
Model	**# of Species**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
AdaBoost	579	0.92 ± 0.02	0.97 ± 0.02	0.79 ± 0.1	0.93 ± 0.03	0.94 ± 0.01
DT	579	0.91 ± 0.02	0.94 ± 0.01	0.84 ± 0.05	0.94 ± 0.02	0.89 ± 0.02
LogitBoost	579	0.92 ± 0.01	0.98 ± 0.01	0.76 ± 0.07	0.92 ± 0.02	0.91 ± 0.04
RF	579	0.98 ± 0.01	1 ± 0	0.93 ± 0.06	0.98 ± 0.02	0.98 ± 0.01
microBiomeGSM: Family	205.76 ± 16.23	0.94 ± 0.02	0.98 ± 0.01	0.86 ± 0.05	0.93 ± 0.05	0.97 ± 0.02
microBiomeGSM: Genus	119.4 ± 15.87	0.93 ± 0.02	0.98 ± 0.01	0.85 ± 0.05	0.93 ± 0.05	0.97 ± 0.02
microBiomeGSM: Order	341.22 ± 15.6	0.93 ± 0.02	0.97 ± 0.02	0.86 ± 0.06	0.93 ± 0.06	0.98 ± 0.03
IBD						
Model	**# of Species**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
AdaBoost	1456	0.88 ± 0.04	0.85 ± 0.12	0.89 ± 0.05	0.84 ± 0.05	0.9 ± 0.04
DT	1456	0.75 ± 0.05	0.72 ± 0.09	0.78 ± 0.06	0.67 ± 0.08	0.75 ± 0.06
LogitBoost	1456	0.85 ± 0.04	0.81 ± 0.1	0.87 ± 0.07	0.8 ± 0.09	0.88 ± 0.04
RF	1456	0.87 ± 0.05	0.91 ± 0.1	0.84 ± 0.05	0.78 ± 0.06	0.92 ± 0.05
microBiomeGSM: Family	260.24 ± 26.92	0.82 ± 0.06	0.85 ± 0.07	0.79 ± 0.1	0.81 ± 0.13	0.91 ± 0.07
microBiomeGSM: Genus	121.78 ± 27.83	0.81 ± 0.06	0.83 ± 0.06	0.79 ± 0.1	0.8 ± 0.12	0.88 ± 0.08
microBiomeGSM: Order	608.27 ± 24.22	0.82 ± 0.07	0.82 ± 0.08	0.81 ± 0.09	0.82 ± 0.15	0.9 ± 0.08
T2D						
Model	**# of Species**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
AdaBoost	1456	0.68 ± 0.08	0.91 ± 0.08	0.39 ± 0.26	0.67 ± 0.09	0.66 ± 0.1
DT	1456	0.57 ± 0.05	0.98 ± 0.06	0.06 ± 0.19	0.57 ± 0.06	0.57 ± 0.09
LogitBoost	1456	0.67 ± 0.08	0.93 ± 0.08	0.36 ± 0.24	0.65 ± 0.08	0.65 ± 0.1
RF	1456	0.72 ± 0.09	0.91 ± 0.09	0.48 ± 0.29	0.71 ± 0.12	0.75 ± 0.1
microBiomeGSM: Family	321.16 ± 36.31	0.65 ± 0.08	0.68 ± 0.09	0.63 ± 0.11	0.65 ± 0.15	0.71 ± 0.08
microBiomeGSM: Genus	129.8 ± 35.03	0.64 ± 0.09	0.64 ± 0.1	0.64 ± 0.13	0.65 ± 0.18	0.69 ± 0.09
microBiomeGSM: Order	596.99 ± 35.14	0.65 ± 0.08	0.65 ± 0.09	0.64 ± 0.12	0.65 ± 0.17	0.72 ± 0.09

The RF classification algorithm achieves this AUC value by incorporating a much larger set of features, 1456 features, into its classification process. In contrast, the

microBiomeGSM tool achieves comparable AUC metric by using a leaner set of 596 features. This difference in feature usage is worth highlighting as it shows that the microBiomeGSM tool is able to deliver competitive results with a lower computational load, making it an efficient and resource-efficient choice for the classification task at hand. These results highlight the nuanced trade-offs in selecting the appropriate tool or algorithm for the specific data analysis requirements.

As shown in Table 4.5, the performance of our proposed method varies depending on the taxonomic level considered. For the order taxonomic level, for all tested datasets, the proposed method outperforms other models in terms of the AUC score, except for the RF classifier. Similarly, for all datasets, at the family and genus taxonomic levels, the AUC values are also highly competitive, outperforming those of the other four machine learning algorithms used in this study, with the sole exception of the RF classifier. These results highlight the robust performance of our method across different taxonomic levels. A remarkable performance of our proposed method was observed when it is applied on the IBDMDB dataset. Here, we obtained an exceptionally high AUC value of 0.98 ± 0.03 at the order taxonomic level using a 100-fold MCCV approach. This remarkable result demonstrates the exceptional performance and the potential of the microBiomeGSM tool.

Table 4.5 Comparative performance evaluation of microBiomeGSM and other machine learning approaches for different microbiome datasets.

Dataset		AdaBoost	DT	LogitBoost	RF	microbiome GSM: Family	microbiome GSM: Genus	microBiome GSM: Order
CRC	AUC	0.78 ± 0.14	0.70 ± 0.04	0.78 ± 0.04	**0.86** ± 0.03	0.81 ± 0.67	0.80 ± 0.68	**0.81** ± 0.77
	# of Species	912	912	912	912	292.88 ± 16.09	161.21 ± 5.17	607.5 ± 188.32
IBDMDB	AUC	0.94 ± 0.01	0.89 ± 0.02	0.91 ± 0.04	**0.98** ± 0.01	0.97 ± 0.02	0.97 ± 0.03	**0.98 ± 0.03**
	# of Species	579	579	579	579	205.76 ± 16.23	119.4 ± 15.87	341.22 ± 15.6
IBD	AUC	0.9 ± 0.04	0.75 ± 0.06	0.88 ± 0.04	**0.92** ± 0.05	**0.91** ± 0.07	0.88 ± 0.08	0.9 ± 0.08
	# of Species	1456	1456	1456	1456	260.24 ± 26.92	121.78 ± 27.83	608.27 ± 24.22
T2D	AUC	0.66 ± 0.1	0.57 ± 0.09	0.65 ± 0.1	**0.75** ± 0.1	0.71 ± 0.08	0.69 ± 0.09	**0.72** ± 0.09
	# of Species	1456	1456	1456	1456	321.16 ± 36.31	129.8 ± 35.03	596.99 ± 35.14

4.6 Discussion

The microbiome is considered as a crucial component of the human body and it is increasingly associated with numerous aspects of development and health. There is growing evidence that the microbiota is essential for understanding, diagnosing, and treating human diseases. In particular, alterations in the gut microbiome community have been linked to a variety of diseases, including CRC [94], T2D [95] and IBD [96]. Several research efforts relied on sample-level feature abundance data to identify predictive microbiome biomarkers using machine learning. In this study, we proposed to perform more effective disease classification and prediction with fewer features. To this end, we developed microBiomeGSM to solve this problem compared to tools that perform predictions with a large amount of data. The success of microBiomeGSM can be explained with the following features of the G-S-M approach:

- For the grouping component of microBiomeGSM, only the features at the similar taxonomic levels are considered.
- microBiomeGSM uses efficient classifiers for the scoring component to identify the key groups for each taxonomic level.
- For the modeling component, significant taxonomic groups are considered cumulatively using effective classifiers.

Via analyzing metagenomic data, this study aims to solve the problem of disease diagnosis using existing taxonomic knowledge; and finally introduces a tool called microBiomeGSM. The proposed tool is based on the G-S-M (Grouping-Scoring-Modeling) approach and uses species-level information by grouping taxonomic features at different taxonomic levels such as genus, family, and order. The performance of microBiomeGSM on four different disease-associated metagenomic datasets was evaluated in comparison to other feature selection methods such as Fast Correlation Based Filter (FCBF), Select Best K (SKB), Extreme Gradient Boosting (XGB), Conditional Mutual Information Maximization (CMIM), Maximum Likelihood and Minimum Redundancy (MRMR), and Information Gain (IG).

The presented microBiomeGSM approach offers several advantages in the field of disease diagnosis via analyzing metagenomic datasets. One significant benefit is its ability to efficiently identify disease-associated taxonomic biomarkers through a robust machine learning model based on the Grouping, Scoring, and Modeling (G-S-M) methodology. Differently from existing approaches, microBiomeGSM identifies groups

of important taxons and detects important species within that taxon for the disease under study. Hence, this innovative approach enables the extraction of valuable insights from microbiome data, shedding light on the influence of specific taxonomic biomarkers on the disease under investigation. Furthermore, the performance evaluation across different diseases, different taxonomic levels (genus, family, order); and the comparative assessment with different feature selection algorithms exhibits the reliability of microBiomeGSM. Finally, the discussions on the biological relevance of the findings of the proposed approach, via drawing evidence from the existing literature, provide valuable context for the identified taxon groups for the disease under study, making microBiomeGSM an informative tool in disease research. Our tool's significance transcends its mere application; it holds the potential for pioneering discoveries. It is geared to discern not isolated microbial entities but entire assemblages of species, paving the way for profound biological interpretations. By spotlighting groups of bacteria and viruses in lieu of singular entities, our tool offers a holistic view, potentially identifying microbial communities implicated in specific diseases.

With this study, we would also like to motivate biologists and the microbiome community to redesign their grouping methods instead of using individual feature selection approaches. We envision that in the future, various biological datasets, including multi-omics, will be used to redefine the groupings. Such innovative grouping strategies, complemented by modeling, promise to provide profound insights into the molecular mechanisms of diseases and the role of microorganisms in disease development.

4.6.1 Biological interpretations of microBiomeGSM's findings

This section discusses the biological relevance of the features discovered by microBiomeGSM at different taxonomic levels for all tested datasets. T2D is a metabolic disease characterized by high glucose levels in blood and caused primarily by cellular resistance to the activity of insulin [97]. There are several studies in the literature that have demonstrated the relation of different microorganisms at the genus, family, and order levels with T2D development. For the T2D dataset, the top 10 microbiomes identified by our method at the genus, family, order levels and the relevant literature can be summarized in Appendix Table A6. On the other hand, inflammatory bowel diseases (IBDs), which include primarily ulcerative colitis and Crohn's disease, but also non-infectious inflammation of the bowel, have puzzled gastroenterologists and immunologists alike since their first modern descriptions around some 75-100 years ago

[98] [7]. For the IBDMDB dataset, the top 10 microbiomes identified by our method at the genus, family, and order levels and the relevant literature can be summarized in Appendix Table A6. CRC is a prevalent malignancy affecting the colon and rectum. It constitutes approximately 10% of all newly diagnosed cancer cases worldwide [99]. For the CRC dataset, the top 10 microbiomes identified by our method at the genus, family, and order levels and the relevant literature can be summarized in Appendix Table A6.

Numerous studies have investigated the relationship between microbiomes and diseases like T2D, CRC, and IBD using similar datasets as used within this study. Upon examination of these studies, it becomes evident that while their experimental designs may vary, they consistently yield comparable results when it comes to identifying microbiomes linked to these diseases. These findings align with the important microbiomes identified by microBiomeGSM for T2D, CRC, and IBD, showcasing the tool's effectiveness in accurately identifying relevant microbiomes associated with these diseases. These congruent findings reinforce the reliability and validity of the microbiome associations detected by the microBiomeGSM tool. It also underscores the tool's capacity to identify microbiomes that are consistently linked to specific diseases, providing valuable insights for disease characterization and prediction. Hassouneh et al. [100] conducted a series of experiments aimed at uncovering microbiomes associated with IBD. In their analysis using the same dataset as used by the microBiomeGSM tool, they observed differences in Clostridium microbiota among IBD patients. Additionally, another microbiome identified for IBD in their study is Ruminococcus. Remarkably, these microbiomes align with the important microbiomes detected for the IBD disease by the microBiomeGSM tool. This correspondence in findings highlights the capacity of microBiomeGSM in identifying relevant microbiomes linked to IBD. Zhang et al. [101] conducted a study with the goal of identifying disease-associated microbiome species for Inflammatory Bowel Disease Microbiome Database (IBDMDB), employing the same dataset (PRJNA289734) as used in microBiomeGSM. In their research, they highlighted the significance of the Bacteroides microbiome. Interestingly, the Bacteroides microbiome is also identified as one of the important microbiomes by the microBiomeGSM tool proposed in our study. This alignment in findings underscores the effectiveness of microBiomeGSM in recognizing key microbiomes associated with diseases like IBD. Bai et al. [102] conducted a series of experiments aimed at identifying microbiomes associated with T2D. In their research, they utilized the SRA4565 data for T2D and highlighted the significance of the methanobacteriales microbiome. Notably,

methanobacteriales is among the top 10 microbiomes identified by the proposed microBiomeGSM tool. This convergence of findings underscores the effectiveness and utility of the proposed tool in uncovering microbiome associations with diseases like T2D. Forslund et al. [103] conducted experiments utilizing the same T2D dataset employed by microBiomeGSM to investigate microbiomes associated with T2D. Upon close examination of their experiments, they underscored the significance of the Clostridiales microbiome in relation to T2D disease. Interestingly, Clostridiales also emerges as one of the important microbiomes identified by microBiomeGSM. This convergence in findings highlights the relevance and effectiveness of microBiomeGSM in identifying crucial microbiomes associated with T2D. MA et al. [104] conducted a study that investigated the microbiomes associated with CRC using the same dataset as in our study. Among the various microbiomes they examined, the Prevotella microbiome stood out as strongly linked to CRC. This association aligns with the findings of microBiomeGSM, underscoring the significance of the Prevotella microbiome in the context of characterizing CRC. Chen et al. [105] conducted research using the same dataset to investigate microbiomes in the context of colorectal cancer, akin to the proposed microBiomeGSM tool. Similar to the findings of microBiomeGSM, their study also identified Peptostreptococcus, Fusobacterium, and Porphyromonas microbiomes as valuable and effective biomarkers for CRC. This convergence in results underscores the potential significance of these specific microbiomes in CRC characterization and their importance as potential biomarkers for the disease.

In summary, via analyzing the raw microbiome data of specific diseases, this study aims to identify taxonomic biomarkers that may have a role in the associated diseases. Three different taxon levels (genus, family, and order) are studied, and disease prediction is performed by building effective machine learning models using the G-S-M approach. 4 different datasets are analyzed and the identified microorganisms at genus, family and order levels are compared with the existing literature.

4.6.2 Limitations of the study

The quality and the scope of our study have been significantly influenced by several primary limiting factors. These factors encompass the nature of the data set, the tools employed for data preprocessing, the specific taxon groups considered, and the overall volume of data under examination. First and foremost, the data set itself plays a pivotal role in shaping the outcomes and conclusions of our study. Its size, diversity, and

representativeness directly impact the generalizability of our findings. Furthermore, the quality of data, its sources, and any potential biases within the dataset significantly affect the reliability of our results. Equally significant is the role of the tools employed for data preprocessing. The choices made in data cleaning, feature selection, and data transformation can introduce variability and influence the robustness of our analytical pipeline. It is paramount to acknowledge how these preprocessing steps can shape the study's outcomes. Additionally, our study's focus on specific taxon groups within the dataset should be considered. The selection of these taxonomic levels and the criteria used for their inclusion or exclusion has bearing on the granularity and relevance of our findings. Finally, the number of data points utilized in our analysis is another crucial factor. A larger dataset provides a broader and potentially more representative sample, which can enhance the reliability and statistical power of our results. Conversely, a smaller dataset may limit the generalizability of our conclusions. A comprehensive understanding of these limiting factors is essential for contextualizing our study's outcomes and conclusions.

Chapter 5

CCPRED

5.1 Motivation

Several studies have attempted to explore the composition and functionality of the gut microbiome in the context of CRC, but a comprehensive study of the gut microbiome in CRC patients has yet to be conducted. The present study attempts to close this gap by developing a robust classification model that facilitates the diagnosis of colorectal cancer. This can be accomplished by carefully analyzing a variety of CRC-related metagenomics datasets using a spectrum of feature selection techniques and machine learning methods. The objectives of the present study include the identification of biomarkers associated with CRC at the species level, as well as at the enzyme and pathway levels that influence host metabolism. The aim of the present study is to identify the most informative features for optimal CRC classification with a reduced feature set containing data on species, enzymes and pathways, leading to better classification results. Essentially, another goal is to develop a classification model that can perform best even with fewer features. To accomplish this, utilization of a dataset comprising metagenomic information from 9 distinct datasets, encompassing case and control groups from 8 diverse populations, is employed. Relative abundance values of the species, enzymes, and pathways are obtained from the same samples and presented as three different datasets. For each population, information at the level of species, enzymes and pathways are used as features and the relative abundance of these structures in the human gut is used as the value of these features to evaluate the performance of machine learning. In the present study, Emphasis is placed on the utilization of feature selection algorithms to optimize feature sets, thereby achieving more accurate classification with fewer features. The effects of the union and intersection of features selected by different feature selection algorithms on disease prediction performance are investigated. Based on the superior performance of the union features selected by the feature selection algorithms, the union features are ranked by

rescaling the importance value (performance of the classification algorithm) for each population and calculating the median of these values to produce a final ranking. Based on this ranking, the features in the top 20 are highlighted as potential biomarkers, and validation of the top 5 features is conducted through a literature search. Furthermore, the present study aims to identify population-specific CRC metagenomic biomarkers by developing population-specific models. These machine learning methods will be applied separately to population-specific datasets related to CRC to identify taxonomic biomarkers, enzymes, and pathways related to CRC that are specific to different populations. In order to evaluate the performance of the models, the following procedures are utilized: i) within-population analysis, ii) cross-population analysis, iii) leave-one-dataset-out (LODO) analysis.

5.2 Proposed Model (CCPRED)

The proposed method of the present study involves a combination of computational analyses and advanced machine learning techniques with the following main objectives: (i) by applying well-known feature selection algorithms to metagenomic data globally, performing a robust CRC prediction is aimed; (ii) analyzing the effect of the identified features on each population is also aimed; (iii) Another objective is attaining a rigorous evaluation of the classification model developed at the global/population-specific scale, having the potential to improve the accuracy and effectiveness of CRC diagnosis globally and population specific manner. (iv) The final aim is to identify global and population-specific metagenomic biomarkers across different molecular levels, including species, enzymes, and pathways. The performance of these biomarkers is systematically assessed using machine learning techniques, providing insight into their diagnostic potential.

To this end, firstly, all features for each molecular level (species, enzyme, and pathway) are analyzed using a global perspective for CRC prediction. Then, feature selection algorithms are applied to reduce the number of features. Finally, CRC prediction is performed, using the union and intersection of the features that are selected by different feature selection algorithms. For population-based experiments, the following three approaches are utilized, i.e., within-population, cross-population and LODO. In these approaches, the intersection and union features that are obtained in the global analyses are also used, and the CRC prediction performance is evaluated. In the present study, to evaluate the results obtained with the proposed method, comparisons are made with the

performance of the microBiomeGSM tool developed by [106]. microBiomeGSM is a novel approach developed for the identification of taxonomic biomarkers from metagenomic data, based on a methodology based on grouping, scoring, and modelling (G-S-M) [107]. This approach aims to detect disease-associated taxonomic biomarkers by developing an efficient machine learning model that analyses taxonomically transformed microbiome sequencing datasets. The microBiomeGSM tool utilizes species-level information and groups taxonomic features at different levels such as genus, family, and order to identify key taxonomic groups associated with specific diseases. Since microBiomeGSM performs best at genus taxon level for species data, the results at this level are included in the comparison. Additional, G-S-M approach [107] forms the basis for developing tools like maTE [108], PriPath [109], GediNET [110], miRcorrNet [111], 3Mint [112], GeNetOntology [113], TextNetTopics [114], TextNetTopics Pro [115] microBiomeGSM [106], miRGediNET [116], miRdisNET [117], miRModuleNet [118], CogNet [119] and AMP-GSM [120], which integrate biological networks and prior knowledge to provide a comprehensive understanding of genetic interactions. For an extensive review of feature selection approaches based on the grouping of features, the reader is referred to [121] and [122].

The following subsections present further details of the methodology of the present study.

5.2.1 Identification of important features (species, enzymes, and pathways), using different feature selection algorithms on the global scale

Using CRC-associated metagenomic datasets, a set of machine learning models is built to discriminate CRC from control samples using different feature selection algorithms and classification models. As illustrated in Figure 5.1, there are two main parts in the workflow: (i) feature selection to identify the most relevant species, enzymes, and pathways for the development of CRC diagnostic model; (ii) model building and classification. As shown in Figure 5.1, four different machine learning algorithms (Random Forest, LogitBoost, AdaBoost, and Decision Tree) are used for the classification task. For feature selection, Information gain (IG), SKB [73], and extreme gradient boosting (XGBoost) [79] are utilized. In addition to using traditional feature selection algorithms individually, a hybrid approach is employed, combining SKB, IG, and XGBoost methods as follows:

1. The importance scores of each feature are calculated using the feature selection algorithms mentioned above. Of the importance scores generated by tree-based classification algorithms, only the score generated by the random forest classification algorithm is used. Only the score generated by the random forest classification algorithm is used.
2. To ensure consistency, min-max scaling is applied to these values.
3. Features with importance scores below a certain threshold (0.5) are discarded.
4. If a feature that passes this filtering process is identified by all three feature selection algorithms, it is assigned to the intersection set.
5. If a feature that passes this filtering process is identified by at least one of the three feature selection algorithms, it is assigned to the union set.

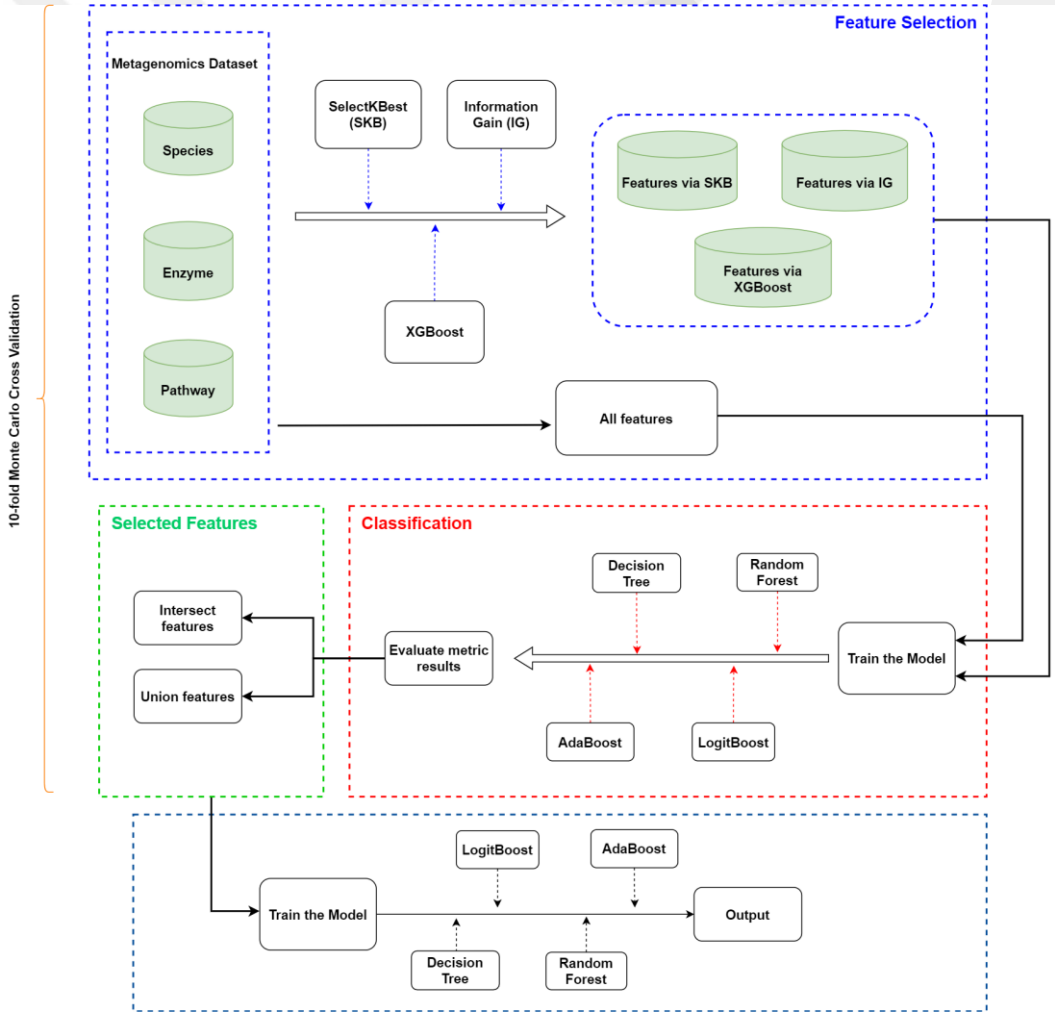


Figure 5.1 Workflow of the methodology (the present study).

Classification is performed using the following three sets of features:

i) Using all features without applying any feature selection algorithm: For each dataset (species, enzyme, and pathway), classification is performed using the above-mentioned machine learning algorithms and all features without applying any feature selection method. There are 917, 2895 and 551 features in species, enzyme, and pathway datasets, respectively.

ii) Using union of features that are selected by at least one of the three different feature selection algorithms: This method evaluates the performance of different machine learning methods by utilizing the union of the features that are selected by at least one of the feature selection algorithms. Since the focus is on the top 100 features selected by different feature selection algorithms and the features with importance scores higher than 0.5, different numbers of features are extracted for the species, enzyme, and pathway datasets. The number of features in the union set is 21, 295, and 38 for species, enzyme, and pathway datasets, respectively.

iii) Using intersection of the features that are commonly identified by all three feature selection algorithms: In this method, the performance of machine learning methods is evaluated based on the intersection of the features that are selected by different feature selection algorithms. Each feature in the intersection set is among the top 100 features and has feature importance score higher than 0.5 for each feature selection algorithm. In the intersection set, different numbers of features were obtained for the species, enzyme, and pathway datasets. The number of features in the intersection set is 9, 25, and 6 for species, enzyme, and pathway datasets, respectively.

In this study, CRC-associated biomarkers are identified globally and in a population-based manner. Thus, the present study goes beyond biomarker identification by aiming to discover population-specific metagenomic biomarkers in three different datasets (species, enzymes, and pathways).

5.2.2 CRC classification by population-specific meta-analysis using metagenomic features (species, enzyme, and pathway)

To identify population-specific taxonomic biomarkers, enzymes, and pathways, the methods described in Section 5.2.1 are applied to population-specific datasets related to CRC. To this end, three different meta-analyses are performed: within-population, cross-population, and Leave One Dataset Out (LODO). This multi-faceted approach provides a deeper understanding of how metagenomic biomarkers differ across populations, and it

also offers valuable insights into the potential applicability of these biomarkers in different contexts. In these experiments, the focus is on the RF algorithm because it outperforms the other classification algorithms in the preliminary analysis (global perspective) and RF is the most commonly used algorithm in human microbiome studies as reported by [55]. Figure 5.2 demonstrates our workflow for the population-specific evaluation.

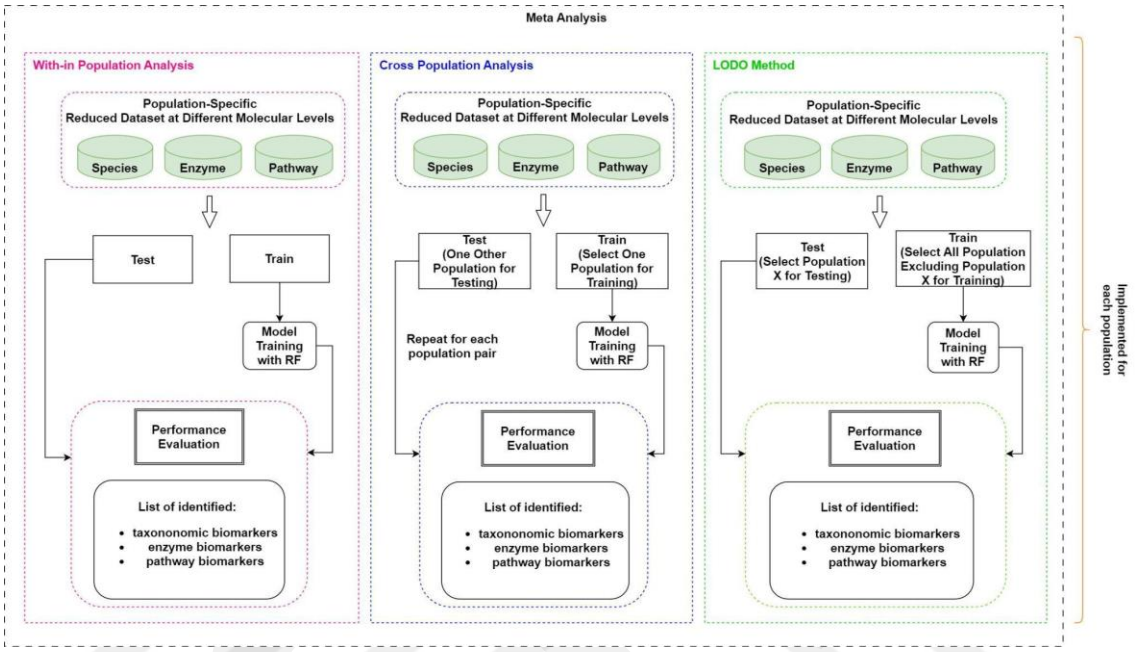


Figure 5.2 Population-specific evaluation of the models that are developed with the intersection / union features of the CRC-associated metagenomics dataset

In order to calculate the performance metrics separately for each population dataset, within-population analysis was applied. A 10-fold Monte Carlo Cross Validation (MCCV) is performed for each population dataset. Data from each population are selected 80% for training and 20% for testing. The average AUC values and standard deviations are calculated.

In the cross-population analysis, the model is trained with the data from a specific population and the developed model is tested with the data from another population that was not used for training. The testing part of this experiment is repeated separately for each population that was not used in the training part. Each dataset is used to train the model, and each time the remaining datasets are used separately as test data.

In Leave One Dataset Out (LODO) analysis, the data from a specific population is kept as the test set, while the data from all other populations that were not selected for testing are combined and used for training the model. This experiment is repeated for each population.

5.2.3. Identification of CRC-associated species, enzymes, and pathways as potential biomarkers across different populations

In order to calculate the final importance score of each feature across different populations, the population-specific contribution of each feature to the Random Forest classifier, which outperforms other classification algorithms, is evaluated. In these evaluations, depending on the best mean scores obtained using the within-population analysis, possible biomarkers obtained by union features are used for species data, intersection features for enzyme data, and union features for pathway data. Then, using the min-max scaling method, the median values of these scaled feature importance scores are obtained and a ranking list is generated. This ranking is simply a ranking of the selected features, i.e., a lower score does not mean that the feature is not important at all. Using this ranking list, the top 20 features for species, enzyme, and pathway datasets are identified, and the top 5 features are explored in depth via referring to the biological literature.

5.3 Implementation of CCPRED

The models are developed using the KNIME platform [93], and the H2O and scikit-learn libraries are utilized. The predictive performance of the models is evaluated using the metrics of accuracy, F1 score, and AUC (Area Under the Receiver Operating Characteristic Curve). As shown in Figures 5.1 and 5.2, the generated models are tested on three separate datasets including the relative abundance values of microorganisms (species), enzymes, pathways that are calculated for CRC patients and healthy samples.

5.4 Results

The main objective of this study is i) to predict CRC with high machine learning performance using few features, and ii) to identify the species, enzymes, and pathways associated with this disease. Another important goal of this study is to provide more accurate and specific information for CRC diagnosis and treatment through more global and population-specific experiments. Using different feature selection algorithms, dominant features will be highlighted and the effect of these features on CRC prediction performance will be investigated in detail using the proposed approaches. In this context, the results are evaluated from two different perspectives. Firstly, using a global dataset,

the performance of the classification algorithms is comparatively evaluated using (i) all features, (ii) the intersection of the features that are selected by all feature selection algorithms (intersection features), and (iii) the union of the features that are selected by at least one feature selection algorithm (union features). Secondly, using population-specific datasets, within-population, cross-population and LODO experiments are performed; and the performance of the generated models using different feature sets, as explained for the global analysis, are comparatively evaluated. This performance evaluation is repeated for three different datasets containing species, enzymes, and pathways as features for the same samples.

5.4.1 Performance evaluation of the global models

The experiments conducted on the global CRC-associated metagenomics dataset, which includes samples from different populations, offer general insights into the CRC development.

5.4.2 Performance evaluation of the global models using all features

Using 10-fold cross-validation, 4 different classifiers (Random Forest, AdaBoost, LogitBoost, and Decision Tree) are run on CRC-associated metagenomic data without applying any feature selection method; and the average performance metrics and standard deviations are shown in Table 5.1. For species, enzyme and pathway datasets, the highest performance is achieved with the RF classification methods. Specificity, sensitivity, F1 measure, precision, accuracy and AUC are used as evaluation criteria. As presented in Table 5.1, the best results were obtained with the RF classifier with an AUC of 0.83 on the species dataset. For the enzyme dataset, the best results were obtained with the RF classifier with an AUC of 0.78. For the pathway dataset, the best results were obtained with the RF classifier with an AUC of 0.76.

Table 5.1 Performance evaluation of classification algorithms using all the features within CRC-associated species, enzyme, and pathway datasets.

CRC-ASSOCIATED SPECIES DATASET						
Model	**Accuracy**	**Sensitivity**	**Specificity**	**F1-Measure**	**AUC**	**Precision**
AdaBoost	0.53 ± 0.008	0.66 ± 0.06	0.76 ± 0.06	0.69 ± 0.002	0.79 ± 0.02	0.53 ± 0.004
DT	0.52 ±9.93E-9	0.64 ± 0.05	0.70 ± 0.05	0.69	0.68 ±0.04	0.53 ±9.93E-9
LogitBoost	0.53 ± 0.003	0.65 ± 0.07	0.76 ± 0.05	0.69 ± 0.001	0.81 ± 0.02	0.53 ± 0.001
RF	0.52 ±9.93E-9	0.66 ± 0.02	0.84 ± 0.06	0.69	0.83 ± 0.03	0.53 ±9.93E-9
CRC-ASSOCIATED ENZYME DATASET						
Model	**Accuracy**	**Sensitivity**	**Specificity**	**F1-Measure**	**AUC**	**Precision**
AdaBoost	0.69 ± 0.03	0.63 ± 0.1	0.73 ± 0.03	0.72 ± 0.01	0.76 ± 0.01	0.65 ± 0.04
DT	0.49	0.60 ± 0.1	0.60 ± 0.05	0.66	0.63 ± 0.01	0.49
LogitBoost	0.68 ± 0.02	0.63 ± 0.06	0.73 ± 0.04	0.72 ± 0.009	0.76 ±0.01	0.64 ± 0.03
RF	0.67 ± 0.02	0.61 ± 0.07	0.75 ± 0.07	0.73 ± 0.01	0.78 ±0.009	0.61 ± 0.02
CRC-ASSOCIATED PATHWAY DATASET						
Model	**Accuracy**	**Sensitivity**	**Specificity**	**F1-Measure**	**AUC**	**Precision**
AdaBoost	0.62 ± 0.05	0.62 ± 0.05	0.69 ± 0.03	0.70 ± 0.01	0.71 ± 0.02	0.58 ± 0.04
DT	0.49	0.59 ± 0.04	0.65 ± 0.07	0.66	0.59 ± 0.03	0.49
LogitBoost	0.61 ± 0.03	0.57 ± 0.08	0.69 ± 0.04	0.70 ± 0.01	0.70 ± 0.006	0.57 ± 0.03
RF	0.69 ±0.02	0.59 ± 0.07	0.76 ± 0.08	0.73 ± 0.008	0.76 ± 0.01	0.65 ± 0.02

5.4.3 Performance evaluation of the global models using feature selection

In this subsection, the results obtained using feature selection algorithms are presented. 4 different feature selection algorithms including traditional feature selection algorithms (SelectKBest (SKB), Information Gain (IG), (XGBoost)) and microBiomeGSM [106] which is a biological domain-knowledge based feature selection technique; and 4 different classification algorithms (AdaBoost, LogitBoost, Decision Tree and Random Forest) are used for CRC prediction. The grouping of the features is performed at the genus level while running the microBiomeGSM tool. Figure 5.3 provides a comparison of the AUC values obtained with different classification algorithms using different feature selection techniques on the global scale. Figure 5.3 (A) shows a comparative assessment of the AUC values obtained using the CRC-associated species dataset. Figure 5.3 (B) and Figure 5.3 (C) show the comparative AUC values that are obtained for the enzyme and pathway datasets, respectively. Since microBiomeGSM uses Random Forest as the classification algorithm, its performance is compared with other feature selection algorithms only using RF classifier. In Figure 5.3, the first two bars in each graph (colored in red and orange) represent the intersection and union

features, respectively. The third bar in each graph (colored in yellow) shows the performance of the model using all features (when no feature selection method is used). The fourth bar in each graph (colored in light blue) shows the results obtained using XGBoost feature selection algorithm. The fifth bar in each graph (colored in green) shows the results obtained using Information Gain (IG) feature selection algorithm. The sixth bar in each graph (colored in dark blue) shows the results obtained using Select K Best feature selection algorithm. The last bar in the Random Forest classifier and in the mean values (colored in steel blue) show the results obtained with the microBiomeGSM tool, where the features are grouped on genus level.

Figure 5.3 (A) illustrates the mean AUC values for different feature selection algorithms, calculated by averaging the AUC values obtained from various classification algorithms. The highest AUC averaged over different classifiers (an AUC of 0.78) is obtained by using all features. Among different classifiers, for each feature selection method except for the SKB feature selection method, the highest AUC values are obtained using the RF algorithm. This underlines the strength of the RF algorithm. Figure 5.3 (A) shows that for the species dataset, the RF classifier achieves an AUC of 0.83 when all features (917 features) are used. For the same dataset, the RF classifier using the intersection of selected features by SKB, IG, and XGBoost algorithms (9 features) results in 0.78 AUC. For the same dataset, the RF classifier using the union of selected features by SKB, IG, and XGBoost algorithms (21 features) yields 0.79 AUC. The union feature set including relative abundance values of only 21 species shows very close performance with the model that uses the relative abundance values of 917 species.

For the enzyme dataset, when the mean AUC values are analyzed in Figure 5.3 (B), it is observed that the highest AUC is obtained by using the union of the features selected by different feature selection methods. In the experiments performed on the enzyme dataset, the highest AUC values are obtained with the Random Forest classification algorithm. Examining Figure 5.3 (B), an AUC of 0.76 is obtained with the RF classification algorithm using all features (2895 features); an AUC of 0.76 is obtained using the intersection of the features (25 features) that are identified by SKB, IG, and XGBoost feature selection algorithms; and an AUC of 0.79 is obtained using the union of the features (295 features) that are detected by different feature selection algorithms.

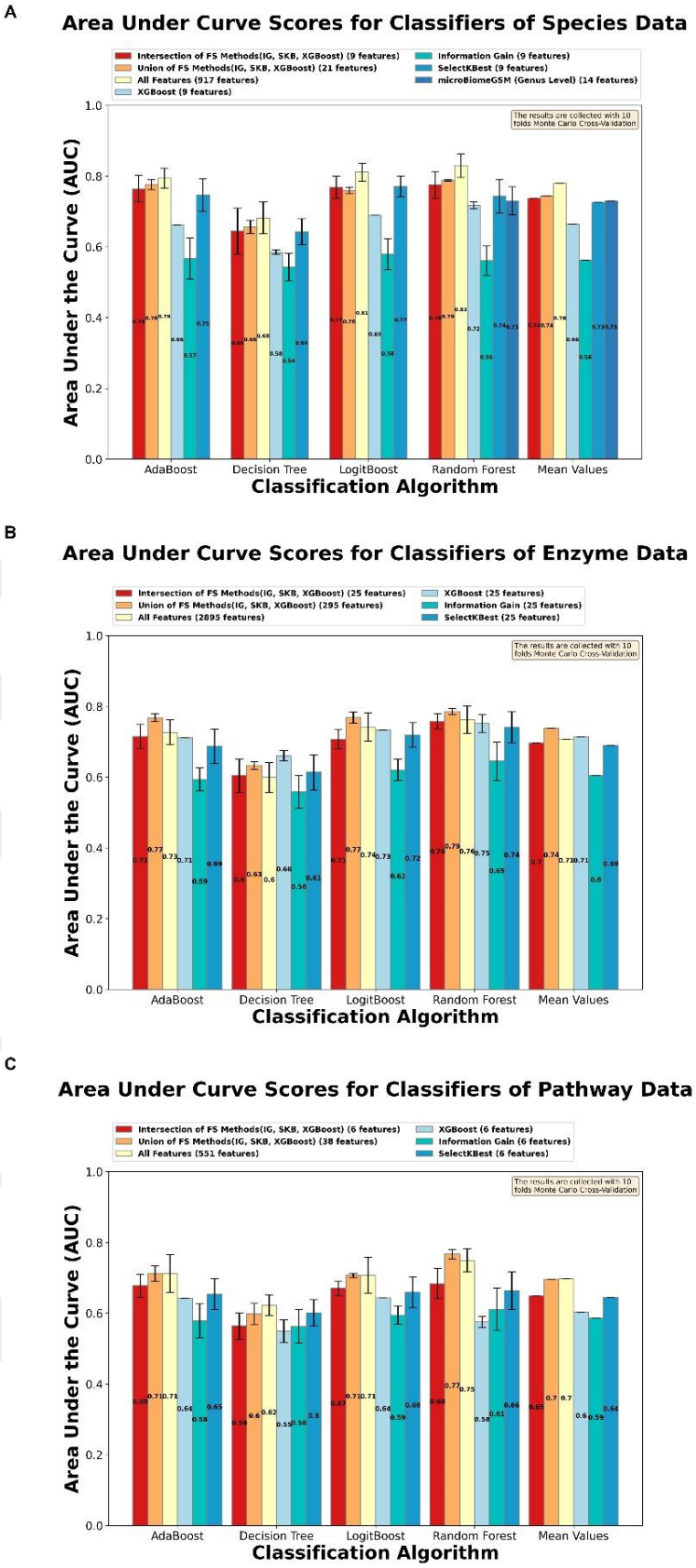


Figure 5.3 Performance evaluation of different feature selection techniques using different classifiers on CRC-associated A) species, B) enzyme and C) pathway datasets.

These results imply that by examining only the abundance values of 25 enzymes in the intersection set, one can perform classification as successfully as analyzing the 2895 enzymes (all features in the enzyme dataset). Hence, it can be deduced that the method of the present study resulted in higher AUC values using fewer features for the CRC-associated enzyme dataset.

When the mean values among different classifiers are analyzed on the pathway dataset, it is seen in Figure 5.3 (C) that the AUC values that are obtained by using all features and by using the union features are similar. When the AUC values on the CRC-associated pathway dataset is analyzed, one can notice that the highest AUC values are obtained using the Random Forest classification algorithm among different classifiers. Examining Figure 5.3 (C), an AUC of 0.75 is obtained with the RF classification algorithm using all features (551 features); an AUC of 0.68 is obtained using 6 features that are commonly identified by SKB, IG and XGBoost feature selection algorithms; and an AUC of 0.77 is obtained by using 38 features that are identified by at least one of the feature selection algorithms (union features). Obtaining higher AUC values using only 38 features that are identified by at least one of the feature selection algorithms emphasizes that these pathways are more informative compared to using all 551 pathways.

5.4.4 Performance evaluation of population-specific models

Table 5.2 and Figure 5.3 show that the Random Forest classifier performs better than the Decision Tree, LogitBoost and AdaBoost algorithms in classifying CRC on the global-scale experiments. Therefore, the Random Forest classifier is deliberately utilized in the population-specific experiments. In particular, the potential of the Random Forest classification algorithm in population-based analysis of species, enzyme, and pathway datasets is investigated using the intersection/union of features identified by different feature selection algorithms. In order to evaluate the performance of the models, the following three methods are applied: i) within-population, ii) cross-population, iii) leave-one-dataset-out (LODO). In this way, population-specific biomarkers associated with CRC are highlighted as a function of populations.

5.4.5 Findings for within-population analysis:

In the within-population analysis, each metagenomics dataset that is specific to a population is analyzed separately. In these experiments, the RF algorithm is used as the

classification algorithm and a 10-fold Monte Carlo cross-validation method is applied. Figure 5.4 shows a comparison of the AUC values obtained for different population datasets using i) all features, and ii) the intersection and union of the features that are detected by the feature selection algorithms. In Figure 5.4, the mean AUC values among different populations are also plotted. Figures 5.4 (A), (B) and (C) illustrate the AUC values of the models for the species, enzyme, and pathway datasets, respectively. In each figure, the first two bars (colored in red and orange) indicate the AUC values of the models, using intersection and union features, respectively. The third bar (colored in yellow) shows the performance of the model when all features are used without applying feature selection methods. For the species dataset, the last bar (colored in steel blue) in Figure 5.4 indicates the findings of microBiomeGSM tool where the features are grouped on the genus level.

Through our within-population experiments using the species, enzyme, and pathway datasets, it was observed that for most of the cases (17 out of 27 comparisons), the AUC values that are obtained using the union features were higher than the AUC values obtained using all features (as shown in Figure 5.4). More specifically, Figure 5.4 (A) shows the AUC values obtained for the CRC-associated species dataset. As can be seen in Figure 5.4 (A), for five out of nine populations ("AUT", "FRA", "ITA", "JPN" and "USA"), the AUC values obtained using the union features (21 features) are higher than the AUC values reported with other approaches (intersection of features, microBiomeGSM, and all features). Among the AUC values that are obtained for different CRC-associated species datasets belonging to different populations, the highest AUC value (an AUC of 0.91) is reported for "AUT" population using the 21 features included in the union set. In only one of the nine different datasets ("JP" population), the AUC value (0.82) that is obtained using the intersection of the features (9 features) is higher than the AUC values reported by other approaches. Using 9 common features from the species datasets, the highest AUC values (0.84) are noted for "AUT" and "FRA" populations. For two of the nine different datasets ("DEU", "IND"), the AUC metrics obtained using all features (917 features) are better than the AUC values reported by other approaches. The highest AUC value among these populations was obtained for "IND" with an AUC of 0.93 using all features. When mean AUC values are analyzed, one can observe from Figure 5.4 (A) that the highest AUC value of 0.82 is obtained with the union features (relative abundance values of 21 species).

Figure 5.4 (B) illustrates the AUC values derived from the CRC-associated enzyme dataset, showcasing outcomes obtained through the utilization of union features, intersection features, and the inclusion of all features. As shown in Figure 5.4 (B), for five of the nine different datasets ("DEU", "FRA", "IND", "JPN", and "JP") the models that are developed using intersection features outperform other models in terms of the AUC metrics. The "IND" population dataset attains the highest AUC value of 0.98. In three out of the nine distinct datasets ("AUT", "CHN", and "ITA"), AUC metrics derived from utilizing all features exhibit a slight improvement over the AUC values obtained from alternative models. Among the AUC values observed across various CRC-associated enzyme datasets from diverse populations, the "AUT" population achieved the highest AUC value of 0.83 when utilizing all features (2895 features). In Figure 5.4 (B), the analysis exclusively focusing on all features (depicted by the yellow bar) reveals that the "IND" population yields the highest AUC value of 0.96 among all AUC values derived from the utilization of all features. In solely one of the nine distinct datasets (specifically, the "USA" population), the AUC value of 0.60 achieved through the utilization of feature union (comprising 295 features) surpasses the AUC values obtained through alternative methodologies. In Figure 5.4 (B), exclusive analysis of union features (represented by the orange bar) comprising 295 features reveals that the "IND" population exhibits the highest AUC value of 0.91 among all AUC values derived from the utilization of union features. Upon examination of mean AUC values depicted in Figure 5.4 (B), it becomes evident that the intersection features (comprising relative abundance values of 25 enzymes) yield the highest AUC value of 0.75. Figure 5.4 (B) indicates that the frequency of superior performances achieved through the intersection of features identified by feature selection methods exceeds that of other methodologies. Figure 5.4 (C) presents the AUC outcomes for pathway data, comparing the utilization of union features identified by various feature selection methods, intersection features identified by different feature selection methods, and all features without any feature selection method. As depicted in Figure 5.4 (C), among six out of nine populations ("AUT", "DEU", "FRA", "ITA", "JPN", and "USA"), the utilization of union features (comprising 38 features) yields higher AUC values compared to other methodologies, including intersection of features, microBiomeGSM, and utilization of all features. Among the AUC values derived from distinct colorectal cancer (CRC)-associated species datasets across various populations, the most elevated AUC value, reaching 0.88, is documented for the "FRA" population. This result is achieved through the utilization of 38 features encompassed within the union set.

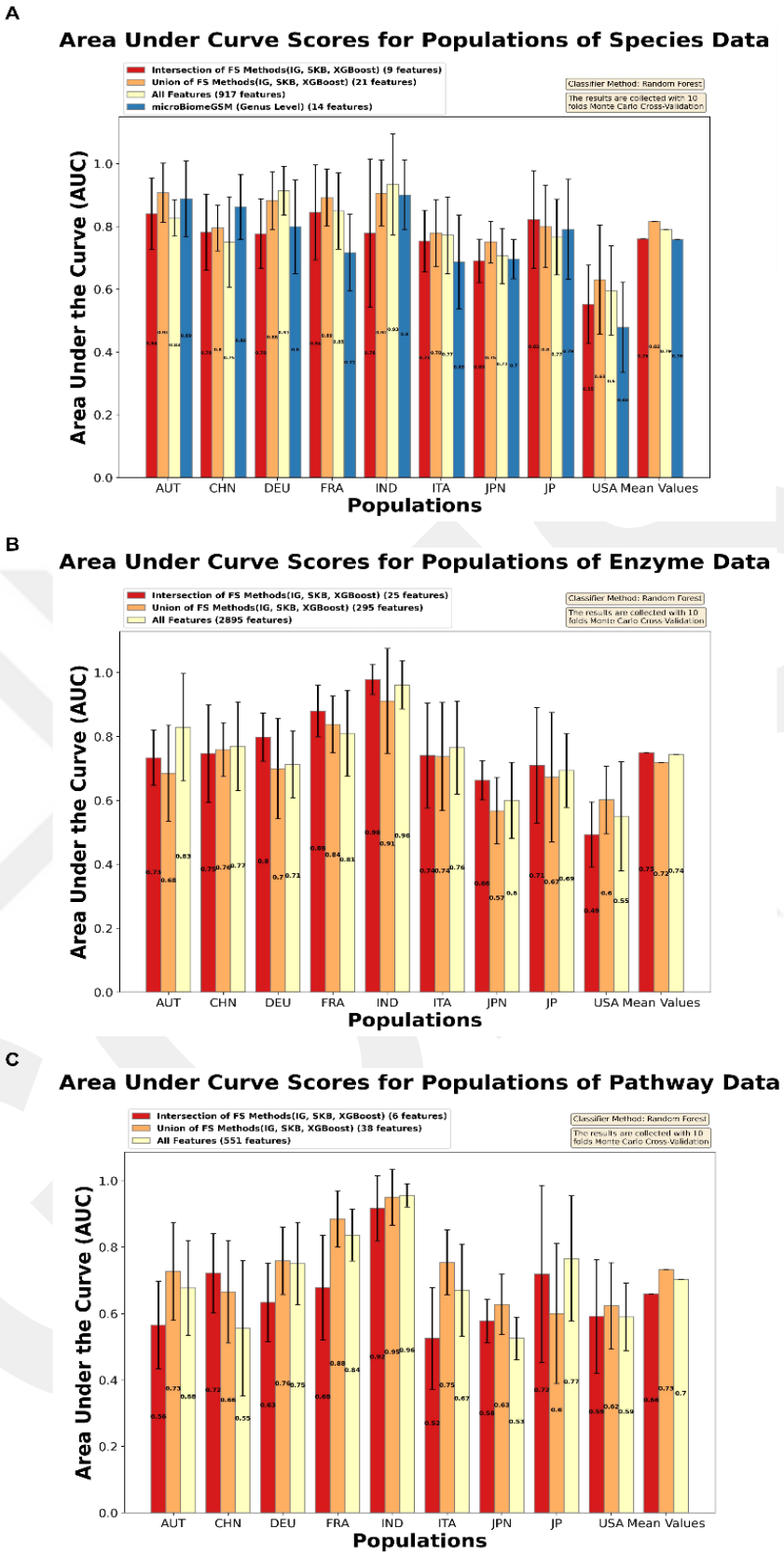


Figure 5.4 AUC values that are obtained as part of the within-population analysis using A) species, B) enzyme and C) pathway features of the CRC-associated population-specific metagenomic datasets. AUC values of different feature selection methods are represented by different colors and comparatively evaluated.

In Figure 5.4 (C), exclusive examination of union features (represented by colored orange bars) reveals the attainment of the most prominent outcome for the "IND" population concerning the AUC. Specifically, the utilization of union features yields an AUC value of 0.95 for this population subgroup. Among the nine distinct datasets examined, namely the "CHN" population, it is noteworthy that the utilization of the intersection of features results in a comparatively higher AUC value of 0.72, surpassing the AUC values obtained through alternative approaches in this specific population cohort. Figure 5.4 (C) demonstrates that exclusive analysis of intersection features (depicted by colored red bars) yields the most superior outcome for the "IND" population concerning the area under the curve (AUC) values obtained using union features. Specifically, an AUC value of 0.92 is observed in this population subgroup. As depicted in Figure 5.4 (C), it is notable that among the nine distinct datasets analyzed, specifically the "IND" and "JP" populations, the models constructed utilizing all features exhibit superior performance compared to alternative models, as evidenced by (AUC) metrics. Among (AUC) values derived from various colorectal cancer (CRC)-associated pathway datasets across diverse populations, the most elevated AUC value of 0.96 is documented for the "IND" population. This notable outcome is achieved through the utilization of all features, encompassing a total of 2895 features. Upon examination of the mean (AUC) values, it becomes apparent from Figure 5.4 (C) that the most notable AUC value, reaching 0.73, is attained through the utilization of union features. These features comprise relative abundance values associated with 38 pathways. Figure 5.4 (C) illustrates a greater frequency of superior performances achieved through the integration of features identified by feature selection methods, compared to alternative approaches.

5.4.6 Findings for cross-population analysis:

In this evaluation method, the model is trained using data from a specific population and the developed model is tested separately using data from another population that was not used for training. This experiment is repeated for each population. One by one, every dataset is used for training the model and each time the remaining datasets are used separately as the test data. Figure 5.5 (A), Figure 5.5 (B) and Figure 5.5 (C) show the AUC values of the machine learning models that are developed using the cross-population technique and applied on species, enzyme, and pathway datasets, respectively.

By examining the average AUC values obtained throughout the within-population experiments, the feature selection approach yielding the highest mean AUC value is

selected for the cross-population analysis. It can be observed from Figure 5.4 (A), for the species dataset, the mean AUC value generated by the union features (21 species) is higher than other tested methods. Hence, we used the reduced dataset including only these 21 features through cross-population experiments of the species dataset. For the enzyme dataset, as shown in Figure 5.4 (B), the models using the intersection features (25 enzymes) resulted in the highest mean AUC value. Hence, the reduced dataset containing solely these 25 features was utilized in cross-population experiments of the enzyme dataset. For the pathway dataset, as shown in Figure 5.4 (C), the models using the union features (38 pathways) generated the highest AUC values when averaged over different populations. Hence, the reduced dataset comprising solely these 38 features was employed in cross-population experiments of the pathway dataset. The Random Forest algorithm was used as the classifier in cross-population analyses since the Random Forest classifier resulted in the highest AUC in our previous experiments with other classifiers tested on the global perspective (as shown in Figure 5.3).

As depicted in Figure 5.5 (A), cross-population experiments revealed that an AUC of 0.80 or higher was achieved in 7 out of 72 instances when employing the union features, consisting of 21 features, within the species dataset. Figure 5 (A) shows that using the relative abundance values of the 21 species within the species dataset, the highest AUC of 0.85 is obtained when "JPN" population is used as the training set and "JP" is used for the test set. In Figure 5.5 (A), it is demonstrated that employing the relative abundance values of the 21 species within the species dataset yields the highest AUC of 0.85. This result is attained when utilizing the "JPN" population as the training set and the "JP" population as the test set.

This high-performance metric between two different datasets collected from the same country but different regions emphasize the success of the proposed approach. The notable performance metric observed between two distinct datasets originating from the same population, but disparate regions underscore the efficacy of the proposed methodology.

The second highest AUC score of 0.83 is achieved when employing "DEU" as the training set and "JP" as the test set, while a similar AUC value of 0.83 is observed with "AUT" as the training set and "JP" as the test set.

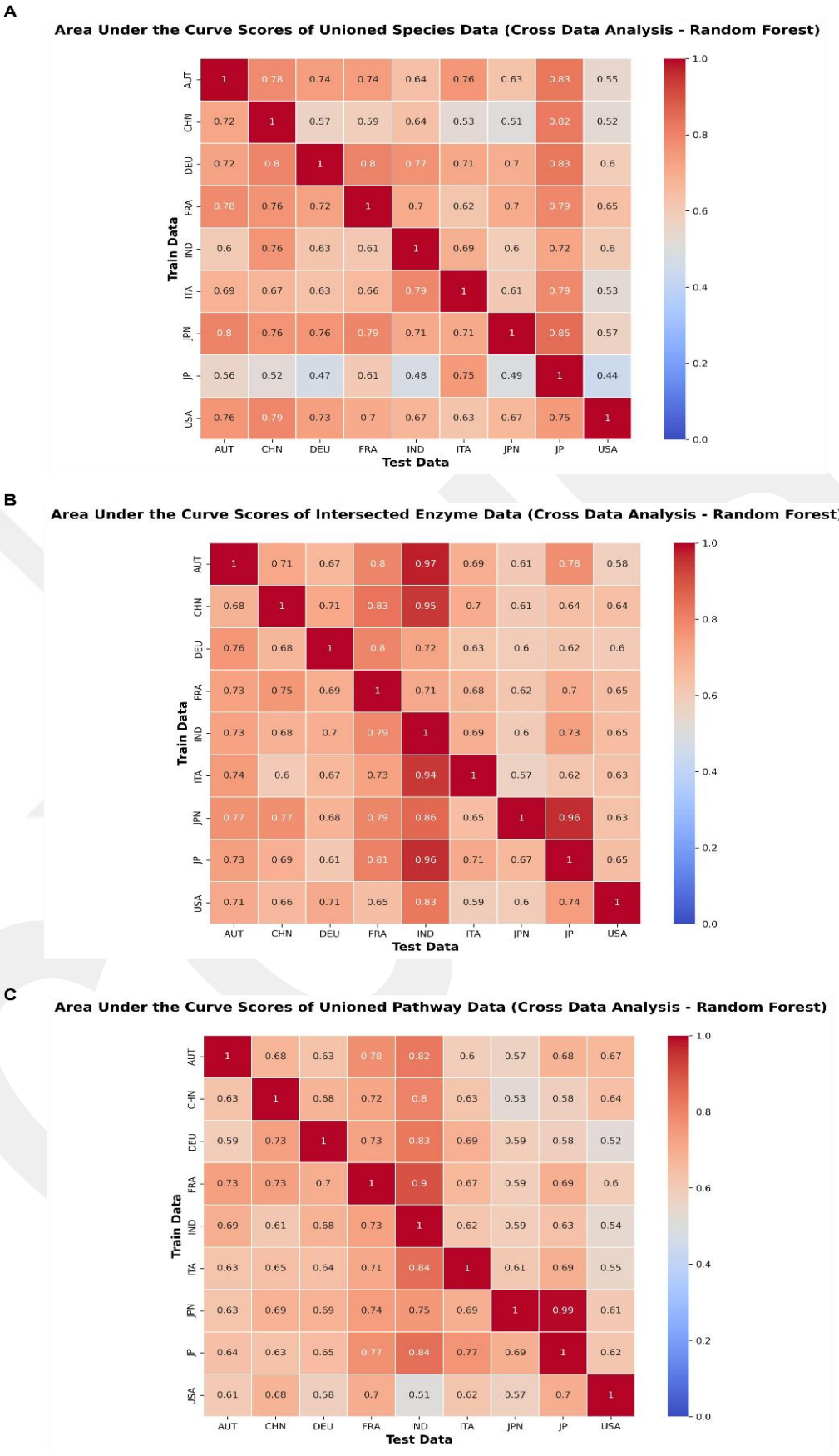


Figure 5.5 Cross-population analysis using union features for species, enzyme, and pathway data

As depicted in Figure 5.5 (B), an AUC of 0.80 and above was obtained for 11/72 cases using the intersection of features for the enzyme dataset. Figure 5.5 (B) shows that the best result is obtained with an AUC of 0.97 when using the intersection features selected by the feature selection methods for the enzyme data (25 enzymes/feature), using "AUT" as the training set and "IND" as the test set. The second-best result is obtained with 0.96 AUC when "JPN" is used as the training set and "JP" as the test set.

In Figure 5.5 (C), AUC of 0.80 and above is obtained for 7/72 cases by using the union of features determined by feature selection methods in cross-population analyses for pathway data. Figure 5.5 (C) shows that when "JPN" is used as the training set and "JP" as the test set, the best result of 0.99 AUC is obtained for the pathway data using the union features (38 features) selected by the feature selection algorithms. This high result between two data from the same population but from different regions underlines the success of the proposed approach. The second-best result is obtained with an AUC of 0.90 using "FRA" as the training set and "IND" as the test data for the pathway data.

5.4.7 Findings for Leave one dataset out (LODO) analysis:

In the Leave one dataset out (LODO) analysis, performance is evaluated by excluding one population dataset for repeated testing for species, enzyme, and pathway datasets, separately. In this experiment, one population is selected for testing and the remaining datasets are combined and used as training data. This experiment is repeated for each population. In these experiments, the RF algorithm is used as the classification algorithm and the 10-fold Monte Carlo cross-validation method is applied. A comparison of the AUC values obtained during the LODO analysis by using different feature selection methods (intersection, union, and all features) is shown for each population in Figure 5.6 (A), Figure 5.6 (B), and Figure 5.6 (C) for the species, enzyme, and pathway datasets, respectively. The first two bars in each graph (colored red and orange) represent the intersection of features identified commonly by all three feature selection algorithms, and the union of features selected by at least one of the three different feature selection algorithms, respectively. The third bar shows the performance of the model when all features were used without applying a feature selection method. The fourth bar for the species data shows the resulting AUC scores determined using the microBiomeGSM tool. As microBiomeGSM is a tool that uses species data, only comparisons with species data are included in these experiments. The evaluation of the LODO results is based on a

comparison of the performance of the proposed approaches (all features, microBiomeGSM intersection of features, and union of features).

In Figure 5.6 (A), it is evident that within the species dataset, the AUC outcomes derived from the union features, selected via feature selection methods, surpass those obtained through alternative approaches in two out of the nine distinct datasets ("DEU" and "FRA"). The highest AUC value among the various CRC-associated species datasets from diverse populations, standing at 0.84, is documented for the "FRA" population, employing the 21 features encompassed within the union set. In Figure 5.6 (A), exclusive examination of the union features (depicted by the orange bar) reveals the most favorable outcome for the "JP" population, exhibiting an AUC of 0.87 among the AUC values derived from these features. As depicted in Figure 5.6 (A), in the case of four out of the nine distinct datasets ("ITA", "JPN", "JP", and "USA"), the models constructed utilizing all features exhibit superior performance compared to alternative models, as evidenced by AUC metrics.

The "JP" population dataset yields the highest AUC value of 0.94. This result also represents the optimal one achieved for the species data. As depicted in Figure 5.6 (A), the models constructed using intersection features do not demonstrate superior performance compared to other models in terms of AUC metrics across any of the nine distinct datasets. The highest achievement observed with the intersection of features identified by the feature selection methods is an AUC of 0.85 for the "IND" dataset. As illustrated in Figure 5.6 (A), among the nine diverse datasets, the models employing microBiomeGSM exhibit superior performance in AUC metrics for three datasets, namely "AUT", "CHN", and "IND". In the array of AUC values derived from various CRC-associated species datasets across diverse populations, the apex AUC value of 0.90 is documented for the "IND" population, utilizing the 14 features amalgamated in the union dataset. Upon scrutiny of the mean AUC values, Figure 5.6 (A) reveals that the utmost AUC value of 0.77 is acquired when employing all features, encompassing the relative abundance values of 917 species.

Figure 5.6 (B) shows the AUC values obtained for the CRC-associated enzyme dataset using union features, intersection features, and all features. As can be seen in Figure 5.6 (B) for the enzyme dataset, in three of the nine different datasets ("AUT", "CHN", and "IND"), the AUC results obtained with the intersection features obtained by the feature selection methods are higher than the AUC values obtained with the other

approaches. Among the AUC values that are obtained for the different CRC-associated enzyme datasets belonging to different populations, the highest AUC value (an AUC of 0.89) is reported for the "IND" population using the 25 features included in the intersection set. In Figure 5.6 (B), when only the intersection features are analyzed (colored red bar), the highest result is obtained for the "JP" population among the AUC values obtained using the intersection features (AUC of 0.94). As shown in Figure 5.6 (B), for three of the nine different datasets ("ITA", "JPN" and "JP"), the models that are developed using all features outperform the other models in terms of AUC metrics. Among the AUC values that are obtained for the different CRC-associated enzyme datasets belonging to different populations, the highest AUC value (an AUC of 0.98) is reported for the "JP" population using the 2895 features included in the all-feature set. This result is also the best result obtained for the LODO approach for enzyme data. When analyzing the mean AUC values, Figure 5.6 (B) shows that the highest AUC value of 0.76 is obtained for the union features and all features.

As can be seen in Figure 5.6 (C) for the pathway dataset, in five of the nine different datasets ("CHN", "DEU", "FRA", "JPN" and "USA"), the AUC results obtained with the union features obtained by the feature selection methods are higher than the AUC values obtained with the other approaches. Among the AUC values that are obtained for the different CRC-associated pathway datasets belonging to different populations, the highest AUC value (an AUC of 0.83) is reported for the "USA" population using the 38 features included in the union set. In Figure 5.6 (C), when only the union features are analyzed (colored orange bar), the highest result is obtained for the "JP" population among the AUC values obtained using the union features (AUC of 0.92). As shown in Figure 5.6 (C), for three of the nine different datasets ("IND", "ITA" and "JP"), the models that are developed using all features outperform the other models in terms of AUC metrics. The highest AUC value of 0.96 is obtained for the "JP" population dataset. This result is also the best result obtained for the pathway data. As shown in Figure 5.6 (C), the models that are developed using intersection features do not outperform the other models in terms of AUC metrics for any of the nine different datasets. The best result obtained with the intersection of features determined by the feature selection methods is an AUC of 0.82 for “JP”. When analyzing the mean AUC values, it can be seen in Figure 5.6 (C) that the highest AUC value of 0.76 is obtained with the union features (relative abundance values of 38 pathways).

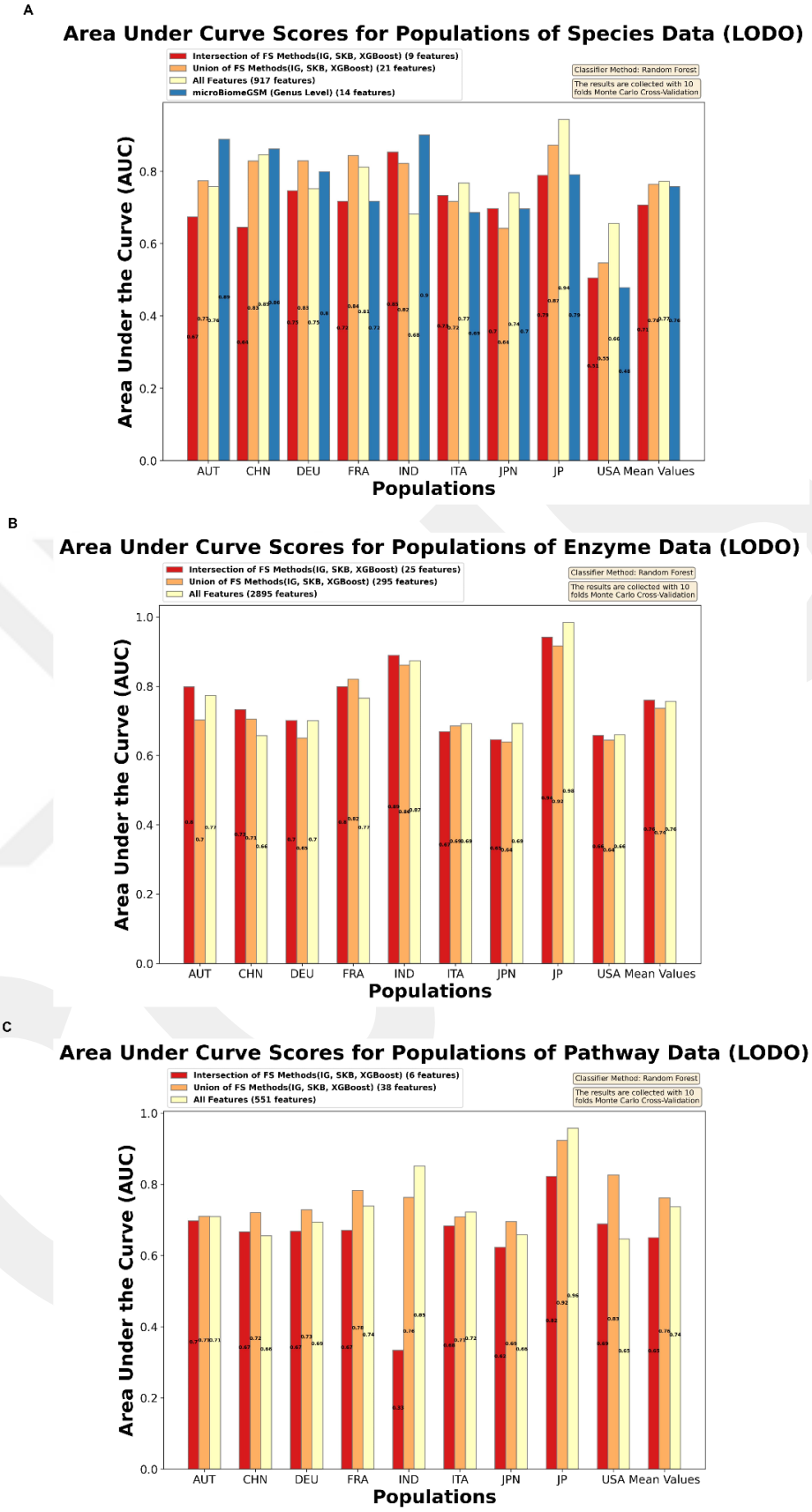


Figure 5.6 AUC values of the LODO analysis using species, enzyme and pathway features of the CRC associated metagenomic dataset. Each feature selection method is represented by a different color.

5.4.8 Potential biomarkers within the union/intersection features, as identified by different feature selection algorithms

In the present study, as shown in Figure 5.3 (A), in the species dataset, union features selected by at least one feature selection algorithm are emphasized as potential biomarkers. Since the performance of the union features is superior to other models that are developed using individual feature selection algorithms and also superior to other models that are generated using the intersection of the features that are selected by all feature selection algorithms; for this dataset, the decision was made to continue the analyses with the union set comprising 21 features. Intersection features selected by feature selection algorithms are highlighted as potential biomarkers in the enzyme dataset (see Figure 5.3 (B)). Since the performance of the intersection features is superior to other models built with individual feature selection algorithms, and superior to other models built with a union of features selected by all feature selection algorithms, we decided to continue the analysis for this dataset with the intersection set containing 25 features. A similar trend can be seen in Figure 5.3 (C), where the union feature set generates higher AUC than other tested feature selection algorithms. Hence, for pathway datasets, the analysis continued, utilizing the union set comprising 38 features. In CRC analyses conducted across global features and population-specific datasets, it becomes evident, upon examining mean values, that AUC scores derived from the union of features selected by feature selection methods yield superior results compared to those obtained from other feature sets (such as intersection features selected by feature selection algorithms and all features).

Among the union features selected by the feature selection algorithms, the importance scores of the features obtained after the classification process using the Random Forest classification algorithm are scaled using the min-max scaling method. The importance scores for each population are scaled between 0 and 100, with 0 denoting the lowest value, while 100 represents the most important value. The union features selected by the feature selection algorithms are assigned the importance value to which they belong in each population. Then the median value of these scores is calculated based on the features and the final ranking is made based on this value. The feature with the highest median value is presented as the most valuable feature. Recently, CRC prediction studies using metagenomic data, especially at the taxonomic level, have considerably increased. In this context, some species might influence risk factors associated with

colorectal cancer or play a role in cancer development. The proposed method revealed a number of species associated with colorectal cancer and their importance has been confirmed by experiments in the literature. The top 20 species include microbiomes that have been shown to be associated with colorectal cancer. The top 20 species obtained with the proposed approach are shown in Figure 5.7. In the "Discussion" section, the top 5 species among these 20 species are analyzed in detail.

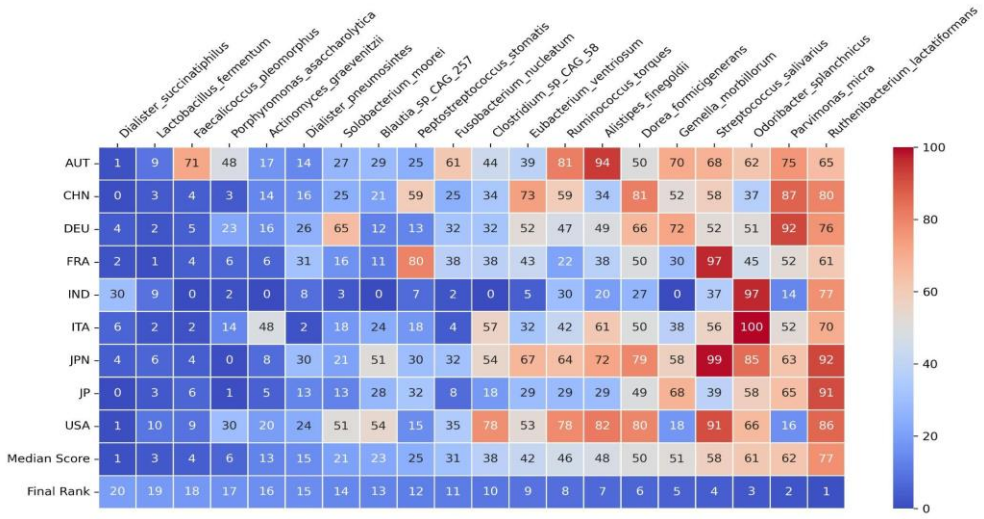


Figure 5.7 Heatmap of the feature importance scores for the identified species in global analysis. The scores of the features in the union set are visualized in different rows for different populations. While higher scores are colored in red, lower scores are colored in blue. The union approach embraces the features that are identified in at least one of the tested feature selection methods. Final ranking of the species is obtained via calculating the median score among different populations.

There are not many studies in the literature on the relationship between enzymes and CRC, but some enzymes may influence risk factors for colorectal cancer or play a role in CRC. The number of these studies is not large enough and the relationship between enzymes and colorectal cancer is still under investigation. Therefore, our discussion of the accuracy of the representative biomarkers that are identified is based on a limited number of studies. Although the number of studies is limited, the enzyme biomarkers identified by the proposed approach highlight a number of enzymes associated with colorectal cancer, some of which have been validated by experiments in the literature. The top 20 enzymes obtained with the proposed approach are shown in Figure 5.8. In the "Discussion" section, the top 5 enzymes among these 20 species are analyzed in detail.

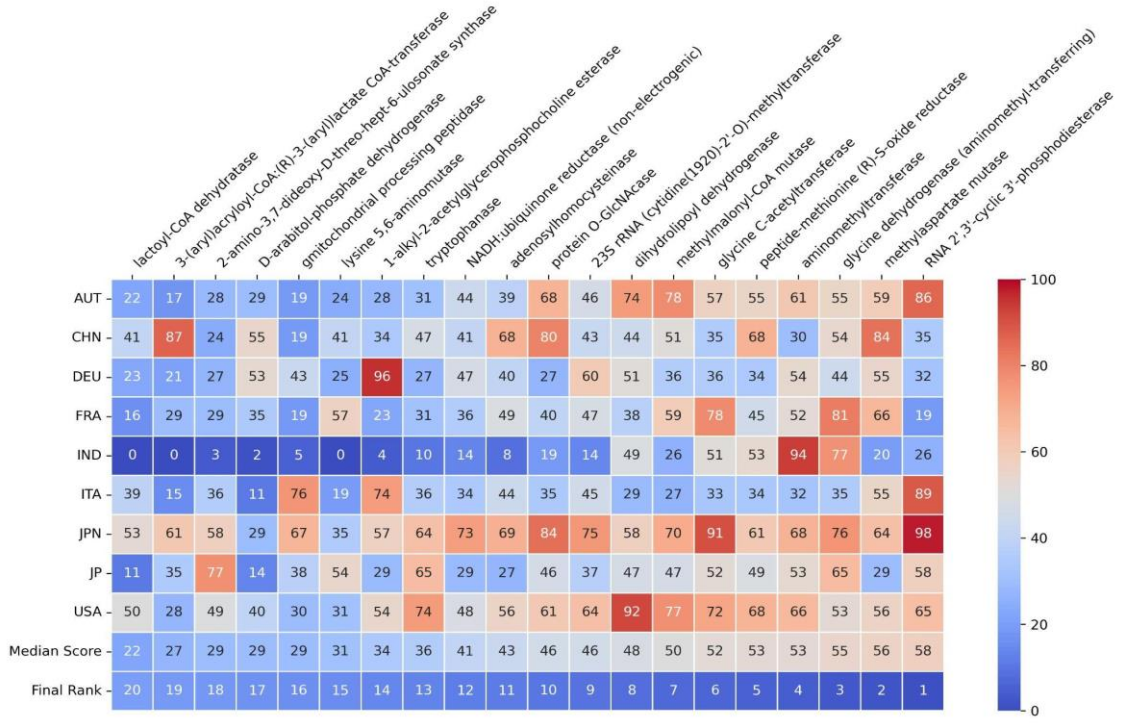


Figure 5.8 Heatmap of the feature importance scores for the identified enzymes in global analysis. The scores of the features in the common set are visualized in different rows for different populations. While higher scores are colored in red, lower scores are colored in blue. The union approach embraces the features that are identified in at least one of the tested feature selection methods. Final ranking of the species is obtained via calculating the median score among different populations.

There are several studies suggesting that pathways may influence risk factors associated with colorectal cancer or may play a role in the development of cancer. The number of these studies is insufficient and the investigation of the relationship between pathways and colorectal cancer is still ongoing. In the present study, the union features selected by the feature selection algorithms determine the usage importance determined by the population-specific classification algorithms, take the median of these rankings, and produce a final ranking. The top 20 pathways identified using the proposed approach are shown in Figure 5.9. In the "Discussion" section, the top 5 pathways among these 20 pathways are analyzed in detail.

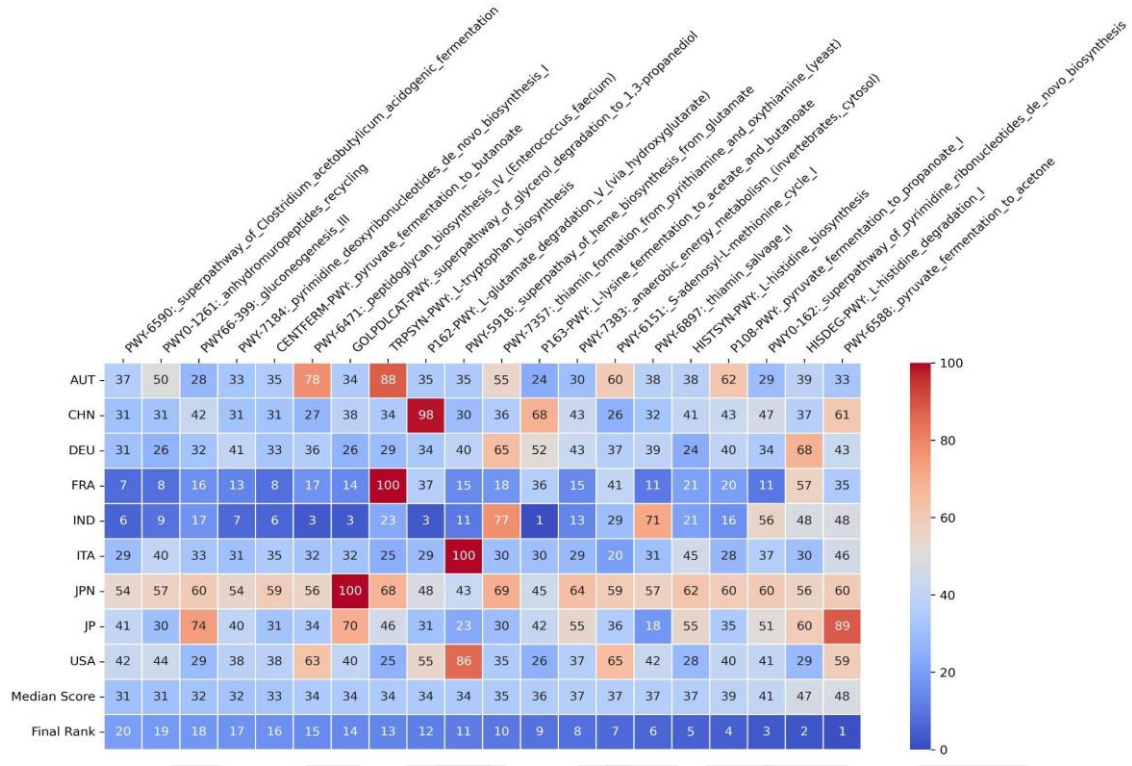


Figure 5.9 Heatmap of the feature importance scores for the identified pathways in global analysis. The scores of the features in the union set are visualized in different rows for different populations. While higher scores are colored in red, lower scores are colored in blue. The union approach embraces the features that are identified in at least one of the tested feature selection methods. Final ranking of the species is obtained via calculating the median score among different populations.

5.5 Discussion

In metagenome-based disease prediction, feature selection algorithms play a crucial role by helping to identify biomarkers at different molecular levels, which not only contributes to a more comprehensive understanding of disease mechanisms at the molecular level, but also has implications for diagnosis and treatment [32]. Our proposed method, using robust feature selection algorithms such as Select K Best (SKB), Information Gain (IG), and Extreme Gradient Boost (XGBoost), and utilizing intersection and union of the features that are selected in multiple feature selection methods, has the potential to streamline microbiota analysis, reduce costs, and improve the effectiveness of CRC diagnosis. The impact of these features on colorectal cancer classification is thoroughly evaluated using state-of-the-art machine learning algorithms, namely, Decision Tree (DT), Random Forest (RF), AdaBoost, and LogitBoost for various metagenomic datasets. Systematic evaluation of the developed models includes a set of performance metrics that ensures a comprehensive assessment of their effectiveness.

5.5.1 Biological Interpretation of the Findings

In recent years, an increasing number of studies have used metagenomics data to identify biomarkers for CRC. Although there is increasing evidence that the gut microbiota is associated with CRC, the collective role of the gut microbiota is still under investigation. In the present study, feature selection algorithms are applied to metagenomic data on species, enzymes, and pathways levels, investigating the effects of the intersection and union features selected by these algorithms on the predictive performance of colorectal cancer. By developing global and population-specific models, the metagenomic data (species, enzymes, and pathways) associated with CRC is comprehensively investigated. Based on the results obtained with the within-population approach, the impact of the union features is investigated from a biological perspective, as they exhibit high performance for the species and pathway datasets, identified by at least one of the feature selection algorithms. For the enzyme data, the impact of intersection features is investigated from a biological perspective, given their high performance across all feature selection algorithms. The potentially colorectal cancer-associated species, enzymes, and pathways identified by the proposed method are validated against previous studies in the related literature. The promising candidates are indicated as possible biomarkers for CRC.

The scores of the top 20 species among different populations and their final ranks are visualized in Figure 5.7. The top 5 important species in this final ranking list are searched in literature for their possible roles in CRC development. Among the top 5 species, each species identified by the proposed method is previously reported by the following studies as associated with colorectal cancer: Ruthenibacterium Lactatiformans [123], Parvimonas Micra [124] [125], Odoribacter Splanchnicus [126], Streptococcus Salivarius [127], and Gemella Morbillorum [128] [129]. Among other taxonomic biomarkers identified in the top 20 species list of the present study, several species (Alistipes Finegoldii [130], Peptostreptococcus stomatis [124], Lactobacillus fermentum [131]) have already been associated with colorectal cancer in the literature. This suggests that the proposed method is an effective method for species metagenomics data and can be used in future studies to investigate species data associated with colorectal cancer.

The scores of the top 20 enzymes among different populations and their final ranks are visualized in Figure 5.8. According to this ranking, the top 5 enzymes are RNA 2',3'-cyclic 3'-phosphodiesterase, methylaspartate mutase [42] [132], glycine dehydrogenase,

aminomethyltransferase [133], and peptide-methionine (R)-S- oxide reductase. Among these top 5 enzymes are two enzymes (methylaspartate mutase and aminomethyltransferase) that have been reported to be directly linked to colorectal cancer. Studies in literature suggest that other enzymes are indirectly associated with colorectal cancer. For example, in the following studies, dihydrolipoyl dehydrogenase [134], methylmalonyl CoA mutase [135] and tryptophanase [136] have not been directly associated with CRC, some analyses have been conducted on other structures with which these enzymes interact, and an association with colorectal cancer has been reported. The enzymes found by the proposed method other than those mentioned above may be inspiring for future studies to reveal the undiscovered relationships between colorectal cancer and candidate enzyme biomarkers.

The scores of the top 20 pathways among different populations and their final ranks are visualized in Figure 5.9. According to this ranking, the top 5 pathway features are PWY-6588: pyruvate fermentation to acetone, HISDEG-PWY: L-histidine degradation I [137], PWY0-162: superpathway of de novo biosynthesis of pyrimidine ribonucleotides, P108-PWY: pyruvate fermentation to propanoate I, and HISTSYN-PWY: L-histidine biosynthesis. Among these top 5 pathways, only one pathway (HISDEG-PWY: L-histidine degradation I) is reported as directly associated with colorectal cancer. In addition, among the top 20 pathways, three pathways, e.g. L-lysine fermentation to acetate and butanoate [138], thiamine salvage II [137] and the 1,3-propanediol-glycerol degradation superpathway [139], are directly associated with colorectal cancer. The pathways found by the proposed method, other than those mentioned above, may be useful for future studies to reveal the relationship between colorectal cancer and the candidate pathway biomarker.

The proposed method is used to predict colorectal cancer using machine learning methods and it identifies potential biomarkers associated with this disease. Research was conducted on 9 different datasets including data from 8 populations. Using CRC-associated metagenomic data on species, enzyme, and pathway levels; global and population-specific analyses were performed. Potential biomarkers identified by the proposed method are validated with different studies in the existing literature. These identified species, enzymes and pathways were suggested as potential biomarkers since the performance of the models developed using these features were superior to other models. As illustrated in Figures 5.7, 5.8, 5.9, the scores of the top 20 species, enzymes and pathways are calculated separately for each population, and median scores are

computed for species, enzyme, and pathway datasets. The top 5 species, enzymes and pathways in the final ranking were studied in detail in comparison to biomedical literature. For most of the species identified by our proposed approach, their possible roles in CRC development have been validated in literature, which highlights the effectiveness of our methodology. In the related literature, the number of studies on the association between enzymes and colorectal cancer is limited. One of the first 5 enzyme biomarkers have been previously reported in literature as associated with colorectal cancer. There are also relatively few studies on the association between pathway data and colorectal cancer. Research to identify pathways associated with colorectal cancer is ongoing. The fact that at least one biomarker from the top 5 biomarkers has been validated for each approach highlights the robustness of the proposed method, despite the limited number of studies available. This study not only provides valuable insights into colorectal cancer-associated species, enzymes, and pathways, but also serves as a guidepost for future research efforts aimed at uncovering previously unknown associations in this field.

Chapter 6

Conclusions and Future Prospects

6.1 Conclusions

In this thesis, using different types of biological data, machine learning methods are applied to comprehensively investigate the predictive performance for different disease types and identify possible biomarkers. In this context, the first study, called miRdisNET, predicts different cancer subgroups using miRNA data and identifies potential biomarkers for these diseases. The second study, called microBiomeGSM, uses metagenomic data to predict CRC at three different taxonomic levels and leverages the power of machine learning. In addition, the proposed approach identifies potential biomarkers associated with CRC. The third study, named CCPRED, performs global and population-dependent CRC prediction at different molecular levels using metagenomic data and identifies potential biomarkers. The experimental results obtained from these studies are described in detail below.

Understanding how miRNAs function on the cellular level provide valuable information for the diagnosis and treatment of human complex diseases. Precise identification of disease-miRNA relationships could accelerate diagnosis, prognosis, and drug development studies. Computational methods are playing an increasingly important role in predicting the potential relationship between disease and miRNA. Machine learning methods are widely used in studies to predict associations between miRNAs and diseases. In this article, we proposed a novel computational method named miRdisNET based on the G-S-M approach to identify associations between miRNAs and diseases. In this study, we developed a novel approach to explore miRNA-disease associations, detect biomarkers of disease-associated miRNAs. The novelty of miRdisNET is that it evaluates the performance of the model, reveals miRNA-disease associations. miRdisNET outperforms state-of-the-art methods with its model performance evaluation. It also identifies the relationships between miRNAs and diseases. In addition, it increases knowledge of disease associations, which can further improve approaches to disease diagnosis, prognosis, and treatment. The strength of miRdisNET is that it achieves high

success based on reliable machine learning methods, predicts possible disease-miRNA associations, and reveals important groups (disease and miRNA) and explores associations between diseases.

Over the past two decades, the number of microbiome studies has increased rapidly thanks to the advances in next generation sequencing (NGS) technologies. Lower costs and increasing computational power have enabled us to obtain enormous amounts of data on the diversity and function of a host or habitat's microbiome. Identifying and accounting for effective taxa in microbiome and disease classification can accelerate disease diagnosis, prognosis, and treatment. Here, we use an efficient machine learning model to identify taxonomic biomarkers that can diagnose diseases. The microBiomeGSM enables researchers to explore the diversity of contributions to disease development by examining metagenomic data at different taxonomic levels. While analyzing microbiome datasets, the microBiomeGSM tool that we present in this study exploits the existing biological knowledge about the taxonomic hierarchy of the species at different levels, such as genus, family, and order. Our results showed that via analyzing different microbiome datasets associated with different diseases, microBiomeGSM builds effective machine learning models to facilitate the diagnosis of diseases. It is anticipated that this thesis will be a guide for future studies and will guide and improve the studies to be conducted on this topic. With this study, we hope to highlight the importance of taxonomic groups in microbiome-based disease prediction and to facilitate the diagnosis of disease using these taxonomic groups.

Using various machine learning algorithms, this thesis identifies the species, enzyme groups and pathways that influence the microbiota of the colorectal cancer patients, hence contributing to the diagnosis and treatment of colorectal cancer. In addition, extensive experiments are performed to evaluate the effect of the identified species, enzyme, and pathway biomarkers on predicting the CRC at the population level. Our population-based analyses include with-in population analysis, cross population analysis (where one population is used as training data and another population is used as test data), and between population analysis (where one population is used as test data and the remaining populations are used as training data (leave one dataset out, LODO)). This comprehensive set of experiments reveals molecular groups (species, enzymes, and pathways) that are effective in CRC diagnosis, customized for populations. For CRC-associated species, enzyme, and pathway datasets, the models that are constructed using i) the union of features that are identified by at least one feature selection algorithm, ii)

the intersection of features identified by all feature selection algorithms, iii) all features are comparatively evaluated. The models that are generated by using union features show higher success rates compared with the success rates of the models that utilize all features and intersection features. Through our global and population-specific experiments, it is observed that XGBoost feature selection algorithm and Random Forest classification algorithm show superior performance compared with other feature selection and classification methods. The species, enzymes, and pathways that we have identified as potential biomarkers are confirmed by studies in the literature, enlightening their association with CRC. It is believed that the method that has been developed will make an important contribution to the future studies of CRC.

6.2 Societal Impact and Contribution to Global Sustainability

Nowadays, early diagnosis of disease is important both to halt the progression of many diseases and to enable effective treatment. Early diagnosis can prevent potential health problems and reduce the number of fatal diseases. This will lead to a significant reduction in healthcare costs. Due to the complex structure of human physiology and the large number and complexity of metabolic and environmental structures that cause diseases, early diagnosis of diseases using conventional methods is a difficult process and leads to errors. Therefore, advancing technology supports the development of effective systems to diagnose diseases and identify biomarkers, which should positively change the future of diseases.

In this thesis, 3 different types of diseases are analyzed, mainly cancer groups, T2D and IBD. In the miRdisNET study, 11 different types of cancer groups that belong to the cancer group are studied. In the microBiomeGSM study, IBD and T2D diseases are analyzed in addition to CRC), one of the cancer groups do not present in the first study. CCPRED, on the other hand, focused exclusively on CRC analyses and conducted global and population-based analyses. Diabetes mellitus (diabetes), which occurs when the glucose (sugar) level in the blood rises above normal and sugar is found in the urine, is one of the most common chronic diseases in the world and in Turkey. Type 2 diabetes mellitus (T2D) is a major global health problem and a major economic burden on healthcare systems. According to a report by the International Diabetes Federation, the

number of adults aged 20 to 79 with T2D worldwide was around 537 million in 2021 and is expected to rise to 783 million by 2045 [140]. The term "inflammatory bowel disease" (IBD) refers to a group of two diseases characterized by persistent inflammation of the gastrointestinal tract: Ulcerative colitis and Crohn's disease. An estimated 2.5to 3 million people in Europe suffer from IBD, which often leads to expenses for treatments such as hospitalization and surgery. This leads to annual costs of 4.6–5.6 billion euros in Europe. In addition, the impact on healthcare costs worldwide is expected to increase as the prevalence of IBD continues to rise [141]. Cancer has one of the highest mortality rates worldwide [142]. About 1-2 million deaths worldwide in 2019 were due to cancer, accounting for more than 25% of all deaths. After cardiovascular disease, cancer is the second most common cause of death in the European Union as a whole. Based on incidence trends before the COVID-19 pandemic, the Joint Research Center estimates that 2–7 million new cancer cases are expected in 2020; between 2020 and 2040, this number is expected to increase by 23%. The financial burden of cancer in Europe is already high; this expected increase will only add to it. In 2018, this burden amounted to €199 billion, representing 1-2% of total healthcare expenditure [143]. CRC is the third most common type of cancer worldwide. More than 1.9 million new cases of CRC and more than 930,000 deaths are estimated for 2020 [144].

This thesis is about predicting diseases and identifying potential biomarkers using machine learning methods for diseases with high mortality rates and high treatment budgets. The G-S-M approach introduced in miRdisNET and microBiomeGSM performs grouping-based feature selection. In this way, possible biomarkers are identified by removing unnecessary and irrelevant features from the data. CCPRED analyses use conventional feature selection methods. Therefore, the experiments performed with the tools and approaches developed for this work are focused on individualized treatments. With individualized treatments, diseases should be easier to detect and treat. The approaches developed for this purpose contribute to the early diagnosis, treatment, and development of the disease in healthcare systems. In addition, these models developed using artificial intelligence methods should contribute to the sustainability of healthcare systems.

6.3 Future Prospects

In this thesis, a total of three different studies on disease prediction and identification of possible biomarkers related to diseases. Among these studies, the miRdisNET and microBiomeGSM were published in the Frontiers media group, which has great importance in the field of health and has high quartiles. CCPRED was submitted to another journal with a high quartile. Studies on disease prediction and identification of potential biomarkers, which are still in their infancy, have gained momentum with the development of technology, but it is believed that they have not yet reached the desired level. The current methods and approaches to be applied in the studies presented in this thesis are briefly summarized below:

- ✓ Future prospects of miRdisNET

This study was mainly applied to cancer. Future studies are planned to apply it to common diseases with high mortality rates such as obesity, T2D and IBD. We also plan to apply modern methods instead of traditional machine learning approaches. Another planned approach is to use a smaller number of features by applying traditional feature selection algorithms instead of grouping-based feature selection.

- ✓ Future prospects of microBiomeGSM

With this study, we would also like to motivate biologists and the microbiome community to redesign their grouping methods instead of using individual feature selection approaches. We envision that in the future, various biological datasets, including multi-omics, will be used to redefine the groupings. Such innovative grouping strategies, complemented by modeling, promise to provide profound insights into the molecular mechanisms of diseases and the role of microorganisms in disease development.

- ✓ Future prospects of CCPRED

The success of the proposed method is one of the research projects with a high success rate among studies on the classification of CRC. However, the feature selection methods used to reduce the number of features is common feature selection algorithms from the literature. The current feature selection algorithms are used instead of these feature selection algorithms. In addition, this method is applied to different types of metagenomic data and different feature types are evaluated for the classification of CRC. This approach developed for CRC will also be applied to other disease types.

BIBLIOGRAPHY

[1] R. Wang, Z. Li, S. Liu, and D. Zhang, "Global, regional and national burden of inflammatory bowel disease in 204 countries and territories from 1990 to 2019: a systematic analysis based on the Global Burden of Disease Study 2019," *BMJ Open*, vol. 13, no. 3, p. e065186, Mar. 2023, doi: 10.1136/bmjopen-2022-065186.

[2] K. S. Lee, Y. Lee, and S.-G. Lee, "Alanine to glycine ratio is a novel predictive biomarker for type 2 diabetes mellitus," *Diabetes, Obesity and Metabolism*, vol. 26, no. 3, pp. 980–988, 2024, doi: 10.1111/dom.15395.

[3] "Cancer Tomorrow." Accessed: Mar. 21, 2024. [Online]. Available: https://gco.iarc.who.int/today/

[4] MunishKhanna, L. K. Singh, and H. Garg, "A novel approach for human diseases prediction using nature inspired computing & machine learning approach," *Multimed Tools Appl*, vol. 83, no. 6, pp. 17773–17809, Feb. 2024, doi: 10.1007/s11042-023-16236-6.

[5] F. Saeed *et al.*, "Enhancing Parkinson's Disease Prediction Using Machine Learning and Feature Selection Methods," *CMC*, vol. 71, no. 3, pp. 5639–5658, 2022, doi: 10.32604/cmc.2022.023124.

[6] I. Slimene, I. Messaoudi, A. Elloumi Oueslati, and Z. Lachiri, "Human disease prediction based on deep and machine learning classification of genes with miRNA binding sites," *Multimed Tools Appl*, Oct. 2023, doi: 10.1007/s11042-023-17457-5.

[7] B. Bakir-Gungor, H. Hacılar, A. Jabeer, O. U. Nalbantoglu, O. Aran, and M. Yousef, "Inflammatory bowel disease biomarkers of human gut microbiota selected via different feature selection methods," *PeerJ*, vol. 10, p. e13205, 2022, doi: 10.7717/peerj.13205.

[8] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Computers in Biology and Medicine*, vol. 112, p. 103375, Sep. 2019, doi: 10.1016/j.compbiomed.2019.103375.

[9] X. Chen, D. Xie, Q. Zhao, and Z.-H. You, "MicroRNAs and complex diseases: from experimental results to computational models," *Briefings in Bioinformatics*, vol. 20, no. 2, pp. 515–539, Mar. 2019, doi: 10.1093/bib/bbx130.

[10] X. Chen, Q.-F. Wu, and G.-Y. Yan, "RKNNMDA: Ranking-based KNN for MiRNA-Disease Association prediction," *RNA Biology*, vol. 14, no. 7, pp. 952–962, Jul. 2017, doi: 10.1080/15476286.2017.1312226.

[11] D. Yao, X. Zhan, and C.-K. Kwoh, "An improved random forest-based computational model for predicting novel miRNA-disease associations," *BMC Bioinformatics*, vol. 20, no. 1, p. 624, Dec. 2019, doi: 10.1186/s12859-019-3290-7.

[12] D. Liu, Y. Huang, W. Nie, J. Zhang, and L. Deng, "SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost," *BMC Bioinformatics*, vol. 22, no. 1, p. 219, Apr. 2021, doi: 10.1186/s12859-021-04135-2.

[13] X. Ding, J.-F. Xia, Y.-T. Wang, J. Wang, and C.-H. Zheng, "Improved Inductive Matrix Completion Method for Predicting MicroRNA-Disease Associations," in *Intelligent Computing Theories and Application*, D.-S. Huang, K.-H. Jo, and Z.-K. Huang, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 247–255. doi: 10.1007/978-3-030-26969-2_23.

[14] S. Zhou, S. Wang, Q. Wu, R. Azim, and W. Li, "Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression," *Computational Biology and Chemistry*, vol. 85, p. 107200, Apr. 2020, doi: 10.1016/j.compbiolchem.2020.107200.

[15] W. Liu *et al.*, "Identification of miRNA–disease associations via deep forest ensemble learning based on autoencoder," *Briefings in Bioinformatics*, vol. 23, no. 3, p. bbac104, May 2022, doi: 10.1093/bib/bbac104.

[16] K. U. Tüfekci, M. G. Öner, R. L. J. Meuwissen, and Ş. Genç, "The Role of MicroRNAs in Human Diseases," in *miRNomics: MicroRNA Biology and Computational Analysis*, M. Yousef and J. Allmer, Eds., in Methods in Molecular Biology. , Totowa, NJ: Humana Press, 2014, pp. 33–50. doi: 10.1007/978-1-62703-748-8_3.

[17] Q. Jiang *et al.*, "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, no. suppl_1, pp. D98–D104, Jan. 2009, doi: 10.1093/nar/gkn714.

[18] A. M. Ardekani and M. M. Naeini, "The Role of MicroRNAs in Human Diseases," *Avicenna J Med Biotechnol*, vol. 2, no. 4, pp. 161–179, 2010.

[19] T.-Y. Ha, "MicroRNAs in Human Diseases: From Cancer to Cardiovascular Disease," *Immune Netw*, vol. 11, no. 3, pp. 135–154, Jun. 2011, doi: 10.4110/in.2011.11.3.135.

[20] A. Lopez-Rincon, M. Martinez-Archundia, G. U. Martinez-Ruiz, A. Schoenhuth, and A. Tonda, "Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection," *BMC Bioinformatics*, vol. 20, no. 1, p. 480, Sep. 2019, doi: 10.1186/s12859-019-3050-8.

[21] Q. Huang *et al.*, "The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis," *Nat Cell Biol*, vol. 10, no. 2, Art. no. 2, Feb. 2008, doi: 10.1038/ncb1681.

[22] M. Gurung *et al.*, "Role of gut microbiota in type 2 diabetes pathophysiology," *eBioMedicine*, vol. 51, Jan. 2020, doi: 10.1016/j.ebiom.2019.11.051.

[23] J. A. Cena, L. G. Reis, A. K. A. de Lima, C. P. Vieira Lima, C. M. Stefani, and N. Dame-Teixeira, "Enrichment of Acid-Associated Microbiota in the Saliva of Type 2 Diabetes Mellitus Adults: A Systematic Review," *Pathogens*, vol. 12, no. 3, Art. no. 3, Mar. 2023, doi: 10.3390/pathogens12030404.

[24] R. Li, F. Shokri, A. L. Rincon, F. Rivadeneira, C. Medina-Gomez, and F. Ahmadizar, "Bi-Directional Interactions between Glucose-Lowering Medications and Gut Microbiome in Patients with Type 2 Diabetes Mellitus: A Systematic Review," *Genes*, vol. 14, no. 8, Art. no. 8, Aug. 2023, doi: 10.3390/genes14081572.

[25] R. Gao *et al.*, "Dysbiosis Signatures of Gut Microbiota Along the Sequence from Healthy, Young Patients to Those with Overweight and Obesity," *Obesity*, vol. 26, no. 2, pp. 351–361, 2018, doi: 10.1002/oby.22088.

[26] R. L. Negrut, A. Cote, and A. M. Maghiar, "Exploring the Potential of Oral Microbiome Biomarkers for Colorectal Cancer Diagnosis and Prognosis: A Systematic Review," *Microorganisms*, vol. 11, no. 6, Art. no. 6, Jun. 2023, doi: 10.3390/microorganisms11061586.

[27] F. H. Zwezerijnen-Jiwa, H. Sivov, P. Paizs, K. Zafeiropoulou, and J. Kinross, "A systematic review of microbiome-derived biomarkers for early colorectal cancer detection," *Neoplasia*, vol. 36, p. 100868, Feb. 2023, doi: 10.1016/j.neo.2022.100868.

[28] I. Huybrechts *et al.*, "The Human Microbiome in Relation to Cancer Risk: A Systematic Review of Epidemiologic Studies," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 29, no. 10, pp. 1856–1868, Oct. 2020, doi: 10.1158/1055-9965.EPI-20-0288.

[29] G. Tabowei *et al.*, "Microbiota Dysbiosis a Cause of Colorectal Cancer or Not? A Systematic Review," *Cureus*, vol. 14, no. 10, Oct. 2022, doi: 10.7759/cureus.30893.

[30] M. Hsu, K. M. Tun, K. Batra, L. Haque, T. Vongsavath, and A. S. Hong, "Safety and Efficacy of Fecal Microbiota Transplantation in Treatment of Inflammatory Bowel Disease in the Pediatric Population: A Systematic Review and Meta-Analysis," *Microorganisms*, vol. 11, no. 5, Art. no. 5, May 2023, doi: 10.3390/microorganisms11051272.

[31] C. Mah *et al.*, "Assessing the Relationship between the Gut Microbiota and Inflammatory Bowel Disease Therapeutics: A Systematic Review," *Pathogens*, vol. 12, no. 2, Art. no. 2, Feb. 2023, doi: 10.3390/pathogens12020262.

[32] N. LaPierre, C. J.-T. Ju, G. Zhou, and W. Wang, "MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction," *Methods*, vol. 166, pp. 74–82, Aug. 2019, doi: 10.1016/j.ymeth.2019.03.003.

[33] L. J. Marcos-Zambrano *et al.*, "Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment," *Frontiers in Microbiology*, vol. 12, 2021, Accessed: Oct. 20, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2021.634511

[34] H. Lim, F. Cankara, C.-J. Tsai, O. Keskin, R. Nussinov, and A. Gursoy, "Artificial intelligence approaches to human-microbiome protein–protein interactions," *Current Opinion in Structural Biology*, vol. 73, p. 102328, Apr. 2022, doi: 10.1016/j.sbi.2022.102328.

[35] H. Soueidan and M. Nikolski, "Machine learning for metagenomics: methods and tools." arXiv, Mar. 08, 2016. doi: 10.48550/arXiv.1510.06621.

[36] T. Deschênes, F. W. E. Tohoundjona, P.-L. Plante, V. Di Marzo, and F. Raymond, "Gene-based microbiome representation enhances host phenotype classification," *mSystems*, vol. 0, no. 0, pp. e00531-23, Jul. 2023, doi: 10.1128/msystems.00531-23.

[37] D. Sharma, A. D. Paterson, and W. Xu, "TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction," *Bioinformatics*, vol. 36, no. 17, pp. 4544–4550, Nov. 2020, doi: 10.1093/bioinformatics/btaa542.

[38] J. Qin *et al.*, "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, vol. 490, no. 7418, Art. no. 7418, Oct. 2012, doi: 10.1038/nature11450.

[39] N. Qin *et al.*, "Alterations of the human gut microbiome in liver cirrhosis," *Nature*, vol. 513, no. 7516, Art. no. 7516, Sep. 2014, doi: 10.1038/nature13568.

[40] R. Giliberti, S. Cavaliere, I. E. Mauriello, D. Ercolini, and E. Pasolli, "Host phenotype classification from human microbiome data is mainly driven by the presence of microbial taxa," *PLOS Computational Biology*, vol. 18, no. 4, p. e1010066, Apr. 2022, doi: 10.1371/journal.pcbi.1010066.

[41] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, "Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights," *PLOS Computational Biology*, vol. 12, no. 7, p. e1004977, Jul. 2016, doi: 10.1371/journal.pcbi.1004977.

[42] C. S. Casimiro-Soriguer, C. Loucera, M. Peña-Chilet, and J. Dopazo, "Towards a metagenomics machine learning interpretable model for understanding the transition from adenoma to colorectal cancer," *Sci Rep*, vol. 12, no. 1, p. 450, Jan. 2022, doi: 10.1038/s41598-021-04182-y.
[43] P. Li, H. Luo, B. Ji, and J. Nielsen, "Machine learning for data integration in human gut microbiome," *Microb Cell Fact*, vol. 21, no. 1, p. 241, Nov. 2022, doi: 10.1186/s12934-022-01973-4.
[44] S. Yachida *et al.*, "Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer," *Nat Med*, vol. 25, no. 6, pp. 968–976, Jun. 2019, doi: 10.1038/s41591-019-0458-7.
[45] H. Zhang, Y. Chang, Q. Zheng, R. Zhang, C. Hu, and W. Jia, "Altered intestinal microbiota associated with colorectal cancer," *Front. Med.*, vol. 13, no. 4, pp. 461–470, Aug. 2019, doi: 10.1007/s11684-019-0695-7.
[46] Q. Yao *et al.*, "Potential of fecal microbiota for detection and postoperative surveillance of colorectal cancer," *BMC Microbiol*, vol. 21, no. 1, p. 156, May 2021, doi: 10.1186/s12866-021-02182-6.
[47] F. Chen *et al.*, "Integrated analysis of the faecal metagenome and serum metabolome reveals the role of gut microbiome-associated metabolites in the detection of colorectal cancer and adenoma," *Gut*, vol. 71, no. 7, pp. 1315–1325, Jul. 2022, doi: 10.1136/gutjnl-2020-323476.
[48] H. Shuwen *et al.*, "Using whole-genome sequencing (WGS) to plot colorectal cancer-related gut microbiota in a population with varied geography," *Gut Pathog*, vol. 14, no. 1, p. 50, Dec. 2022, doi: 10.1186/s13099-022-00524-x.
[49] J. Yang *et al.*, "Development of a colorectal cancer diagnostic model and dietary risk assessment through gut microbiome analysis," *Exp Mol Med*, vol. 51, no. 10, Art. no. 10, Oct. 2019, doi: 10.1038/s12276-019-0313-4.
[50] Q. Feng *et al.*, "Gut microbiome development along the colorectal adenoma–carcinoma sequence," *Nat Commun*, vol. 6, no. 1, Art. no. 1, Mar. 2015, doi: 10.1038/ncomms7528.
[51] L. Zhou, Z. Jiang, Z. Zhang, J. Xing, D. Wang, and D. Tang, "Progress of gut microbiome and its metabolomics in early screening of colorectal cancer," *Clin Transl Oncol*, Feb. 2023, doi: 10.1007/s12094-023-03097-6.
[52] A. Dokht Khosravi, S. Seyed-Mohammadi, A. Teimoori, and A. Asarehzadegan Dezfuli, "The role of microbiota in colorectal cancer," *Folia Microbiol*, vol. 67, no. 5, pp. 683–691, Oct. 2022, doi: 10.1007/s12223-022-00978-1.
[53] X. Fan, Y. Jin, G. Chen, X. Ma, and L. Zhang, "Gut Microbiota Dysbiosis Drives the Development of Colorectal Cancer," *DIG*, vol. 102, no. 4, pp. 508–515, 2021, doi: 10.1159/000508328.
[54] A. Artemev *et al.*, "The Association of Microbiome Dysbiosis With Colorectal Cancer," *Cureus*, vol. 14, no. 2, Feb. 2022, doi: 10.7759/cureus.22156.
[55] L. J. Marcos-Zambrano *et al.*, "Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment," *Frontiers in Microbiology*, vol. 12, 2021, Accessed: Nov. 01, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2021.634511
[56] M. Bose *et al.*, "Analysis of an Indian colorectal cancer faecal microbiome collection demonstrates universal colorectal cancer-associated patterns, but closest correlation with other Indian cohorts," *BMC Microbiology*, vol. 23, no. 1, p. 52, Mar. 2023, doi: 10.1186/s12866-023-02805-0.

[57] J. Zhen *et al.*, "The global research of microbiota in colorectal cancer screening: a bibliometric and visualization analysis," *Frontiers in Oncology*, vol. 13, 2023, Accessed: Sep. 28, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2023.1169369

[58] C. Yu *et al.*, "Investigation of trends in gut microbiome associated with colorectal cancer using machine learning," *Frontiers in Oncology*, vol. 13, 2023, Accessed: Sep. 28, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2023.1077922

[59] B. Xie, Q. Ding, H. Han, and D. Wu, "miRCancer: a microRNA–cancer association database constructed by text mining on literature," *Bioinformatics*, vol. 29, no. 5, pp. 638–644, Mar. 2013, doi: 10.1093/bioinformatics/btt014.

[60] R. Mitra, C. M. Adams, W. Jiang, E. Greenawalt, and C. M. Eischen, "Pan-cancer analysis reveals cooperativity of both strands of microRNA that regulate tumorigenesis and patient survival," *Nat Commun*, vol. 11, no. 1, Art. no. 1, Feb. 2020, doi: 10.1038/s41467-020-14713-2.

[61] F. Beghini *et al.*, "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3," *eLife*, vol. 10, p. e65088, May 2021, doi: 10.7554/eLife.65088.

[62] A. Marco-Ramell *et al.*, "Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data," *BMC Bioinformatics*, vol. 19, no. 1, p. 1, Jan. 2018, doi: 10.1186/s12859-017-2006-0.

[63] A. M. Thomas *et al.*, "Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation," *Nat Med*, vol. 25, no. 4, Art. no. 4, Apr. 2019, doi: 10.1038/s41591-019-0405-7.

[64] G. Ditzler, R. Polikar, and G. Rosen, "Multi-Layer and Recursive Neural Networks for Metagenomic Classification," *IEEE Transactions on NanoBioscience*, vol. 14, no. 6, pp. 608–616, Sep. 2015, doi: 10.1109/TNB.2015.2461219.

[65] H. Alatawi *et al.*, "Attributes of intestinal microbiota composition and their correlation with clinical primary non-response to anti-TNF-α agents in inflammatory bowel disease patients," *Biomolecules and Biomedicine*, vol. 22, no. 3, Art. no. 3, Jun. 2022, doi: 10.17305/bjbms.2021.6436.

[66] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinform. Comput. Biol.*, vol. 03, no. 02, pp. 185–205, Apr. 2005, doi: 10.1142/S0219720005001004.

[67] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[68] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[69] C. Duvallet, S. M. Gibbons, T. Gurry, R. A. Irizarry, and E. J. Alm, "Meta-analysis of gut microbiome studies identifies disease-specific and shared responses," *Nat Commun*, vol. 8, no. 1, Art. no. 1, Dec. 2017, doi: 10.1038/s41467-017-01973-8.

[70] M. Oudah and A. Henschel, "Taxonomy-aware feature engineering for microbiome classification," *BMC Bioinformatics*, vol. 19, no. 1, p. 227, Jun. 2018, doi: 10.1186/s12859-018-2205-3.

[71] F. Fleuret and E. Ch, "Fast Binary Feature Selection with Conditional Mutual Information".

[72] B. Senliol, G. Gulgezen, L. Yu, and Z. Cataltepe, "Fast Correlation Based Filter (FCBF) with a different search strategy," in *2008 23rd International Symposium*

on Computer and Information Sciences, Oct. 2008, pp. 1–4. doi: 10.1109/ISCIS.2008.4717949.

[73] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *MACHINE LEARNING IN PYTHON*.

[74] B. Bakir-Gungor, O. Bulut, A. Jabeer, O. U. Nalbantoglu, and M. Yousef, "Discovering Potential Taxonomic Biomarkers of Type 2 Diabetes From Human Gut Microbiota via Different Feature Selection Methods," *Frontiers in Microbiology*, vol. 12, 2021, Accessed: Dec. 21, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2021.628426

[75] A. Jabeer *et al.*, "Identifying Taxonomic Biomarkers of Colorectal Cancer in Human Intestinal Microbiota Using Multiple Feature Selection Methods," in *2022 Innovations in Intelligent Systems and Applications Conference (ASYU)*, IEEE, 2022, pp. 1–6.

[76] B.-Q. Li, L.-L. Hu, L. Chen, K.-Y. Feng, Y.-D. Cai, and K.-C. Chou, "Prediction of Protein Domain with mRMR Feature Selection and Analysis," *PLOS ONE*, vol. 7, no. 6, p. e39308, Jun. 2012, doi: 10.1371/journal.pone.0039308.

[77] M. Toğaçar, B. Ergen, Z. Cömert, and F. Özyurt, "A Deep Feature Learning Model for Pneumonia Detection Applying a Combination of mRMR Feature Selection and Machine Learning Models," *IRBM*, vol. 41, no. 4, pp. 212–222, Aug. 2020, doi: 10.1016/j.irbm.2019.10.006.

[78] R. Alzubi, N. Ramzan, H. Alzoubi, and A. Amira, "A Hybrid Feature Selection Method for Complex Diseases SNPs," *IEEE Access*, vol. 6, pp. 1292–1301, 2018, doi: 10.1109/ACCESS.2017.2778268.

[79] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[80] Z. Jiang, J. Che, M. He, and F. Yuan, "A CGRU multi-step wind speed forecasting model based on multi-label specific XGBoost feature selection and secondary decomposition," *Renewable Energy*, vol. 203, pp. 802–827, Feb. 2023, doi: 10.1016/j.renene.2022.12.124.

[81] J. T. KENT, "Information gain and a general measure of correlation," *Biometrika*, vol. 70, no. 1, pp. 163–173, Apr. 1983, doi: 10.1093/biomet/70.1.163.

[82] M. Mera-Gaona, U. Neumann, R. Vargas-Canas, and D. M. López, "Evaluating the impact of multivariate imputation by MICE in feature selection," *PLOS ONE*, vol. 16, no. 7, p. e0254720, Jul. 2021, doi: 10.1371/journal.pone.0254720.

[83] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856–863. Accessed: Mar. 14, 2024. [Online]. Available: https://cdn.aaai.org/ICML/2003/ICML03-111.pdf

[84] A. F. Amiri, H. Oudira, A. Chouder, and S. Kichou, "Faults detection and diagnosis of PV systems based on machine learning approach using random forest classifier," *Energy Conversion and Management*, vol. 301, p. 118076, Feb. 2024, doi: 10.1016/j.enconman.2024.118076.

[85] A. Kumari, M. Akhtar, M. Tanveer, and M. Arshad, "Diagnosis of breast cancer using flexible pinball loss support vector machine," *Applied Soft Computing*, p. 111454, Mar. 2024, doi: 10.1016/j.asoc.2024.111454.

[86] D. Colledani, P. Anselmi, and E. Robusto, "Machine learning-decision tree classifiers in psychiatric assessment: An application to the diagnosis of major

depressive disorder," *Psychiatry Research*, vol. 322, p. 115127, Apr. 2023, doi: 10.1016/j.psychres.2023.115127.
[87] S. Ma, "Churn Prediction in Business Using Logistic Regression and Logit Boost," in *2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)*, Oct. 2023, pp. 363–366. doi: 10.1109/ICDSCA59871.2023.10392658.
[88] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771–780, p. 1612, 1999.
[89] R. Kolde, S. Laur, P. Adler, and J. Vilo, "Robust rank aggregation for gene list integration and meta-analysis," *Bioinformatics*, vol. 28, no. 4, pp. 573–580, Feb. 2012, doi: 10.1093/bioinformatics/btr709.
[90] Y. Yang *et al.*, "The role of microRNA in human lung squamous cell carcinoma," *Cancer Genetics and Cytogenetics*, vol. 200, no. 2, pp. 127–133, Jul. 2010, doi: 10.1016/j.cancergencyto.2010.03.014.
[91] V. Petkova *et al.*, "MiRNA expression profiling in adenocarcinoma and squamous cell lung carcinoma reveals both common and specific deregulated microRNAs," *Medicine (Baltimore)*, vol. 101, no. 33, p. e30027, Aug. 2022, doi: 10.1097/MD.0000000000030027.
[92] X.-W. Wang and Y.-Y. Liu, "Comparative study of classifiers for human microbiome data," *Med Microecol*, vol. 4, p. 100013, Jun. 2020, doi: 10.1016/j.medmic.2020.100013.
[93] M. R. Berthold *et al.*, "KNIME - the Konstanz information miner: version 2.0 and beyond," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 26–31, Nov. 2009, doi: 10.1145/1656274.1656280.
[94] M. Song, A. T. Chan, and J. Sun, "Influence of the Gut Microbiome, Diet, and Environment on Risk of Colorectal Cancer," *Gastroenterology*, vol. 158, no. 2, pp. 322–340, Jan. 2020, doi: 10.1053/j.gastro.2019.06.048.
[95] D. Salamon *et al.*, "Characteristics of the gut microbiota in adult patients with type 1 and 2 diabetes based on the analysis of a fragment of 16S rRNA gene using next-generation sequencing," *Polish Archives of Internal Medicine*, vol. 128, Apr. 2018, doi: 10.20452/pamw.4246.
[96] M. T. Alam, G. C. A. Amos, A. R. J. Murphy, S. Murch, E. M. H. Wellington, and R. P. Arasaradnam, "Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels," *Gut Pathogens*, vol. 12, no. 1, p. 1, Jan. 2020, doi: 10.1186/s13099-019-0341-6.
[97] M. Sedighi *et al.*, "Comparison of gut microbiota in adult patients with type 2 diabetes and healthy individuals," *Microbial Pathogenesis*, vol. 111, pp. 362–369, Oct. 2017, doi: 10.1016/j.micpath.2017.08.038.
[98] Y. Ni, C. Mu, X. He, K. Zheng, H. Guo, and W. Zhu, "Characteristics of gut microbiota and its response to a Chinese Herbal Formula in elder patients with metabolic syndrome," *Drug Discoveries & Therapeutics*, vol. 12, no. 3, pp. 161–169, 2018, doi: 10.5582/ddt.2018.01036.
[99] X. Li, J. Feng, Z. Wang, G. Liu, and F. Wang, "Features of combined gut bacteria and fungi from a Chinese cohort of colorectal cancer, colorectal adenoma, and post-operative patients," *Frontiers in Microbiology*, vol. 14, 2023, Accessed: Sep. 21, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1236583
[100] S. A.-D. Hassouneh, M. Loftus, and S. Yooseph, "Linking Inflammatory Bowel Disease Symptoms to Changes in the Gut Microbiome Structure and Function,"

Frontiers in Microbiology, vol. 12, 2021, Accessed: Sep. 20, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2021.673632

[101] Y. Zhang *et al.*, "Discovery of bioactive microbial gene products in inflammatory bowel disease," *Nature*, vol. 606, no. 7915, Art. no. 7915, Jun. 2022, doi: 10.1038/s41586-022-04648-7.

[102] X. Bai *et al.*, "Landscape of the gut archaeome in association with geography, ethnicity, urbanization, and diet in the Chinese population," *Microbiome*, vol. 10, no. 1, p. 147, Sep. 2022, doi: 10.1186/s40168-022-01335-7.

[103] K. Forslund *et al.*, "Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota," *Nature*, vol. 528, no. 7581, Art. no. 7581, Dec. 2015, doi: 10.1038/nature15766.

[104] Y. Ma *et al.*, "Metagenome Analysis of Intestinal Bacteria in Healthy People, Patients With Inflammatory Bowel Disease and Colorectal Cancer," *Frontiers in Cellular and Infection Microbiology*, vol. 11, 2021, Accessed: Sep. 20, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcimb.2021.599734

[105] F. Chen *et al.*, "Meta-analysis of fecal viromes demonstrates high diagnostic potential of the gut viral signatures for colorectal cancer and adenoma risk assessment," *Journal of Advanced Research*, vol. 49, pp. 103–114, Jul. 2023, doi: 10.1016/j.jare.2022.09.012.

[106] B. Bakir-Gungor, M. Temiz, A. Jabeer, D. Wu, and M. Yousef, "microBiomeGSM: the identification of taxonomic biomarkers from metagenomic data using grouping, scoring and modeling (G-S-M) approach," *Front Microbiol*, vol. 14, p. 1264941, Nov. 2023, doi: 10.3389/fmicb.2023.1264941.

[107] M. Yousef, J. Allmer, Y. İnal, and B. B. Gungor, "G-S-M: A Comprehensive Framework for Integrative Feature Selection in Omics Data Analysis and Beyond." Apr. 01, 2024. doi: 10.1101/2024.03.30.585514.

[108] M. Yousef, L. Abdallah, and J. Allmer, "maTE: discovering expressed interactions between microRNAs and their targets," *Bioinformatics*, vol. 35, no. 20, pp. 4020–4028, Oct. 2019, doi: 10.1093/bioinformatics/btz204.

[109] M. Yousef, F. Ozdemir, A. Jaaber, J. Allmer, and B. Bakir-Gungor, "PriPath: Identifying Dysregulated Pathways from Differential Gene Expression via Grouping, Scoring and Modeling with an Embedded Machine Learning Approach," In Review, preprint, Apr. 2022. doi: 10.21203/rs.3.rs-1449467/v1.

[110] E. Qumsiyeh, L. Showe, and M. Yousef, "GediNET for discovering gene associations across diseases using knowledge based machine learning approach," *Sci Rep*, vol. 12, no. 1, p. 19955, Nov. 2022, doi: 10.1038/s41598-022-24421-0.

[111] M. Yousef, G. Goy, R. Mitra, C. M. Eischen, A. Jabeer, and B. Bakir-Gungor, "miRcorrNet: machine learning-based integration of miRNA and mRNA expression profiles, combined with feature grouping and ranking," *PeerJ*, vol. 9, p. e11458, May 2021, doi: 10.7717/peerj.11458.

[112] M. Unlu Yazici, J. S. Marron, B. Bakir-Gungor, F. Zou, and M. Yousef, "Invention of 3Mint for feature grouping and scoring in multi-omics," *Front. Genet.*, vol. 14, p. 1093326, Mar. 2023, doi: 10.3389/fgene.2023.1093326.

[113] N. S. Ersoz, B. Bakir-Gungor, and M. Yousef, "GeNetOntology: Identifying Affected Gene Ontology Groups via Grouping, Scoring and Modelling from Gene Expression Data utilizing Biological Knowledge Based Machine Learning," *Frontiers in Genetics*, vol. 14, p. 1139082.

[114] M. Yousef and D. Voskergian, "TextNetTopics: Text Classification Based Word Grouping as Topics and Topics' Scoring," *Front. Genet.*, vol. 13, p. 893378, Jun. 2022, doi: 10.3389/fgene.2022.893378.
[115] D. Voskergian, B. Bakir-Gungor, and M. Yousef, "TextNetTopics Pro, a topic model-based text classification for short text by integration of semantic and document-topic distribution information," *Front. Genet.*, vol. 14, p. 1243874, Oct. 2023, doi: 10.3389/fgene.2023.1243874.
[116] E. Qumsiyeh, Z. Salah, and M. Yousef, "miRGediNET: A comprehensive examination of common genes in miRNA-Target interactions and disease associations: Insights from a grouping-scoring-modeling approach," *Heliyon*, vol. 9, no. 12, p. e22666, Dec. 2023, doi: 10.1016/j.heliyon.2023.e22666.
[117] A. Jabeer, M. Temiz, B. Bakir-Gungor, and M. Yousef, "miRdisNET: Discovering microRNA biomarkers that are associated with diseases utilizing biological knowledge-based machine learning," *Front. Genet.*, vol. 13, p. 1076554, Jan. 2023, doi: 10.3389/fgene.2022.1076554.
[118] M. Yousef, G. Goy, and B. Bakir-Gungor, "miRModuleNet: Detecting miRNA-mRNA Regulatory Modules," *Front. Genet.*, vol. 13, p. 767455, Apr. 2022, doi: 10.3389/fgene.2022.767455.
[119] M. Yousef, E. Ülgen, and O. Uğur Sezerman, "CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis," *PeerJ Computer Science*, vol. 7, p. e336, Feb. 2021, doi: 10.7717/peerj-cs.336.
[120] Ü. G. Söylemez, M. Yousef, and B. Bakir-Gungor, "AMP-GSM: Prediction of Antimicrobial Peptides via a Grouping–Scoring–Modeling Approach," *Applied Sciences*, vol. 13, no. 8, p. 5106, Apr. 2023, doi: 10.3390/app13085106.
[121] C. Kuzudisli, B. Bakir-Gungor, N. Bulut, B. Qaqish, and M. Yousef, "Review of feature selection approaches based on grouping of features," *PeerJ*, vol. 11, p. e15666, Jul. 2023, doi: 10.7717/peerj.15666.
[122] M. Yousef, A. Kumar, and B. Bakir-Gungor, "Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data," *Entropy*, vol. 23, no. 1, 2021, doi: 10.3390/e23010002.
[123] N. Trivieri *et al.*, "BRAFV600E mutation impinges on gut microbial markers defining novel biomarkers for serrated colorectal cancer effective therapies," *J Exp Clin Cancer Res*, vol. 39, no. 1, p. 285, Dec. 2020, doi: 10.1186/s13046-020-01801-w.
[124] M. A. Osman *et al.*, "Parvimonas micra, Peptostreptococcus stomatis, Fusobacterium nucleatum and Akkermansia muciniphila as a four-bacteria biomarker panel of colorectal cancer," *Sci Rep*, vol. 11, no. 1, Art. no. 1, Feb. 2021, doi: 10.1038/s41598-021-82465-0.
[125] L. Zhao *et al.*, "Parvimonas micra promotes colorectal tumorigenesis and is associated with prognosis of colorectal cancer patients," *Oncogene*, vol. 41, no. 36, Art. no. 36, Sep. 2022, doi: 10.1038/s41388-022-02395-7.
[126] C.-W. Png, Y.-K. Chua, J.-H. Law, Y. Zhang, and K.-K. Tan, "Alterations in co-abundant bacteriome in colorectal cancer and its persistence after surgery: a pilot study," *Sci Rep*, vol. 12, no. 1, Art. no. 1, Jun. 2022, doi: 10.1038/s41598-022-14203-z.
[127] C. C. Wong and J. Yu, "Gut microbiota in colorectal cancer development and therapy," *Nat Rev Clin Oncol*, pp. 1–24, May 2023, doi: 10.1038/s41571-023-00766-x.

[128] K. B. Laupland, F. Edwards, L. Furuya-Kanamori, D. L. Paterson, and P. N. A. Harris, "Bloodstream infection and colorectal cancer risk in Queensland Australia, 2000-2019," *The American Journal of Medicine*, May 2023, doi: 10.1016/j.amjmed.2023.05.003.
[129] Y. Shimomura *et al.*, "Mediation effect of intestinal microbiota on the relationship between fiber intake and colorectal cancer," *International Journal of Cancer*, vol. 152, no. 9, pp. 1752–1762, 2023, doi: 10.1002/ijc.34398.
[130] B. J. Parker, P. A. Wearsch, A. C. M. Veloo, and A. Rodriguez-Palacios, "The Genus Alistipes: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health," *Front. Immunol.*, vol. 11, Jun. 2020, doi: 10.3389/fimmu.2020.00906.
[131] J.-E. Lee *et al.*, "Characterization of the Anti-Cancer Activity of the Probiotic Bacterium Lactobacillus fermentum Using 2D vs. 3D Culture in Colorectal Cancer Cells," *Biomolecules*, vol. 9, no. 10, Art. no. 10, Oct. 2019, doi: 10.3390/biom9100557.
[132] W. Gou *et al.*, "Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features Associated With Type 2 Diabetes," *Diabetes Care*, vol. 44, no. 2, pp. 358–366, Dec. 2020, doi: 10.2337/dc20-1536.
[133] E. Odin, A. Sondén, G. Carlsson, B. Gustavsson, and Y. Wettergren, "Folate pathway genes linked to mitochondrial biogenesis and respiration are associated with outcome of patients with stage III colorectal cancer," *Tumour Biol.*, vol. 41, no. 6, p. 1010428319846231, Jun. 2019, doi: 10.1177/1010428319846231.
[134] P. J. Lee, S. J. Woo, H. M. Yoo, N. Cho, and H. P. Kim, "Differential Mechanism of ATP Production Occurs in Response to Succinylacetone in Colon Cancer Cells," *Molecules*, vol. 24, no. 19, Art. no. 19, Jan. 2019, doi: 10.3390/molecules24193575.
[135] V. Lacombe, G. Lenaers, and G. Urbanski, "Diagnostic and Therapeutic Perspectives Associated to Cobalamin-Dependent Metabolism and Transcobalamins' Synthesis in Solid Cancers," *Nutrients*, vol. 14, no. 10, Art. no. 10, Jan. 2022, doi: 10.3390/nu14102058.
[136] M. Wyatt and K. L. Greathouse, "Targeting Dietary and Microbial Tryptophan-Indole Metabolism as Therapeutic Approaches to Colon Cancer," *Nutrients*, vol. 13, no. 4, Art. no. 4, Apr. 2021, doi: 10.3390/nu13041189.
[137] J.-W. Huh *et al.*, "Enterotypical Prevotella and three novel bacterial biomarkers in preoperative stool predict the clinical outcome of colorectal cancer," *Microbiome*, vol. 10, no. 1, p. 203, Nov. 2022, doi: 10.1186/s40168-022-01388-8.
[138] E. Russo *et al.*, "From adenoma to CRC stages: the oral-gut microbiome axis as a source of potential microbial and metabolic biomarkers of malignancy," *Neoplasia*, vol. 40, p. 100901, Jun. 2023, doi: 10.1016/j.neo.2023.100901.
[139] F. Bellerba *et al.*, "Colorectal cancer, Vitamin D and microbiota: A double-blind Phase II randomized trial (ColoViD) in colorectal cancer patients," *Neoplasia*, vol. 34, p. 100842, Dec. 2022, doi: 10.1016/j.neo.2022.100842.
[140] P.-S. Yu *et al.*, "Association Between Trimethylamine N-oxide and Adverse Kidney Outcomes and Overall Mortality in Type 2 Diabetes Mellitus," *The Journal of Clinical Endocrinology & Metabolism*, p. dgae009, Jan. 2024, doi: 10.1210/clinem/dgae009.
[141] M. Barreiro-de Acosta *et al.*, "Epidemiological, Clinical, Patient-Reported and Economic Burden of Inflammatory Bowel Disease (Ulcerative colitis and Crohn's disease) in Spain: A Systematic Review," *Adv Ther*, vol. 40, no. 5, pp. 1975–2014, May 2023, doi: 10.1007/s12325-023-02473-6.

[142] N. Azamjah, Y. Soltan-Zadeh, and F. Zayeri, "Global Trend of Breast Cancer Mortality Rate: A 25-Year Study," *Asian Pac J Cancer Prev*, vol. 20, no. 7, pp. 2015–2020, 2019, doi: 10.31557/APJCP.2019.20.7.2015.
[143] C. Berchet, G. Dedet, N. Klazinga, and F. Colombo, "Inequalities in cancer prevention and care across Europe," *The Lancet Oncology*, vol. 24, no. 1, pp. 10–11, Jan. 2023, doi: 10.1016/S1470-2045(22)00746-X.
[144] E. Morgan *et al.*, "Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN," *Gut*, vol. 72, no. 2, pp. 338–344, Feb. 2023, doi: 10.1136/gutjnl-2022-327736.
[145] X. Wu and S. Park, "Fecal Bacterial Community and Metagenome Function in Asians with Type 2 Diabetes, According to Enterotypes," Biomedicines, vol. 10, no. 11, Art. no. 11, Nov. 2022, doi: 10.3390/biomedicines10112998.
[146] A. O. Afolayan, L. A. Adebusoye, E. O. Cadmus, and F. A. Ayeni, "Insights into the gut microbiota of Nigerian elderly with type 2 diabetes and non-diabetic elderly persons," *Heliyon*, vol. 6, no. 5, p. e03971, May 2020, doi: 10.1016/j.heliyon.2020.e03971.
[147] P. Therdtatha *et al.*, "Gut Microbiome of Indonesian Adults Associated with Obesity and Type 2 Diabetes: A Cross-Sectional Study in an Asian City, Yogyakarta," *Microorganisms*, vol. 9, no. 5, Art. no. 5, May 2021, doi: 10.3390/microorganisms9050897.
[148] A. Metwaly, S. Reitmeier, and D. Haller, "Microbiome risk profiles as biomarkers for inflammatory and metabolic disorders," *Nat Rev Gastroenterol Hepatol*, vol. 19, no. 6, Art. no. 6, Jun. 2022, doi: 10.1038/s41575-022-00581-2.
[149] Z. Khudhair *et al.*, "Administration of Hookworm Excretory/Secretory Proteins Improves Glucose Tolerance in a Mouse Model of Type 2 Diabetes," *Biomolecules*, vol. 12, no. 5, Art. no. 5, May 2022, doi: 10.3390/biom12050637.
[150] Z. Zhao *et al.*, "Myricetin relieves the symptoms of type 2 diabetes mice and regulates intestinal microflora," *Biomedicine & Pharmacotherapy*, vol. 153, p. 113530, Sep. 2022, doi: 10.1016/j.biopha.2022.113530.
[151] C. Wang *et al.*, "Uygur type 2 diabetes patient fecal microbiota transplantation disrupts blood glucose and bile acid levels by changing the ability of the intestinal flora to metabolize bile acids in C57BL/6 mice," *BMC Endocrine Disorders*, vol. 22, no. 1, p. 236, Sep. 2022, doi: 10.1186/s12902-022-01155-8.
[152] X. Du *et al.*, "Alteration of gut microbial profile in patients with diabetic nephropathy," *Endocrine*, vol. 73, no. 1, pp. 71–84, Jul. 2021, doi: 10.1007/s12020-021-02721-1.
[153] C. Geisler *et al.*, "Gut microbiome alterations are differentially associated with hand grip strength and type 2 diabetes mellitus.," in *Diabetologie und Stoffwechsel*, Georg Thieme Verlag, Apr. 2023, p. P 143. doi: 10.1055/s-0043-1768005.
[154] G. Gradisteanu Pircalabioru, M.-C. Chifiriuc, A. Picu, L. M. Petcu, M. Trandafir, and O. Savu, "Snapshot into the Type-2-Diabetes-Associated Microbiome of a Romanian Cohort," *International Journal of Molecular Sciences*, vol. 23, no. 23, Art. no. 23, Jan. 2022, doi: 10.3390/ijms232315023.
[155] K.-A. Lê *et al.*, "Alterations in fecal Lactobacillus and Bifidobacterium species in type 2 diabetic patients in Southern China population," *Frontiers in Physiology*, vol. 3, 2013, Accessed: Jul. 04, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphys.2012.00496
[156] K. Hosomi *et al.*, "Oral administration of Blautia wexlerae ameliorates obesity and type 2 diabetes via metabolic remodeling of the gut microbiota," *Nat Commun*, vol. 13, no. 1, Art. no. 1, Aug. 2022, doi: 10.1038/s41467-022-32015-7.

[157] Q. Li, Y. Chang, K. Zhang, H. Chen, S. Tao, and Z. Zhang, "Implication of the gut microbiome composition of type 2 diabetic patients from northern China," *Sci Rep*, vol. 10, no. 1, Art. no. 1, Mar. 2020, doi: 10.1038/s41598-020-62224-3.
[158] Y. Nam *et al.*, "Heat-Killed Lactiplantibacillus plantarum LRCC5314 Mitigates the Effects of Stress-Related Type 2 Diabetes in Mice via Gut Microbiome Modulation," vol. 32, no. 3, pp. 324–332, Mar. 2022, doi: 10.4014/jmb.2111.11008.
[159] R. Lin *et al.*, "Gut Microbiota Mediate Melatonin Signaling in Association With Type 2 Diabetes," *Current Developments in Nutrition*, vol. 6, p. 1019, Jun. 2022, doi: 10.1093/cdn/nzac069.024.
[160] A. P. Doumatey *et al.*, "Gut Microbiome Profiles Are Associated With Type 2 Diabetes in Urban Africans," *Frontiers in Cellular and Infection Microbiology*, vol. 10, 2020, Accessed: Jul. 04, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcimb.2020.00063
[161] D. Neri-Rosario *et al.*, "Dysbiosis signatures of gut microbiota and the progression of type 2 diabetes: a machine learning approach in a Mexican cohort," *Frontiers in Endocrinology*, vol. 14, Jun. 2023, doi: 10.3389/fendo.2023.1170459.
[162] M. Liu *et al.*, "Oxymatrine ameliorated experimental colitis via mechanisms involving inflammatory DCs, gut microbiota and TLR/NF-κB pathway," *International Immunopharmacology*, vol. 115, p. 109612, Feb. 2023, doi: 10.1016/j.intimp.2022.109612.
[163] M. F. Neurath, "Host–microbiota interactions in inflammatory bowel disease," *Nat Rev Gastroenterol Hepatol*, vol. 17, no. 2, Art. no. 2, Feb. 2020, doi: 10.1038/s41575-019-0248-1.
[164] A. Carstens *et al.*, "The Gut Microbiota in Collagenous Colitis Shares Characteristics With Inflammatory Bowel Disease-Associated Dysbiosis," *Clin Transl Gastroenterol*, vol. 10, no. 7, p. e00065, Jul. 2019, doi: 10.14309/ctg.0000000000000065.
[165] A. Teofani *et al.*, "Intestinal Taxa Abundance and Diversity in Inflammatory Bowel Disease Patients: An Analysis including Covariates and Confounders," *Nutrients*, vol. 14, no. 2, Art. no. 2, Jan. 2022, doi: 10.3390/nu14020260.
[166] S. Stojanov, A. Berlec, and B. Štrukelj, "The Influence of Probiotics on the Firmicutes/Bacteroidetes Ratio in the Treatment of Obesity and Inflammatory Bowel disease," *Microorganisms*, vol. 8, no. 11, Art. no. 11, Nov. 2020, doi: 10.3390/microorganisms8111715.
[167] D. A. Muñiz Pedrogo *et al.*, "An Increased Abundance of Clostridiaceae Characterizes Arthritis in Inflammatory Bowel Disease and Rheumatoid Arthritis: A Cross-sectional Study," *Inflammatory Bowel Diseases*, vol. 25, no. 5, pp. 902–913, Apr. 2019, doi: 10.1093/ibd/izy318.
[168] C. Colquhoun, M. Duncan, and G. Grant, "Inflammatory Bowel Diseases: Host-Microbial-Environmental Interactions in Dysbiosis," *Diseases*, vol. 8, no. 2, Art. no. 2, Jun. 2020, doi: 10.3390/diseases8020013.
[169] N. L. Zitomersky *et al.*, "Characterization of Adherent Bacteroidales from Intestinal Biopsies of Children and Young Adults with Inflammatory Bowel Disease," *PLOS ONE*, vol. 8, no. 6, p. e63686, Jun. 2013, doi: 10.1371/journal.pone.0063686.
[170] H. Tye *et al.*, "NLRP1 restricts butyrate producing commensals to exacerbate inflammatory bowel disease," *Nat Commun*, vol. 9, no. 1, Art. no. 1, Sep. 2018, doi: 10.1038/s41467-018-06125-0.

[171] C. M. Rands, H. Brüssow, and E. M. Zdobnov, "Comparative genomics groups phages of Negativicutes and classical Firmicutes despite different Gram-staining properties," *Environmental Microbiology*, vol. 21, no. 11, pp. 3989–4001, 2019, doi: 10.1111/1462-2920.14746.
[172] A. N. Ananthakrishnan *et al.*, "Gut Microbiome Function Predicts Response to Anti-integrin Biologic Therapy in Inflammatory Bowel Diseases," *Cell Host & Microbe*, vol. 21, no. 5, pp. 603-610.e3, May 2017, doi: 10.1016/j.chom.2017.04.010.
[173] R. M. Shobar *et al.*, "The Effects of Bowel Preparation on Microbiota-Related Metrics Differ in Health and in Inflammatory Bowel Disease and for the Mucosal and Luminal Microbiota Compartments," *Clin Transl Gastroenterol*, vol. 7, no. 2, p. e143, Feb. 2016, doi: 10.1038/ctg.2015.54.
[174] S. M. Bloom *et al.*, "Commensal Bacteroides Species Induce Colitis in Host-Genotype-Specific Fashion in a Mouse Model of Inflammatory Bowel Disease," *Cell Host & Microbe*, vol. 9, no. 5, pp. 390–403, May 2011, doi: 10.1016/j.chom.2011.04.009.
[175] X. Zhang, Y. Tong, X. Lyu, J. Wang, Y. Wang, and R. Yang, "Prevention and Alleviation of Dextran Sulfate Sodium Salt-Induced Inflammatory Bowel Disease in Mice With Bacillus subtilis-Fermented Milk via Inhibition of the Inflammatory Responses and Regulation of the Intestinal Flora," *Frontiers in Microbiology*, vol. 11, 2021, Accessed: Jan. 06, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2020.622354
[176] T. Zhang *et al.*, "[Changes of fecal flora and its correlation with inflammatory indicators in patients with inflammatory bowel disease]," *Nan Fang Yi Ke Da Xue Xue Bao*, vol. 33, no. 10, pp. 1474–1477, Oct. 2013.
[177] H. Faden, "The Role of Faecalibacterium, Roseburia, and Butyrate in Inflammatory Bowel Disease," *DDI*, vol. 40, no. 6, pp. 793–795, 2022, doi: 10.1159/000522247.
[178] M. Kverka *et al.*, "Oral administration of Parabacteroides distasonis antigens attenuates experimental murine colitis through modulation of immunity and microbiota composition," *Clinical and Experimental Immunology*, vol. 163, no. 2, pp. 250–259, Feb. 2011, doi: 10.1111/j.1365-2249.2010.04286.x.
[179] M. Schirmer, A. Garner, H. Vlamakis, and R. J. Xavier, "Microbial genes and pathways in inflammatory bowel disease," *Nat Rev Microbiol*, vol. 17, no. 8, Art. no. 8, Aug. 2019, doi: 10.1038/s41579-019-0213-6.
[180] K. A. Shaw *et al.*, "Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease," *Genome Medicine*, vol. 8, no. 1, p. 75, Jul. 2016, doi: 10.1186/s13073-016-0331-y.
[181] K. Nishino *et al.*, "Analysis of endoscopic brush samples identified mucosa-associated dysbiosis in inflammatory bowel disease," *J Gastroenterol*, vol. 53, no. 1, pp. 95–106, Jan. 2018, doi: 10.1007/s00535-017-1384-4.
[182] M. Issa, A. N. Ananthakrishnan, and D. G. Binion, "Clostridium difficile and inflammatory bowel disease," *Inflammatory Bowel Diseases*, vol. 14, no. 10, pp. 1432–1442, Oct. 2008, doi: 10.1002/ibd.20500.
[183] X. Sui *et al.*, "The relationship between KRAS gene mutation and intestinal flora in tumor tissues of colorectal cancer patients," *Annals of Translational Medicine*, vol. 8, no. 17, Art. no. 17, Sep. 2020, doi: 10.21037/atm-20-5622.

[184] S. Han *et al.*, "Intestinal microorganisms involved in colorectal cancer complicated with dyslipidosis," *Cancer Biology & Therapy*, vol. 20, no. 1, pp. 81–89, Jan. 2019, doi: 10.1080/15384047.2018.1507255.
[185] A. A. Alhhazmi *et al.*, "Gut Microbial and Associated Metabolite Markers for Colorectal Cancer Diagnosis," *Microorganisms*, vol. 11, no. 8, Art. no. 8, Aug. 2023, doi: 10.3390/microorganisms11082037.
[186] H. Li, D. Sheng, C. Jin, G. Zhao, and L. Zhang, "Identifying and ranking causal microbial biomarkers for colorectal cancer at different cancer subsites and stages: a Mendelian randomization study," *Frontiers in Oncology*, vol. 13, 2023, Accessed: Sep. 20, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2023.1224705
[187] H. Yu *et al.*, "Fecal microbiota transplantation inhibits colorectal cancer progression: Reversing intestinal microbial dysbiosis to enhance anti-cancer immune responses," *Frontiers in Microbiology*, vol. 14, 2023, Accessed: Sep. 20, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1126808
[188] Y. Yang *et al.*, "Prospective study of oral microbiome and colorectal cancer risk in low-income and African American populations," *International Journal of Cancer*, vol. 144, no. 10, pp. 2381–2389, 2019, doi: 10.1002/ijc.31941.
[189] O. Phipps *et al.*, "Oral and Intravenous Iron Therapy Differentially Alter the On- and Off-Tumor Microbiota in Anemic Colorectal Cancer Patients," *Cancers*, vol. 13, no. 6, Art. no. 6, Jan. 2021, doi: 10.3390/cancers13061341.
[190] C. Hatcher *et al.*, "Application of Mendelian randomization to explore the causal role of the human gut microbiome in colorectal cancer," *Sci Rep*, vol. 13, no. 1, Art. no. 1, Apr. 2023, doi: 10.1038/s41598-023-31840-0.
[191] L. Sun *et al.*, "The difference of human gut microbiome in colorectal cancer with and without metastases," *Frontiers in Oncology*, vol. 12, 2022, Accessed: Sep. 20, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2022.982744
[192] H. Zhang *et al.*, "Microbiome analysis reveals universal diagnostic biomarkers for colorectal cancer across populations and technologies," *Frontiers in Microbiology*, vol. 13, 2022, Accessed: Sep. 20, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2022.1005201
[193] M. W. Dougherty and C. Jobin, "Intestinal bacteria and colorectal cancer: etiology and treatment," *Gut Microbes*, vol. 15, no. 1, p. 2185028, Dec. 2023, doi: 10.1080/19490976.2023.2185028.
[194] J. Zhang *et al.*, "Expansion of Colorectal Cancer Biomarkers Based on Gut Bacteria and Viruses," *Cancers*, vol. 14, no. 19, Art. no. 19, Jan. 2022, doi: 10.3390/cancers14194662.
[195] S. Rezasoltani *et al.*, "Oral Microbiota as Novel Biomarkers for Colorectal Cancer Screening," *Cancers*, vol. 15, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/cancers15010192.
[196] A. Elkholy *et al.*, "Microbiome diversity in African American, European American, and Egyptian colorectal cancer patients," *Heliyon*, vol. 9, no. 7, p. e18035, Jul. 2023, doi: 10.1016/j.heliyon.2023.e18035.
[197] M. Gutierrez-Angulo, M. de la L. Ayala-Madrigal, J. M. Moreno-Ortiz, J. Peregrina-Sandoval, and F. D. Garcia-Ayala, "Microbiota composition and its impact on DNA methylation in colorectal cancer," *Frontiers in Genetics*, vol. 14, 2023, Accessed: Sep. 20, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fgene.2023.1037406

[198] N. Avuthu and C. Guda, "Meta-Analysis of Altered Gut Microbiota Reveals Microbial and Metabolic Biomarkers for Colorectal Cancer," *Microbiology Spectrum*, vol. 10, no. 4, pp. e00013-22, Jun. 2022, doi: 10.1128/spectrum.00013-22.

[199] A. R. Bourgonje *et al.*, "Patients With Inflammatory Bowel Disease Show IgG Immune Responses Towards Specific Intestinal Bacterial Genera," *Frontiers in Immunology*, vol. 13, 2022, Accessed: Sep. 19, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fimmu.2022.842911

[200] A. Volkova and K. V. Ruggles, "Predictive Metagenomic Analysis of Autoimmune Disease Identifies Robust Autoimmunity and Disease Specific Microbial Signatures," *Frontiers in Microbiology*, vol. 12, 2021, Accessed: Sep. 20, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2021.621310

[201] Y. E. Park *et al.*, "Microbial changes in stool, saliva, serum, and urine before and after anti-TNF-α therapy in patients with inflammatory bowel diseases," *Sci Rep*, vol. 12, no. 1, Art. no. 1, Apr. 2022, doi: 10.1038/s41598-022-10450-2.

[202] S. Cheng *et al.*, "Altered gut microbiome in FUT2 loss-of-function mutants in support of personalized medicine for inflammatory bowel diseases," *Journal of Genetics and Genomics*, vol. 48, no. 9, pp. 771–780, Sep. 2021, doi: 10.1016/j.jgg.2021.08.003.

[203] Y. Zhou *et al.*, "Gut Microbiota Offers Universal Biomarkers across Ethnicity in Inflammatory Bowel Disease Diagnosis and Infliximab Response Prediction," *mSystems*, vol. 3, no. 1, p. 10.1128/msystems.00188-17, Jan. 2018, doi: 10.1128/msystems.00188-17.

[204] D. E. Bosch, R. Abbasian, B. Parajuli, S. B. Peterson, and J. D. Mougous, "Structural disruption of Ntox15 nuclease effector domains by immunity proteins protects against type VI secretion system intoxication in Bacteroidales," *mBio*, vol. 14, no. 4, pp. e01039-23, Jun. 2023, doi: 10.1128/mbio.01039-23.

[205] Y. Chen *et al.*, "Exploiting lactic acid bacteria for inflammatory bowel disease: A recent update," *Trends in Food Science & Technology*, vol. 138, pp. 126–140, Aug. 2023, doi: 10.1016/j.tifs.2023.06.007.

[206] J. Zhang, Y. Guo, and L. Duan, "Features of Gut Microbiome Associated With Responses to Fecal Microbiota Transplantation for Inflammatory Bowel Disease: A Systematic Review," *Frontiers in Medicine*, vol. 9, 2022, Accessed: Sep. 19, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmed.2022.773105

[207] J. D. Ravichandar *et al.*, "Strain level and comprehensive microbiome analysis in inflammatory bowel disease via multi-technology meta-analysis identifies key bacterial influencers of disease," *Frontiers in Microbiology*, vol. 13, 2022, Accessed: Sep. 20, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2022.961020

[208] M. Kulecka *et al.*, "Diarrheal-associated gut dysbiosis in cancer and inflammatory bowel disease patients is exacerbated by Clostridioides difficile infection," *Frontiers in Cellular and Infection Microbiology*, vol. 13, 2023, Accessed: Sep. 20, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcimb.2023.1190910

APPENDIX

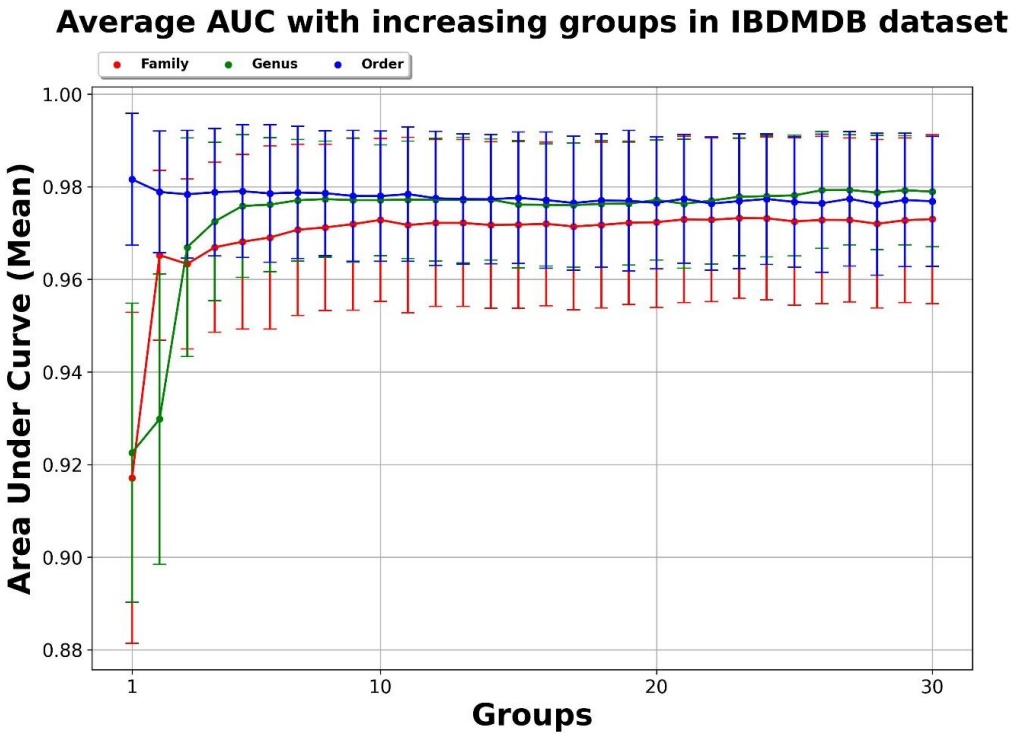


Figure A1. Changes in average AUC scores as the number of groups increases from 1 to 30, tested on the IBDMDB dataset across three different taxonomic levels.

Table A1. The miRNAs of the top 10 disease groups important for LUSC, which were determined by comparing and validating the miRNAs with external databases.

Group Name (disease name)	Score	# of miRNAs	# of validated miRNAs in the Groups for LUSC	Validated miRNAs in the Groups for LUSC
acute cerebral ischemia	0.003164	1	0	---
aortic stenosis	0.006329	39	5	hsa-miR-30a, hsa-miR-133a, hsa-miR-193a, hsa-miR-21, hsa-miR-195
bladder neoplasms	0.00949	134	14	hsa-miR-144, hsa-miR-126, hsa-miR-133a, hsa-miR-21, hsa-miR-140, hsa-miR-195, hsa-miR-15a, hsa-miR-30a, hsa-miR-193a, hsa-miR-125a, hsa-miR-101, hsa-miR-218, hsa-miR-223, hsa-miR-185
carcinoma, lung, non-small-cell	0.01265	282	20	hsa-miR-144, hsa-miR-126, hsa-miR-133a, hsa-miR-195, hsa-miR-30d, hsa-let-7i, hsa-miR-101, hsa-miR-375, hsa-miR-223, hsa-miR-185, hsa-miR-7, hsa-miR-21, hsa-miR-140, hsa-miR-15a, hsa-miR-30a, hsa-miR-193a, hsa-miR-125a, hsa-miR-218, hsa-miR-372, hsa-miR-95
idiopathic pulmonary fibrosis	0.01582	13	3	hsa-miR-30a, hsa-miR-21, hsa-miR-185
melanoma	0.01582	256	17	hsa-miR-144, hsa-miR-126, hsa-miR-650, hsa-miR-195, hsa-miR-30d, hsa-let-7i, hsa-miR-101, hsa-miR-375, hsa-miR-223, hsa-miR-185, hsa-miR-7, hsa-miR-21, hsa-miR-15a, hsa-miR-30a, hsa-miR-193a, hsa-miR-125a, hsa-miR-218
neoplasms [unspecific]	0.02215	327	19	hsa-miR-144, hsa-miR-126, hsa-miR-650, hsa-miR-195, hsa-miR-30d, hsa-let-7i, hsa-miR-101, hsa-miR-375, hsa-miR-223, hsa-miR-185, hsa-miR-7, hsa-miR-21, hsa-miR-140, hsa-miR-15a, hsa-miR-30a, hsa-miR-193a, hsa-miR-125a, hsa-miR-218, hsa-miR-372
colorectal carcinoma	0.02531	349	21	hsa-miR-144, hsa-miR-126, hsa-miR-650, hsa-miR-133a, hsa-miR-195, hsa-miR-30d, hsa-let-7i, hsa-miR-101, hsa-miR-375, hsa-miR-223, hsa-miR-185, hsa-miR-7, hsa-miR-21, hsa-miR-140, hsa-miR-15a, hsa-miR-30a, hsa-miR-193a, hsa-miR-125a, hsa-miR-218, hsa-miR-372, hsa-miR-95
eosinophilic esophagitis	0.02848	29	6	hsa-miR-223, hsa-miR-30a, hsa-miR-193a, hsa-miR-21, hsa-miR-144, hsa-miR-375
heart failure	0.03164	195	12	hsa-miR-126, hsa-miR-650, hsa-miR-133a, hsa-miR-21, hsa-miR-195, hsa-miR-30d, hsa-miR-30a, hsa-miR-125a, hsa-let-7i, hsa-miR-375, hsa-miR-223, hsa-miR-372

Table A2. Performance metrics obtained for all dataset using microBiomeGSM. The effect of grouping at different taxonomic levels (i.e., Order, Family and Genus) is shown. G represents Group.

CRC											
Taxa Rank	**Metric**	**10 G**	**9 G**	**8 G**	**7 G**	**6 G**	**5 G**	**4 G**	**3 G**	**2 G**	**1 G**
Order	**Accuracy**	0,74	0,74	0,74	0,75	0,75	0,75	0,74	0,74	0,72	0,69
Family		0,74	0,74	0,74	0,74	0,75	0,75	0,73	0,71	0,70	0,68
Genus		0,73	0,74	0,74	0,73	0,73	0,73	0,72	0,70	0,69	0,66
Order	**Sensitivity**	0,73	0,73	0,72	0,73	0,72	0,72	0,72	0,70	0,69	0,63
Family		0,69	0,69	0,69	0,69	0,69	0,64	0,56	0,50	0,46	0,42
Genus		0,67	0,67	0,65	0,65	0,60	0,58	0,53	0,50	0,47	0,40
Order	**Specificity**	0,76	0,76	0,77	0,76	0,77	0,77	0,77	0,77	0,76	0,76
Family		0,79	0,79	0,80	0,80	0,81	0,87	0,90	0,91	0,93	0,94
Genus		0,79	0,80	0,83	0,82	0,85	0,88	0,90	0,90	0,91	0,92
Order	**AUC**	0,81	0,81	0,81	0,82	0,82	0,82	0,82	0,82	0,80	0,77
Family		0,83	0,82	0,82	0,82	0,83	0,79	0,76	0,74	0,71	0,69
Genus		0,78	0,79	0,79	0,79	0,78	0,77	0,76	0,75	0,73	0,68
IBDMDB											
Taxa Rank	**Metric**	**10 G**	**9 G**	**8 G**	**7 G**	**6 G**	**5 G**	**4 G**	**3 G**	**2 G**	**1 G**
Order	**Accuracy**	0,93	0,94	0,94	0,94	0,94	0,95	0,94	0,94	0,94	0,96
Family		0,95	0,94	0,94	0,94	0,94	0,94	0,94	0,93	0,93	0,93
Genus		0,92	0,93	0,93	0,93	0,93	0,93	0,92	0,92	0,93	0,92
Order	**Sensitivity**	0,97	0,98	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98
Family		0,98	0,98	0,98	0,98	0,99	0,98	0,98	0,97	0,97	0,99
Genus		0,98	0,98	0,97	0,98	0,98	0,98	0,97	0,97	0,98	0,98
Order	**Specificity**	0,86	0,86	0,87	0,86	0,87	0,88	0,87	0,87	0,87	0,93
Family		0,87	0,86	0,86	0,86	0,86	0,86	0,86	0,84	0,83	0,81
Genus		0,82	0,82	0,84	0,82	0,83	0,84	0,83	0,82	0,82	0,8
Order	**AUC**	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98
Family		0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,97	0,97	0,93
Genus		0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,96	0,92	0,91
T2D											
Taxa Rank	**Metric**	**10 G**	**9 G**	**8 G**	**7 G**	**6 G**	**5 G**	**4 G**	**3 G**	**2 G**	**1 G**
Order	**Accuracy**	0,66	0,66	0,68	0,67	0,67	0,68	0,66	0,67	0,67	0,64
Family		0,65	0,66	0,65	0,67	0,65	0,66	0,67	0,63	0,63	0,60
Genus		0,62	0,62	0,62	0,63	0,62	0,59	0,59	0,58	0,58	0,55
Order	**Sensitivity**	0,65	0,64	0,68	0,64	0,62	0,66	0,64	0,64	0,67	0,66
Family		0,65	0,66	0,66	0,65	0,64	0,64	0,69	0,66	0,65	0,66
Genus		0,69	0,67	0,66	0,67	0,66	0,63	0,65	0,65	0,64	0,65
Order	**Specificity**	0,67	0,69	0,67	0,70	0,71	0,69	0,67	0,69	0,67	0,61
Family		0,64	0,65	0,64	0,69	0,66	0,69	0,65	0,60	0,61	0,54
Genus		0,56	0,56	0,59	0,59	0,58	0,54	0,52	0,50	0,51	0,46
Order	**AUC**	0,73	0,74	0,73	0,74	0,73	0,74	0,73	0,74	0,70	0,66
Family		0,70	0,70	0,69	0,68	0,67	0,70	0,71	0,71	0,68	0,66
Genus		0,68	0,67	0,67	0,69	0,67	0,65	0,62	0,59	0,58	0,59
IBD											
Taxa Rank	**Metric**	**10 G**	**9 G**	**8 G**	**7 G**	**6 G**	**5 G**	**4 G**	**3 G**	**2 G**	**1 G**
Order	**Accuracy**	0,82	0,82	0,83	0,82	0,83	0,85	0,85	0,83	0,83	0,86
Family		0,81	0,83	0,82	0,81	0,81	0,82	0,80	0,81	0,82	0,79
Genus		0,82	0,82	0,80	0,79	0,78	0,79	0,76	0,75	0,73	0,69
Order	**Sensitivity**	0,84	0,83	0,83	0,83	0,83	0,85	0,85	0,84	0,85	0,87
Family		0,83	0,84	0,83	0,84	0,84	0,83	0,81	0,83	0,82	0,77
Genus		0,85	0,84	0,79	0,79	0,80	0,78	0,75	0,75	0,71	0,63
Order	**Specificity**	0,81	0,82	0,83	0,81	0,84	0,85	0,85	0,81	0,81	0,84
Family		0,79	0,81	0,81	0,79	0,78	0,80	0,79	0,79	0,82	0,81
Genus		0,79	0,79	0,80	0,78	0,75	0,81	0,76	0,76	0,75	0,75
Order	**AUC**	0,92	0,93	0,93	0,92	0,93	0,93	0,93	0,92	0,93	0,91
Family		0,88	0,88	0,87	0,88	0,87	0,89	0,89	0,88	0,88	0,86
Genus		0,89	0,88	0,88	0,86	0,85	0,85	0,83	0,82	0,80	0,74

Table A3. Top 5 significant groups that are identified by microBiomeGSM for family taxonomic level for all datasets.

CRC	
Top 5 Family groups	**List of species**
PEPTOSTREPTOCOCCACEAE	Clostridioides_difficile, Criibacterium_bergeronii, Filifactor_alocis, Intestinibacter_bartlettii ……
PEPTONIPHILACEAE	Anaerococcus_lactolyticus, Anaerococcus_tetradius, Anaerococcus_vaginalis, Finegoldia_magna ……..
FUSOBACTERIACEAE	Cetobacterium_somerae, Fusobacterium_equinum, Fusobacterium_gonidiaformans, Fusobacterium_hwasookii, Fusobacterium_mortiferum …….
BACILLALES_UNCLASSIFIED	Gemella_asaccharolytica, Gemella_bergeri, Gemella_morbillorum, Gemella_sanguinis
VEILLONELLACEAE	Allisonella_histaminiformans, Anaeroglobus_geminatus, Dialister_invisus, Dialister_micraerophilus ……
IBDMDB	
Top 5 Family groups	**List of species**
BACTEROIDACEAE	Bacteroides_caccae,Bacteroides_cellulosilyticus, Bacteroides_clarus, Bacteroides_coprocola
LACHNOSPIRACEAE	Anaerocolumna_aminovalerica, Anaerosporobacter_mobilis, Anaerostipes_caccae, Anaerostipes_hadrus
RUMINOCOCCACEAE	Agathobaculum_butyriciproducens, Anaerofilum_sp_An201, Anaeromassilibacillus_sp_An172, Anaeromassilibacillus_sp_An250
RIKENELLACEAE	Alistipes_finegoldii, Alistipes_indistinctus, Alistipes_inops, Alistipes_onderdonkii
FIRMICUTES_UNCLASSIFIED	Firmicutes_bacterium_CAG_110, Firmicutes_bacterium_CAG_145, Firmicutes_bacterium_CAG_170, Firmicutes_bacterium_CAG_238
IBD	
Top 5 Family groups	**List of species**
LACHNOSPIRACEAE	Anaerostipes_hadrus, Blautia_hydrogenotrophica, Ruminococcus_gnavus, Ruminococcus_obeum……..
BIFIDOBACTERIACEAE	Bifidobacterium_adolescentis, Bifidobacterium_angulatum, Bifidobacterium_bifidum, Bifidobacterium_catenulatum……
CORIOBACTERIACEAE	Adlercreutzia_equolifaciens, Atopobium_parvulum, Atopobium_rimae, Collinsella_aerofaciens ….
RUMINOCOCCACEAE	Anaerotruncus_colihominis, Anaerotruncus_unclassified, Faecalibacterium_prausnitzii, Ruminococcaceae_bacterium_D16…..
ERYSIPELOTRICHACEAE	Catenibacterium_mitsuokai, Coprobacillus_sp_29_1, Coprobacillus_sp_D6, Clostridium_innocuum….
T2D	
Top 5 Family groups	**List of species**
LACHNOSPIRACEAE	Anaerostipes_hadrus, Blautia_hydrogenotrophica, Ruminococcus_gnavus, Ruminococcus_obeum ……
BIFIDOBACTERIACEAE	Bifidobacterium_adolescentis, Bifidobacterium_angulatum, Bifidobacterium_bifidum, Bifidobacterium_catenulatum ……..
RUMINOCOCCACEAE	Anaerotruncus_colihominis, Anaerotruncus_unclassified, Faecalibacterium_prausnitzii, Ruminococcaceae_bacterium_D16, Ruminococcus_albus, …….
EUBACTERIACEAE	Eubacterium_brachy, Eubacterium_eligens, Eubacterium_hallii, Eubacterium_ramulus, Eubacterium_rectale ……
CORIOBACTERIACEAE	Adlercreutzia_equolifaciens, Atopobium_parvulu, Atopobium_rimae, Collinsella_aerofaciens ……

Table A4. Top 5 significant groups that are identified by microBiomeGSM for order taxonomic level for all datasets.

CRC	
Top 5 Order groups	**List of species**
CLOSTRIDIALES	Catabacter_hongkongensis, Christensenella_minuta, Butyricicoccus_pullicaecorum, Butyribacterium_methylotrophicum, Clostridium_baratii, …….
TISSIERELLALES	Anaerococcus_lactolyticus,Anaerococcus_tetradius, Anaerococcus_vaginalis, Finegoldia_magna, Parvimonas_micra…
BACTEROIDALES	Bacteroides_caccae, Bacteroides_caecimuris, Bacteroides_cellulosilyticus, Bacteroides_clarus, Bacteroides_coprocola ……
FUSOBACTERIALES	Cetobacterium_somerae, Fusobacterium_equinum, Fusobacterium_gonidiaformans, Fusobacterium_hwasookii, Fusobacterium_mortiferum, ……
BACILLALES	Bacillus_aerius, Bacillus_sp_FJAT_27916, Gemella_asaccharolytica, Gemella_bergeri, Gemella_haemolysans, ……
IBDMDB	
Top 5 Order groups	**List of species**
BACTEROIDALES	Bacteroides_caccae, Bacteroides_cellulosilyticus, Bacteroides_clarus, Bacteroides_coprocola
CLOSTRIDIALES	Catabacter_hongkongensis, Christensenella_minuta, Butyricicoccus_pullicaecorum, Butyribacterium_methylotrophicum
FIRMICUTES_UNCLASSIFIED	Firmicutes_bacterium_CAG_110, Firmicutes_bacterium_CAG_145, Firmicutes_bacterium_CAG_170, Firmicutes_bacterium_CAG_238
VEILLONELLALES	Allisonella_histaminiformans, Anaeroglobus_geminatus, Dialister_invisus, Dialister_micraerophilus
BURKHOLDERIALES	Oxalobacter_formigenes, Parasutterella_excrementihominis, Sutterella_parvirubra, Turicimonas_muris
IBD	
Top 5 Order groups	**List of species**
CLOSTRIDIALES	Clostridium_asparagiforme, Clostridium_bolteae, Clostridium_citroniae, Clostridium_clostridioforme, Clostridium_hathewayi, …….
CORIOBACTERIALES	Adlercreutzia_equolifaciens, Atopobium_parvulum, Atopobium_rimae, Collinsella_aerofaciens, Collinsella_unclassified ……
BIFIDOBACTERIALES	Bifidobacterium_adolescentis, Bifidobacterium_angulatum, Bifidobacterium_bifidum, Bifidobacterium_catenulatum, Bifidobacterium_longum……
ERYSIPELOTRICHALES	Coprobacillus_sp_D6, Catenibacterium_mitsuokai,Clostridium_innocuum, Erysipelotrichaceae_bacterium_2_2_44A, Coprobacillus_sp_29_1 ……
BACTEROIDALES	Bacteroides_barnesiae, Bacteroides_caccae, Bacteroides_cellulosilyticus, Bacteroides_clarus …
T2D	
Top 5 Order groups	**List of species**
CLOSTRIDIALES	Clostridium_asparagiforme, Clostridium_bolteae, Clostridium_citroniae, Clostridium_clostridioforme, Clostridium_hathewayi, …….
BIFIDOBACTERIALES	Bifidobacterium_adolescentis, Bifidobacterium_angulatum, Bifidobacterium_bifidum, Bifidobacterium_catenulatum,
CORIOBACTERIALES	Adlercreutzia_equolifaciens, Atopobium_parvulum, Atopobium_rimae, Collinsella_aerofaciens, Collinsella_unclassified ……
BACTEROIDALES	Bacteroides_barnesiae, Bacteroides_caccae, Bacteroides_cellulosilyticus, Bacteroides_clarus, Bacteroides_coprocola, ……
LACTOBACILLALES	Granulicatella_unclassified, Enterococcus_faecium, Lactobacillus_animalis, Lactobacillus_delbrueckii, Lactobacillus_fermentum, ……

Table A5. Top 5 significant groups that are identified by microBiomeGSM for genus taxonomic level for all datasets.

CRC	
Top 5 Genus groups	**List of species**
PARVIMONAS	Parvimonas_micra, Parvimonas_sp_KA00067, Parvimonas_sp_oral_taxon_110, Parvimonas_sp_oral_taxon_393 ……
PEPTOSTREPTOCOCCUS	Peptostreptococcus_anaerobius, Peptostreptococcus_sp_MV1, Peptostreptococcus_stomatis
FUSOBACTERIUM	Fusobacterium_equinum, Fusobacterium_gonidiaformans, Fusobacterium_hwasookii, Fusobacterium_mortiferum, Fusobacterium_naviforme ……
GEMELLA	Gemella_asaccharolytica, Gemella_bergeri, Gemella_haemolysans, Gemella_morbillorum……
DIALISTER	Dialister_invisus, Dialister_micraerophilus, Dialister_pneumosintes, Dialister_sp_CAG_357, Dialister_succinatiphilus
IBDMDB	
Top 5 Genus groups	**List of species**
BACTEROIDES	Bacteroides_caccae, Bacteroides_cellulosilyticus, Bacteroides_clarus, Bacteroides_coprocola
ALISTIPES	Alistipes_finegoldii, Alistipes_indistinctus, Alistipes_inops, Alistipes_onderdonkii
EUBACTERIUM	Eubacterium_coprostanoligenes, Eubacterium_dolichum_CAG_375, Eubacterium_eligens, Eubacterium_hallii
ROSEBURIA	Roseburia_faecis, Roseburia_hominis, Roseburia_intestinalis, Roseburia_inulinivorans
FIRMICUTES_UNCLASSIFIED	Firmicutes_bacterium_CAG_110, Firmicutes_bacterium_CAG_145, Firmicutes_bacterium_CAG_170, Firmicutes_bacterium_CAG_238
IBD	
Top 5 Genus groups	**List of species**
BLAUTIA	Blautia_hydrogenotrophica, Ruminococcus_gnavus, Ruminococcus_obeum, Ruminococcus_torques, Blautia_hansenii ……
BIFIDOBACTERIUM	Bifidobacterium_adolescentis, Bifidobacterium_angulatum, Bifidobacterium_bifidum, Bifidobacterium_catenulatum …….
EUBACTERIUM	Eubacterium_brachy, Eubacterium_eligens, Eubacterium_hallii, Eubacterium_ramulus, Eubacterium_rectale ……
DOREA	Dorea_formicigenerans, Dorea_longicatena, Dorea_unclassified…
COLLINSELLA	Collinsella_aerofaciens, Collinsella_unclassified, Collinsella_intestinalis, Collinsella_stercoris, Collinsella_tanakaei ……
T2D	
Top 5 Genus groups	**List of species**
EUBACTERIUM	Eubacterium_brachy, Eubacterium_eligens, Eubacterium_hallii, Eubacterium_ramulus, Eubacterium_rectale ……
BIFIDOBACTERIUM	Bifidobacterium_adolescentis, Bifidobacterium_angulatum, Bifidobacterium_bifidum, Bifidobacterium_catenulatum …….
BLAUTIA	Ruminococcus_torques, Blautia_hydrogenotrophica, Ruminococcus_obeum, Ruminococcus_gnavus, Blautia_hansenii ……
DOREA	Dorea_formicigenerans, Dorea_longicatena, Dorea_unclassified
LACHNOSPIRACEAE_NONAME	lachnospiraceae_bacterium_1_1_57FAA, Lachnospiraceae_bacterium_1_4_56FAA, Lachnospiraceae_bacterium_2_1_58FAA, Lachnospiraceae_bacterium_3_1_46FAA, ……

Table A6. List of biomarkers derived by microBiomeGSM for all dataset for T2D and IBDMDB.

T2D		
Rank	**Name of Microbiome (Family taxonomic level)**	**Reference**
1	LACHNOSPIRACEAE	[145]
2	BIFIDOBACTERIACEAE	[146]
3	RUMINOCOCCACEAE	[147]
4	EUBACTERIACEAE	[145]
5	CORIOBACTERIACEAE	[145]
6	CLOSTRIDIALES FAMILY XIII INCERTAE SEDIS	–
7	ERYSIPELOTRICHACEAE	[145]
8	PEPTOSTREPTOCOCCACEAE	[145]
9	CARNOBACTERIACEAE	–
10	BACTEROIDACEAE	[147]
Rank	**Name of Microbiome (Order taxonomic level)**	**Reference**
1	CLOSTRIDIALES	[148]
2	BIFIDOBACTERIALES	–
3	CORIOBACTERIALES	[149]
4	BACTEROIDALES	[150]
5	LACTOBACILLALES	–
6	ERYSIPELOTRICHALES	[151]
7	SELENOMONADALES	[152]
8	VERRUCOMICROBIALES	[153]
9	METHANOBACTERIALES	–
10	BACILLALES	–
Rank	**Name of Microbiome (Genus taxonomic level)**	**Reference**
1	EUBACTERIUM	[154]
2	BIFIDOBACTERIUM	[155]
3	BLAUTIA	[156]
4	DOREA	[157]
5	LACHNOSPIRACEAE_NONAME	–
6	RUMINOCOCCUS	[158]
7	COPROCOCCUS	[159]
8	PEPTOSTREPTOCOCCUS	[160]
9	ERYSIPELOTRICHACEAE_NONAME	–
10	GRANULICATELLA	[161]
IBDMDB		
Rank	**Name of Microbiome (Family taxonomic level)**	**Reference**
1	BACTEROIDACEAE	[162]
2	LACHNOSPIRACEAE	[163]
3	RUMINOCOCCACEAE	[164]
4	RIKENELLACEAE	[165]
5	FIRMICUTES_UNCLASSIFIED	[166]
6	TANNERELLACEAE	[165]
7	EUBACTERIACEAE	[65]
8	CLOSTRIDIACEAE	[167]
9	VEILLONELLACEAE	[96]
10	ODORIBACTERACEAE	[168]
Rank	**Name of Microbiome (Order taxonomic level)**	**Reference**
1	BACTEROIDALES	[169]
2	CLOSTRIDIALES	[170]
3	FIRMICUTES_UNCLASSIFIED	[166]
4	VEILLONELLALES	[171]
5	BURKHOLDERIALES	[172]
6	METHANOMASSILIICOCCALES	–
7	DESULFOVIBRIONALES	–
8	ERYSIPELOTRICHALES	–
9	BIFIDOBACTERIALES	[173]
10	EGGERTHELLALES	–
Rank	**Name of Microbiome (Genus taxonomic level)**	**Reference**
1	BACTEROIDES	[174]
2	ALISTIPES	[175]
3	EUBACTERIUM	[176]
4	ROSEBURIA	[177]
5	FIRMICUTES_UNCLASSIFIED	[166]
6	PARABACTEROIDES	[178]
7	RUMINOCOCCUS	[179]
8	COPROCOCCUS	[180]
9	BLAUTIA	[181]
10	CLOSTRIDIUM	[182]

Table A7. List of biomarkers derived by microBiomeGSM for all dataset for CRC and IBD.

CRC		
Rank	**Name of Microbiome (Family taxonomic level)**	**Reference**
1	PEPTOSTREPTOCOCCACEAE	--
2	PEPTONIPHILACEAE	--
3	FUSOBACTERIACEAE	--
4	BACILLALES_UNCLASSIFIED	[183]
5	VEILLONELLACEAE	[184]
6	LACHNOSPIRACEAE	[185]
7	ERYSIPELOTRICHACEAE	[186]
8	RUMINOCOCCACEAE	[185]
9	PREVOTELLACEAE	[187]
10	STREPTOCOCCACEAE	[188]
Rank	**Name of Microbiome (Order taxonomic level)**	**Reference**
1	CLOSTRIDIALES	[189]
2	TISSIERELLALES	--
3	BACTEROIDALES	[190]
4	FUSOBACTERIALES	--
5	BACILLALES	[183]
6	VEILLONELLALES	--
7	ERYSIPELOTRICHALES	--
8	LACTOBACILLALES	[191]
9	ACTINOMYCETALES	--
10	DESULFOVIBRIONALES	--
Rank	**Name of Microbiome (Genus taxonomic level)**	**Reference**
1	PARVIMONAS	[185]
2	PEPTOSTREPTOCOCCUS	[192]
3	FUSOBACTERIUM	[193]
4	GEMELLA	[194]
5	DIALISTER	[195]
6	LACHNOCLOSTRIDIUM	[185]
7	PREVOTELLA	[196]
8	STREPTOCOCCUS	[197]
9	PORPHYROMONAS	[185]
10	SOLOBACTERIUM	[198]
IBD		
Rank	**Name of Microbiome (Family taxonomic level)**	**Reference**
1	LACHNOSPIRACEAE	[199]
2	BIFIDOBACTERIACEAE	[200]
3	CORIOBACTERIACEAE	--
4	RUMINOCOCCACEAE	[201]
5	ERYSIPELOTRICHACEAE	[202]
6	CLOSTRIDIALES_FAMILY_XIII_INCERTAE_SEDIS	--
7	EUBACTERIACEAE	[65]
8	PEPTOSTREPTOCOCCACEAE	--
9	CARNOBACTERIACEAE	--
10	CLOSTRIDIACEAE	--
Rank	**Name of Microbiome (Order taxonomic level)**	**Reference**
1	CLOSTRIDIALES	[203]
2	CORIOBACTERIALES	--
3	BIFIDOBACTERIALES	--
4	ERYSIPELOTRICHALES	--
5	BACTEROIDALES	[204]
6	LACTOBACILLALES	--
7	SELENOMONADALES	--
8	VERRUCOMICROBIALES	--
9	CANDIDATUS_SACCHARIBACTERIA_NONAME	--
10	BACILLALES	--
Rank	**Name of Microbiome (Genus taxonomic level)**	**Reference**
1	BLAUTIA	[199]
2	BIFIDOBACTERIUM	[205]
3	EUBACTERIUM	[206]
4	DOREA	[199]
5	COLLINSELLA	--
6	PEPTOSTREPTOCOCCUS	--
7	COPROCOCCUS	[180]
8	ERYSIPELOTRICHACEAE_NONAME	--
9	LACHNOSPIRACEAE_NONAME	[207]
10	BACTEROIDES	[208]

CURRICULUM VITAE

2010 – 2015	Baclehor, Computer Engineering, Erciyes University, Kayseri, TURKEY
2015 – 2018	Master, Management Information Systems, Sivas Cumhuriyet University, Sivas, TURKEY
2019 – 2024	Doctoral Candidate, Electrical and Computer Engineering, Abdullah Gul University, Kayseri, TÜRKİYE
2017-	Research Assistant, Management Information Systems, Sivas Cumhuriyet University, Sivas, TURKEY

SELECTED PUBLICATIONS AND PRESENTATIONS

J1) Jabeer, A., **Temiz, M.**, Bakir-Gungor, B., & Yousef, M. "miRdisNET: discovering microRNA biomarkers that are associated with diseases utilizing biological knowledge-based machine learning", Frontiers in Genetics (2023).

J2) Bakir-Gungor, B., **Temiz, M.**, Jabeer, A., Wu, D., & Yousef, M. "microBiomeGSM: the identification of taxonomic biomarkers from metagenomic data using grouping, scoring and modeling (GSM) approach", Frontiers in Microbiology (2023).

C1) Temiz, M., Yousef, M., & Bakir-Gungor, B. "Population Specific Classification of Colorectal Cancer with Meta-Analysis of Metagenomic Data". In 2023 Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE, pp. 1-5, (2023)