Beyza CANAKCIMAKSUTOGLU

M.Sc. Thesis

AGU 2025

IDENTIFYING POTENTIAL TAXONOMIC BIOMARKERS OF GASTROINTESTINAL CANCERS FROM HUMAN MICROBIOTA USING THE GROUPING-SCORING-MODELING (G-S-M) AND TRADITIONAL FEATURE SELECTION APPROACHES

M.Sc. THESIS
SUBMITTED TO THE DEPARTMENT OF BIOENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Beyza ÇANAKCIMAKSUTOĞLU
May 2025

IDENTIFYING POTENTIAL TAXONOMIC BIOMARKERS OF GASTROINTESTINAL CANCERS FROM HUMAN MICROBIOTA USING THE GROUPING-SCORING-MODELING (G-S-M) AND TRADITIONAL FEATURE SELECTION APPROACHES

A THESIS
SUBMITTED TO THE DEPARTMENT OF BIOENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF
ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Beyza ÇANAKCIMAKSUTOĞLU
May 2025

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: Beyza ÇANAKCIMAKSUTOĞLU
Signature :

REGULATORY COMPLIANCE

M.Sc. thesis titled Identifying Potential Taxonomic Biomarkers of Gastrointestinal Cancers from Human Microbiota Using the Grouping-Scoring-Modeling (G-S-M) and Traditional Feature Selection Approaches 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
Beyza Çanakcımaksutoğlu	Assoc. Prof. Burcu Güngör	Prof. Dr. Malik Yousef

Head of the Bioengineering Program
Asst. Prof. Emel Başak Gencer Akçok

ACCEPTANCE AND APPROVAL

M.Sc. thesis titled Identifying Potential Taxonomic Biomarkers for Gastrointestinal Cancers from Human Microbiota Using the Grouping-Scoring-Modeling (G-S-M) and Traditional Feature Selection Approaches and prepared by Beyza Canakcimaksutoglu has been accepted by the jury in the Bioengineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

22 /05/ 2025

(Thesis Defense Exam Date)

JURY:

Advisor : Assoc. Prof. Burcu Güngör

Member : Assoc. Prof. Müşerref Duygu Saçar Demirci

Member : Assoc. Prof. Özkan Ufuk Nalbantoğlu

APPROVAL:

The acceptance of this M.Sc. thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Engineering & Science, Executive Board dated ….. …../….. /2025 and numbered ……..…… .

…….. /…….. / 2025

Graduate School Dean
Prof. İrfan ALAN

ABSTRACT

IDENTIFYING POTENTIAL TAXONOMIC BIOMARKERS FOR GASTROINTESTINAL CANCERS FROM HUMAN MICROBIOTA USING THE GROUPING-SCORING-MODELING (G-S-M) AND TRADITIONAL FEATURE SELECTION APPROACHES

Beyza CANAKCIMAKSUTOGLU
MSc. in Bioengineering
Advisor: Assoc. Prof. Burcu Gungor
Co-advisor : Prof. Dr. Malik Yousef
May 2025

Analysis of microbial abundance values holds potential for cancer prediction. This study aims to identify shared microbial biomarkers among gastrointestinal (GI) cancer patients using both tissue and blood samples—an area not previously studied in parallel. This study analyzed blood and tissue samples, focusing on head and neck, esophagus, stomach, colon, and colorectal cancers, processing them individually. By performing decontamination steps, processing non-human genetic codes, determining microorganisms and their abundances at the species level, the TCMA data set was created from the "Cancer Genome Atlas", which collected tissue and blood samples from cancer patients. Traditional feature selection algorithms (CMIM, mRMR, FCBF, IG, XGB, and SKB) reduced the high-dimensional feature space. Classification performance was evaluated using a forest classifier with 100-fold Monte Carlo cross-validation. Moreover, the MicrobiomeGSM model, which was created to decrease the feature size and prediction time via a grouping method, was trained, and the generalizability of the MicrobiomeGSM model was showcased. Traditional feature selection methods and the biological data-based MicrobiomeGSM model were applied, and their performance was compared. In the future, common biomarker candidates may help to understand the possibility of metastasis, and medical doctors can decide their treatment path of patients.

Keywords: Taxonomic Biomarker Selections, Machine Learning, Grouping Scoring Modeling, Metagenomics, Feature Selection

ÖZET

GRUPLAMA PUANLAMA MODELLEME (G-S-M) VE GELENEKSEL ÖZELLİK SEÇİM YAKLAŞIMINI KULLANARAK İNSAN GASTROİNTESTİNAL KANSER MİKROBİYOTALARINDAKİ POTANSİYEL TAKSONOMİK BİYOBELİRTEÇLERİN BELİRLENMESİ

Beyza ÇANAKCIMAKSUTOĞLU
Biyomühendislik Anabilim Dalı Yüksek Lisans
Tez Danışmanı: Doç. Dr. Burcu Güngör
İkinci Tez Danışmanı: Prof. Dr. Malik Yousef
Mayıs 2025

Mikrobiyal bolluk değerlerinin analizi, kanser tahmini için bir potansiyel taşır. Bu çalışma, daha önce paralel olarak incelenmemiş bir alan olan hem doku hem de kan örnekleri kullanarak gastrointestinal (GI) kanser hastaları arasında paylaşılan mikrobiyal biyobelirteçleri belirlemeyi amaçlamaktadır. Bu çalışma, baş ve boyun, yemek borusu, mide, kolon ve kolorektal kanserlere odaklanarak kan ve doku örneklerini analiz etti. Dekontaminasyon adımları gerçekleştirilerek, insan olmayan genetik kodlar işlenerek, tür düzeyinde mikroorganizmalar ve bollukları belirlenerek, kanser hastalarından doku ve kan örnekleri toplayan "Kanser Genom Atlası"ndan TCMA veri seti oluşturuldu. Geleneksel özellik seçimi algoritmaları (CMIM, mRMR, FCBF, IG, XGB ve SKB) yüksek boyutlu özellik alanını daralttı. Sınıflandırma performansı, 100-kat Monte Carlo çapraz doğrulaması olan bir Random Forest kullanılarak değerlendirildi. Ayrıca, gruplama yöntemi ile özellik boyutunu ve tahmin süresini azaltmak için oluşturulan MicrobiomeGSM modeli, hem kan hem de dokudan türetilen örnekler kullanılarak eğitildi ve MicrobiomeGSM modelinin genelleştirilebilirliği sergilendi. Geleneksel özellik seçimi yöntemleri ve biyolojik veri tabanlı MicrobiomeGSM modellerinin performansları karşılaştırıldı. Gelecekte, ortak biyobelirteç adayları doktorların metastaz olasılığını anlamasına yardımcı olabilir ve tedavi yollarına buna göre karar verilebilir.

Anahtar Kelimeler: Taksonomik Biyobelirteç Seçimleri, Makine Öğrenimi, Gruplama Puanlama Modellemesi, Metagenomik, Özellik Seçimi

Acknowledgements

To my beloved parents, your unwavering support, endless love, and sacrifices have been the foundation of everything I've accomplished. You never stopped believing in me, even when I doubted myself. You gave me strength when I felt weak, and reminded me of my worth when things got tough. Every step I've taken in this journey has been possible because of the values, discipline, and encouragement you instilled in me. Thank you for being my pillars, my safe place, and my greatest motivation.

To my siblings, who have stood by me not just as family, but as true companions — your presence, patience, and support made even the hardest days bearable. Thank you for always being there, sometimes without saying a word, but always making me feel understood and never alone.

I would also like to express my heartfelt appreciation to my advisor, Assoc. Prof. Burcu Bakır-Güngör, and Prof. Dr. Malik Yousef. I'm grateful for their time, insight, and unwavering belief in the project.

I thank TÜBİTAK BİDEB for supporting my research under the 2210-C Domestic Graduate Scholarship Program (Project No: 2210), and for providing me with financial assistance during my master's studies

To all my friends — especially Didem Bayraktaroğlu, Zehra Keskin and Meryem Öztürk— thank you for your patience, for listening, for reminding me of life outside of research, and for standing by me in ways that truly mattered. Your support has been a quiet force behind this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CMIM	Conditional Mutual Information Maximization
COAD	Colon adenoma
CRC	Colorectal cancer
ESCA/EC	Esophageal Cancer
FCBF	Fast Correlation Based Filter
GI	Gastrointestinal Cancers
IG	Information Gain
HNSC	Head and Neck Cancer
MRMR	Maximum Likelihood and Minimum Redundancy
READ	Rectum Cancer
RF	Random Forest
SEN/SN	Sensitivity
SKB	Select K Best
SPE/SP	Specificity
STAD/GC	Stomach Cancer
TCGA	The Cancer Genome Atlas
TCMA	The Cancer Microbiome Atlas
XGB	Extreme Gradient Boosting

To all child
who wants to
be scientist

Chapter 1

Introduction

1.1 Gastrointestinal Cancers

Gastrointestinal (GI) cancers represent more than a quarter of all cancers, including stomach, liver, esophageal, pancreatic, and colorectal cancers, and the prevalence of such cancers is constantly increasing. Gastrointestinal cancers have a variety of etiologies and clinical management. For example, microbiota is one of these diversities. The human microbiome, consisting of trillions of microorganisms such as bacteria, viruses, eukaryotic fungi, and protozoa, significantly affects human health and metabolism [1,2,3]. Imbalanced microbiomes cause diverse diseases by disrupting the health of the organs [4]. During the progression from the adenoma stage to carcinoma, disruptions in the microbiome significantly influence this transition, accompanied by shifts in the abundance of specific microorganisms [3]. Certain pathogenic bacteria contribute to inflammation by producing genotoxic compounds, such as colibactin, which negatively impact host health. Additionally, they generate microbiome-specific metabolites, including secondary bile acids and short-chain fatty acids, which are implicated in the initiation and progression of cancer [3, 4, 5]. In addition to cancer, metabolic disorders, infectious diseases, and digestive tract diseases have a relationship with the human microbiome, and there is strong evidence for this link [6]. The same article showed that the microorganisms were mainly located in cancer tissue and cells of the immune system. So, these microorganisms are not located in the extracellular compartment [6]. In the same article, it was stated that some bacteria were found close to the nuclear membrane in breast cancer tissue. It has been stated that the *Fusobacterium nucleatum*, on which many studies have been conducted, has a close interaction with colorectal cancer cells. It is important to find out which microorganisms play a role in disease development. It is crucial to use taxonomic profiles obtained from metagenomic data, which is taken from the microbiota

of patients and contains many living genetic residues in disease prediction, and many articles are working on this idea [7].

According to the study conducted by Kayali and his teammates, it is estimated that gastrointestinal cancers will cause 7.5 million new cases and 5.6 million deaths in 20 years [8]. According to GLOBOCAN databases, it is estimated that the number of people diagnosed with esophageal cancer in 2020 exceeded 604.000, while the number of deaths from these patients exceeded 544.000. The same database reported that the estimated number of stomach cancer cases exceeded 1 million, while the number of deaths associated with it was over 768.000. In the same report, the number of people suffering from colorectal cancer is estimated to be around 1.9 million, and up to 935.000 deaths [9].

1.1.1 Colorectal Cancer

Since colon cancer and rectal cancer are assumed to develop in a single organ, they are represented as a single tumor entity called colorectal cancer [10]. According to GLOBACAN [11] data, colorectal cancer is the 3rd most common type of cancer worldwide. Studies have highlighted that the risk of colorectal cancer has increased rapidly in adults under the age of 50 in the United States for 53 years, starting from the 1980s [12], [13]. New approaches are needed in the detection and treatment of colorectal cancer due to the increased risk of this cancer, especially in young adults. It has been stated in articles [13,14] that the roles of microorganisms (bacteria, fungi, viruses, archaea) living in the intestine include food fermentation, nutrient absorption, vitamin biosynthesis, formation of the intestinal epithelium and protection of the host against pathogens, and that in addition to that, the disrupted microbiota also can cause disorders like colorectal cancers [13], Study [5] has shown that the different number or variety microorganisms living in the microbiota can serve as biomarkers in the detection of CRC [13]. Studies [13, 14] have shown that Fusobacterium nucleatum and *Peptostreptococcus anaerobius* bacteria, which are highly enriched in CRC, play a role in the evolution of adenomas in the intestines into cancer and metastasis. According to the article written by Cheng et al. in 2020 [13], it was understood that *Streptococcus bovis* bacteria are another type of microorganism known to

be pathogenic for CRC, and while the number of *Fusobacterium nucleatum* increases significantly in early-stage CRC cases, it can also lead to a poor prognosis. In the same article [13], it is predicted that *Enterococcus faecalis* has high abundance values in CRC patients and may damage host DNA by mediating the production of superoxide by epithelial cells.

1.1.2 Stomach Cancer

Years ago, it was thought that stomach acid made the stomach sterile. However, it was discovered that the bacteria called *Helicobacter pylori* lived in gastritis and peptic ulcer patients, and it was proven that the stomach was not sterile [15]. According to the same research, gastric juice and gastric mucosa have different microbiota. For example, the same study [15] claims that *Firmicutes, Actinobacteria,* and *Bacteroidetes* bacterial species colonize the gastric juice, while *Proteobacteria* and *Firmicutes* bacterial species are abundant in the gastric mucosa. Studies [16, 17] mention that although *Veillonella, Lactobacillus*, and *Clostridium* bacteria are abundant in the mouth and duodenum, they can also live in the stomach, but they are temporary. According to a study conducted in 2019 [18], *Helicobacter pylori* strains were found to trigger gastric cancer because these bacteria have virulence factors. Also, the study by [15] mentioned that *Helicobacter pylori* stimulates the production of pro-inflammatory cytokines and inhibits the proliferation of immune cells, thus triggering diseases such as stomach cancer and peptic ulcers. In the study in 2021 [19], it was determined that *Peptostreptococcus stomatis, Dialister pneumosintes, Spodoptera exigua, Parvimonas micra,* and *Streptococcus anginosus* species were found more in the stomach microbiota of patients with gastric carcinoma compared to the gastric microbiota of healthy people. The same study [19] states that *Desulfovibrio, Escherichia, Faecalibacterium,* or *Oscillospira* species may be predictive of stomach cancer with accuracy values of 0.9.

1.1.3 Head and Neck Cancer

In the statement of the article [20], it is known that head and neck cancer is associated with periodontitis, which occurs with a high immune response caused by increased microorganisms in the mouth. In the study [21], *Streptococcus anginosus* living in dental plaques can increase the damage to host DNA by increasing nitric oxide production. While the same study explained that microorganisms in the upper respiratory tract may contribute to the development of cancer by affecting the lower respiratory tract, they also stated that the pathogenic bacteria detected in the mouths of lung cancer patients were actually colonized around the lung tumor. Also, *Fusobacterium nucleatum* and *Prophyromonas gingivalis* were accepted as early diagnostic biomarkers of head and neck cancer in a recent study [22]. The same research [22] observed that *L. johnsonii*|*gasseri* complex, *L. fermentum, L. rhamnosus*, and *L. vaginalis,* which are *Lactobacillus* strains, are highly enriched in the microbiota of head and neck cancer.

1.1.4 Esophageal Cancers

According to esophageal research [23], early diagnosis of Esophageal cancer is difficult as its symptoms are not obvious, and the incidence of this cancer ranks 7th among all cancers. This research mentioned that the abundance of the phyla *Proteobacteria*, *Firmicutes,* and *Bacteroidetes* exceeds that of other phyla. In the same study, it was stated that *Enterobacteriaceae, Neisseria, Streptococcus, Actinobacillus, Sphingomonas, Haemophilus, Prevotella, Veillonella,* and *Porphyromonas* species were the species that ranked first in terms of abundance in Esophageal cancer, and these abundance values were obtained as a result of comparisons with healthy people. Another study [24] explained that the *Lactobacillus* bacterial species plays an important role in esophageal cancer and that it can trigger cancer development by preventing the growth of other acid-producing bacteria it produces. One recent study [25] revealed the abundance of *Rhizobiales* and *Sphingomonadales* species in tumor tissue, while *Rhodospirillales* species were found to be abundant in tissues around the tumor.

1.1.5 Colon Cancer

Colon and rectal cancers are the second leading cause of death. Although the incidence of colon cancer is decreasing thanks to colonoscopy, the incidence of these diseases is increasing in young adults. Colon cancer occurs over a few years by following the path of accumulation of genetic mutations in cells, formation of adenomas, and cancer (carcinogenesis)[26].

1.2 Machine learning approach in GI cancers

Determining the microorganisms involved in disease development is essential, and utilizing taxonomic profiles from metagenomic data, which captures diverse genetic residues from patient microbiota, is crucial for disease prediction, as supported by various studies [7]. It is possible to see wet-lab and dry-lab studies in this area. The disease can be predicted from taxonomic profiles obtained using machine learning models, but not only that, these models can also rank potential biomarkers by predicting which microorganism is more important for that disease. These methods allow the microbiome-disease relationship to be recognized more clearly [7]. In a study conducted by the Bakır-Gungor-Yousef lab, various feature selection methods were utilized to identify taxonomic biomarkers in the intestinal microbiota associated with colon cancer. Besides feature selection methods, multiple machine learning algorithms were employed, with the Random Forest classifier outperforming Adaboost, Support Vector Machine, Decision Tree, and Logitboost algorithms [27]. In a study in 2022 [28], β-diversity between samples was calculated using principal coordinate analysis in the evaluation of biomarkers of gastric and colorectal cancers. A study [29] compiled machine learning approaches and found that Gradient Boosting models were used to determine the type of cancer using microbiome data, Random Forest (RF) [30] was generally used to detect bacteria from CRC stool samples, and deep learning and a deep neural network were used to evaluate biomarkers. An appropriate classification model was developed [31] to detect biomarkers of esophageal cancer, and logical regression, support vector machine (SVM), artificial neural network, random forest, and XGBoost classification models were used for this model.

1.3 Research Questions

This thesis aims to investigate gastrointestinal cancers, including colon, colorectal, stomach, and esophageal cancers, as well as head and neck cancers. The primary objective is to identify both common and unique biomarker candidates across these cancer types. This is important because malignant cancers have a high potential to metastasize, but such metastasis cannot be detected by Positron Emission Tomography until the cancer forms a detectable cell cluster. Since microbiota dysbiosis supports cancer development, disrupted microbiota can be a significant indicator. Pathogenic common biomarkers may act as triggers for cancer; therefore, identifying shared biomarkers among GI cancers may help reassess the likelihood of metastasis to another GI region. This can support the development of more accurate treatment strategies. If these biomarkers are present, they may allow detection of cancer before it reaches a clustered stage.

The second aim is to train the MicrobiomeGSM model with blood, normal and tumor tissues belonging to cancer patients, and at the end of this training, to predict the potantial biomarker candidates by grouping them at many taxonomic levels.

Another objective is to develop models using traditional feature selection methods on the same datasets and to evaluate their performance. Finally, the performances of models created with conventional methods will be compared to that of the MicrobiomeGSM model.

Chapter 2

Material and Methods

Primary tumors, blood-derived normal, and solid tissue normal have a total of 14,492 microbial species, meaning the dimension of the total data.

2.1. Dataset

The Cancer Genome Atlas (TCGA) database contains bacterial and viral reads from 10,481 patients, 33 cancer types, and 18,116 samples [32]. The Cancer Microbiome Atlas (TCMA) database aimed to identify microbial signatures that distinguish cancer types within the data in the TCGA database [33]. The TCGA database includes a total of 620 samples and 221 taxa, including gastrointestinal cancers [esophageal cancer (ESCA), stomach cancer (STAD), colon cancer (COAD), and rectal cancer (READ), as well as head and neck cancer (HNSC)]. These taxa include the relative abundance values of bacteria at the species level. These data consist of samples obtained from primary tumor tissue and tumor-free tissues close to the tumor tissue and belonging to the same organ. The TCMA database (https://tcma.pratt.duke.edu/) is a database that retrieves metagenomic data of cancer diseases (obtained by the next-generation sequencing method) from the TCGA database re-research [33]. Bioinformatics pipelines are used during this process. Contaminations in the data are cleaned using the PathSeq tool, and bacterial genes are detected [33]. The database in question analyzes bacterial genes from the kingdom level to the species level. The project used the relative abundance values of bacteria belonging to targeted cancer types in the database.

2.2. Data manipulation

The data in the TCMA database includes data obtained from primary tumor tissue, blood, and tumor-free tissues and relative abundance values of bacteria at the family-

order-genus-species levels. Patient samples with microorganisms at the species level were selected from the main dataset because having microorganisms at the species level can provide more detailed and reliable information. The database used separate filtering methods to examine specific cancer types. COAD, HNSC, STAD, ESCA, and COAD-READ cancer types were focused on one by one. READ (rectal cancer) and COAD (colon cancer) data constitute colorectal cancer, so these two cancers were grouped and turned into a single file for colorectal cancer. The data from the database needs to be organized in an appropriate format for use in the MicrobiomeGSM and traditional feature selection approaches. Therefore, the desired cancers were distinguished from the data of bacterial species taken from the "TCMA" database and normalized by means of sample ID numbers. Primary tumor (pos)-Solid Tissue normal (neg), blood-derived normal (neg)-primary tumor (pos) sample groups were used, and all samples belonged to cancer patients. For example, if a doctor takes a sample of primary tumor from the cancer patient, the same doctor may take a sample of normal tissue, which is far away from the tumor area of the same patient, and/or take a blood sample from the same patients. A separate file was created for each cancer and sample type. Invalid, incorrect/missing, or irrelevant data were removed from the obtained data. Very small numbers represented by Euler were replaced with 0, and as a result, a clean dataset was obtained.

At the end of data manipulation, there are 684 features in the solid tissue normal sample (as well as primary tumor sample) COAD dataset, 697 features in the HNSC dataset, 685 features in the COAD-READ dataset, 320 features in the STAD dataset, and 314 features in the ESCA dataset. There are 682 features in blood-derived samples in the COAD dataset, 698 features in the HNSC dataset, 683 features in the COAD-READ dataset, 216 features in the STAD dataset, and 312 in the ESCA dataset. Non-informative features were deleted, and then studies were conducted on the remaining features. For this purpose, different approaches were applied. The final dataset overview are represented by Table 2.2.1.

Table 2.2.1.1: Overview of 5 datasets used in this thesis

#	Cancer types	# of samples of blood	# of samples of solid	# of positives
1	COAD	99	21	125
2	COAD-READ (CRC)	140	25	170
3	ESCA	42	22	62
4	HNSC	130	21	157
5	STAD	89	39	128

2.3. Grouping Scoring Modelling Approach

One of the approaches used in this study is MicrobiomeGSM, which is a machine learning-based model. The MicrobiomeGSM tool was designed by [7]. The microbiomeGSM tool uses metagenomic data as a feature in this thesis. In the introductory article of the model and its other articles, metagenomic data of patients with diseases such as CRC and IBD were used, and biomarkers that may be important for these diseases were determined with a feature selection based on pre-existing biological data-based grouping [7, 34, 34]. The performance of this model was calculated separately for each disease data. The main purpose of the model is to increase its prediction performance by using fewer features and machine learning techniques and to predict biomarkers of the disease at the taxonomic level [7]. The feature selection is based on grouping groups of microorganisms, especially at three main levels: family, order, and genus. It eliminates a large number of less important features from the dataset, having levels at the species level and tries to provide a stronger classification by obtaining species that may be important at the end of this grouping [7]. The model in question produces a separate result for each taxonomic class.

During this process, three main steps are implemented: grouping, scoring, and modeling approaches and general idea of tool was illustrated in the referenced Figure 2.3.1.

2.3.1 Grouping Component

This part of the model arranges features into smaller groups using existing biological knowledge to guide the process. In this step, the dataset with three main different taxonomic levels is grouped one by one as family, order, and genus according to the same taxonomic levels. Each specific species has its own abundance values found in that sample. In this component, Levene's test statistical approach, which tests the variance homogeneity, was used to pick meaningful groups based on the p-value.

2.3.2 Scoring Component

After grouping, the 'G' (Grouping) component generates multiple microorganism groups, which are evaluated by the 'S' (Scoring) component. For each group, the training data is split into 90% for training and 10% for testing. Random Forest classifier is trained with the training set and tested with the remaining data to calculate performance metrics such as accuracy, specificity, and precision. This approach was applied for each group with 5-fold Monte Carlo cross-validation and average scores are used to rank the groups. Each group was assigned with higher average scores. A ranking was made between the groups according to the this average value. The "k" groups with the highest value, for example, 10 top groups, are selected.

2.3.3 Modeling Component

These selected 10 groups are entitled to proceed to the next stage, the "modeling" step. MicrobiomeGSM is re-trained with Random Forest in the 100-fold Monte Carlo cross-validation setting on the best groups. In the modeling step, to handle the variation in taxonomic group rankings obtained from multiple runs, rank aggregation method (Robust Rank Aggregation algorithm [35]) are used. The RobustRankAggreg R package helps combine these rankings and assigns a p-value to each group, allowing researchers to identify the most statistically significant ones. The RobusRankAggreg algorithm aims to check if there is any random sequence for the features.

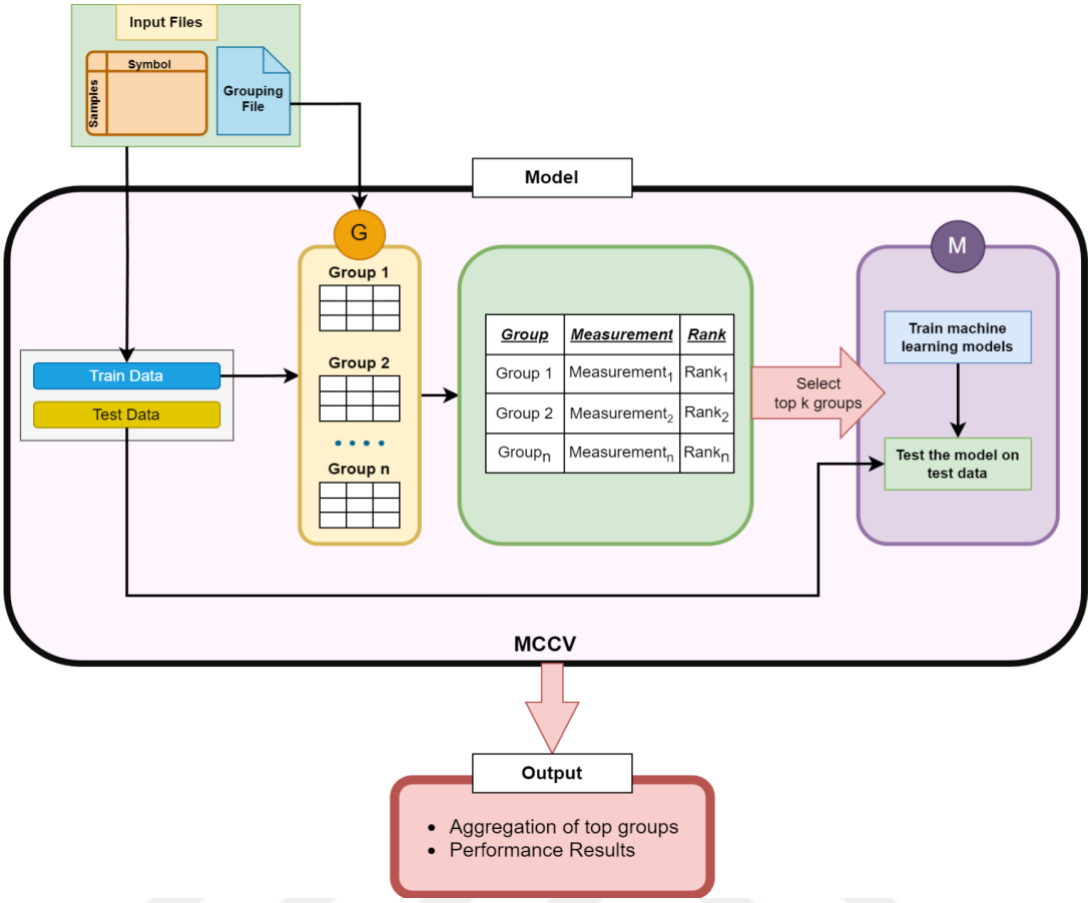


Figure 2.3. 1: Working principle of MicrobiomeGSM tool and explaining the grouping-scoring-modeling approach simply.

2.4 Traditional Feature Selection Approach

The total dimension is large for classification and prediction operations. Therefore, feature selection and data cleaning processes can be important steps to reduce the dimensionality. When literature research is completed about reducing the dimension, it is said that Min Redundancy Max Relevance (mRMR) [36], Lasso [37], Elastic Net [38], and the iterative sure select algorithm [39] methods are used. In this study, as feature selection algorithms, Peng's mRMR [40], Conditional Mutual Information Maximization (CMIM) [41], Fast Correlation Based Filter (FCBF) [42], IG, and XGB were used. The aim of using these algorithms is to increase model performance and identify important biomarker candidates that may be necessary for disease prediction.

The main logic of the mRMR [40] algorithm is to minimize the interdependence among features (min redundancy) while maximizing the connection between each feature and the labeled class element (max relevance). This method weighs the important features and does this with mutual information. In addition, the method applies Forward selection techniques using the sequential search method.

The selection of important features is possible thanks to conditional entropy and mutual information with the class. This situation is called Conditional Mutual Information Maximization [41].

The FCBF method [42] ranks features according to their mutual information with the target class and eliminates those that fall below a defined threshold. This method utilizes the concept of "predominant correlation". In this method, features should have a strong correlation with the target variable but a limited correlation with other features, and this feature selection is made independent of any classifier. Instead of traditional Pearson or Spearman correlations, this method uses Symmetrical Uncertainty, which is based on information theory and relies on Shannon entropy and information gain. Trying to reduce redundancy among features while selecting features is an important process that constitutes FCBF. FCBF can be applied during the analysis of biological science data, thus, meaningful biological insights can be revealed, and possible biomarkers can be determined. This can be achieved by increasing the classification accuracy thanks to FCBF.

In the select K best score method, [43] described a process where features are scored against the class label using a specific function, with the k highest-scoring features being selected. In Python 3, feature selection techniques like CMIM, mRMR, FCBF, and SKB are implemented through the skfeature and sklearn libraries. In the research conducted by Bakir-Gungor and her teammates [44], it was stated that some of these feature selection methods were used, that they achieved good performance, and that these methods were tested on the inflammatory bowel disease-associated metagenomic dataset.

According to the literature review of [45], in the human microbiome research, the mostly previously mentioned future selection methods were applied, the data were analyzed, and favorable outcomes were achieved.

2.4.1 Classification Model Construction

Due to the interpretability of tree-based models and their ability to be easily transformed into rule sets, Random Forest proceeded in subsequent experiments. Moreover, Random Forest is widely regarded as one of the most frequently employed algorithms and has the best performance in human microbiome research, as highlighted by [45, 46].

Specifically, because of the RF's success, the RF predictor node from the H20 library within KNIME was utilized for the solid normal tissue samples and blood normal tissue of all cancer types to compare the model performances. Any extra feature selection process were added for this Random Forest model.

In the traditional feature selection approach, the 6 best-studied feature selection algorithms, namely CMIM, mRMR, FCBF, IG, XGB, and SKB, were applied on KNIME 4.6. In the preliminary analysis, Decision Tree, RF, LogitBoost, AdaBoost, an ensemble SVM with kNN, and an ensemble Logitboost with kNN were applied to evaluate the effects of different classification methods. The main purpose of this approach is to select the important features with multiple selections and try to achieve high performance with these features.

Our experimental design involved utilizing 100-fold Monte Carlo cross-validation (MCCV). This method involves randomly partitioning a subset of the dataset (without replacement) to form the training set, while the remaining data is allocated to the test set, as described by [47]. The process is repeated numerous times, with each iteration generating a new set of training and testing data [47]. For these experiments, 90% of the data was selected for training and 10% for testing. As depicted in Figure 2.4.1.1, feature selection techniques were applied only to the training set. The methodology was implemented using the Konstanz Information Miner (KNIME) platform, as described by [48].

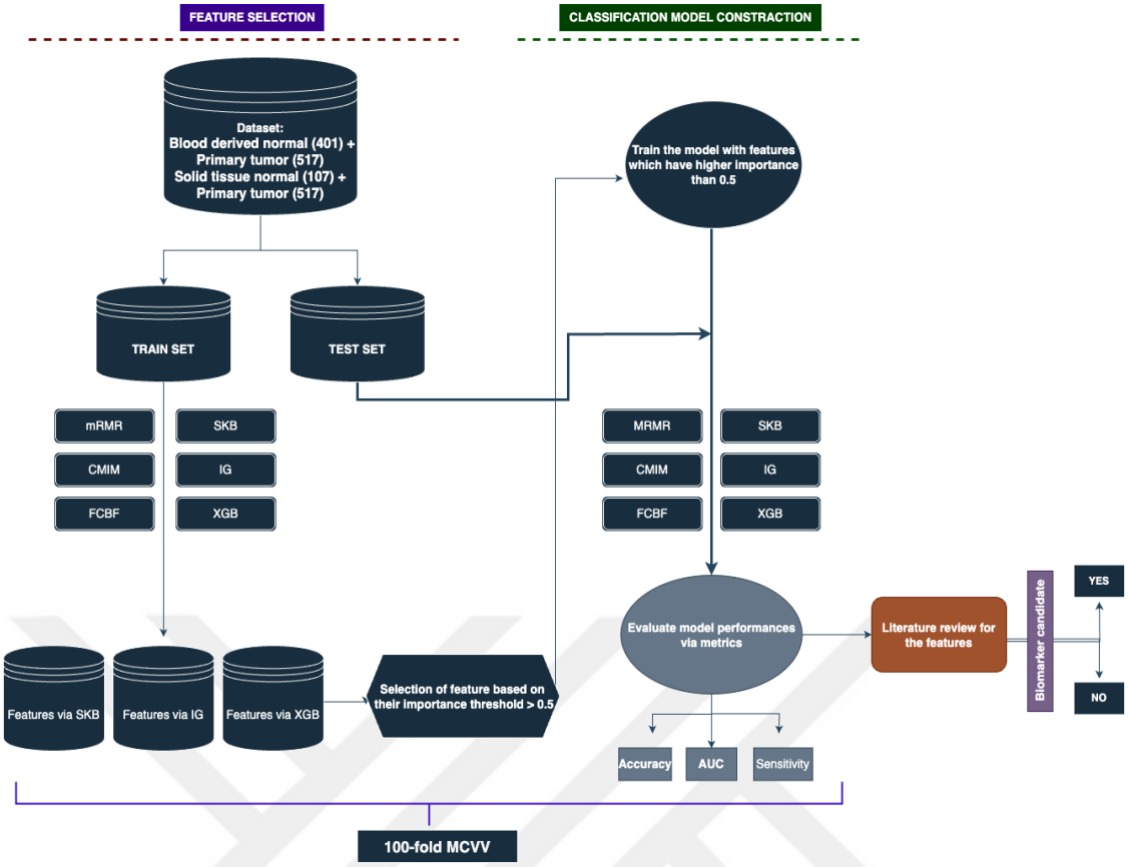


Figure 2.4.1. 1: Shows the process of the traditional feature selection approach. In all steps, Decision Tree, RF, LogitBoost, AdaBoost, an ensemble SVM with kNN, and an ensemble Logitboost with kNN were used.

As a first step, the model performance without specifying any number of features is shown in Appendix Table 1. As seen in the results, SKB, IG, and XGB feature selections produced the best performance results among all feature selection models, and it was decided to use these feature selections in the subsequent experiments. The best (threshold of MinMax Scaled Importance > 0.5) features were obtained based on SKB, IG, and XGB feature selections, and these features were passed through the 6 feature selection algorithms and classification methods mentioned above. The obtained model performances were shown in the Appendix. These steps were performed one by one for both solid normal tissue and blood-derived data of the patient. In addition, the model performances of all cancer data increased to the highest level in the experiments conducted with features above the 0.5 thresholds.

2.5. Model Performance Evaluation

All datasets were tested with both the combination of solid tissue normal and primary tumor to create positive and negative classes. In addition, another positive-negative class was created with the combination of normal blood tissue and primary tumor. These datasets were tested with both traditional feature selection/classification models and the pre-existing biological data-based MicrobiomeGSM tool and also Random Forest model. As shown in Figure 2.5.1, the performance metrics include Accuracy, Precision, Specificity and F1-measure, which are calculated based on true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). In addition, the Area Under the Curve (AUC) is used to evaluate how well the classifier can distinguish between classes. Although MicrobiomeGSM promises high performance, its main purpose is to try to correctly detect biomarker candidates by reducing the number of features it uses during prediction.

The traditional feature selection algorithms that we compared with MicrobiomeGSM are XGBoost, SKB, IG, mRMR, CMIM, and FCBF. These eliminate noise pollution by reducing the number of features and the features that passed the best-performing SKB, XGB, and IG algorithms and had at least 0.5 feature importance value were passed through six feature selection algorithms again.

$$\text{Sensitivity (SEN)} = \text{TP}/\left(\text{TP} + \text{FN}\right),$$

$$\text{Specificity (SPE)} = \text{TN}/\left(\text{TN} + \text{FP}\right),$$

$$\text{Accuracy (ACC)} = \left(\text{TP} + \text{TN}\right)/\left(\text{TP} + \text{TN} + \text{FP} + \text{FN}\right), \quad (2.5.1)$$

Figure 2.5. 1: Shows the formula of accuracy, specificity, and sensitivity provided by all model was used.

2.6. Jaccard Similarity Index Between Species of Cancer Types

Two different word clusters may contain similar or different words or even more than one of the same words [49]. The similarities of these two different sets of words can be measured with the similarity scores obtained from the Jaccard similarity index. It is also known as Intersection-Over-Union and is the common word of two sets divided by all words [50]. The similarity score takes a value between 0 and 1, and the similarity increases as the value approaches 1 [49, 50]. The Jaccard similarity index was applied between the results of five types of cancers from traditional and biological based feature selection approaches. We can see the clear differences and similarities between gastrointestinal cancers based on their possible biomarkers in the results part.

Chapter 3

Results

3.1 Traditional Feature Selection Results

A traditional feature selection approach was utilized for the STAD dataset, and all features (320) were used and evaluated without feature selection as in Table 3.1.1. Among the AUC values, the XGB, IG, and SKB feature selection models have the highest values: 68%, 64%, and 78%, respectively. Then, the experiment was repeated by selecting a total of 12 features from the SKB, IG, and XGB feature selection models, with a threshold of 0.5, as seen in Table 3.1.2. The model performances reached the top point with 12 features. Among the AUC values of the prediction made with fewer features, the XGB, IG, and SKB feature selection models again have the highest values, 84%, 83%, and 83%, respectively. Controlling the number of features means less time and cost. In this respect, using only 12 features for STAD produced comparable evaluation criteria. All approaches, including this, were utilized for all dataset and results were showed in the appendix of this research.

Table 3.1. 1: Performance of the traditional feature selection approach with 320 features in the solid tissue normal and primary tumor results of STAD

FS Model	ML Model	Accuracy	SEN	SPE	F1	Precision	AUC
CMIM	Adaboost	0.67	0.80	0.40	0.76	0.73	0.55
FCBF	Adaboost	0.58	0.68	0.4	0.68	0.69	0.59
IG	DT	0.69	0.81	0.47	0.77	0.76	0.64
SKB	XGBoost	**0.74**	**0.88**	0.46	**0.82**	**0.78**	**0.78**
XGB	Stacking Logitboost with KNN	0.67	0.74	**0.52**	0.74	0.76	0.68
MRMR	Stacking Logitboost with KNN	0.60	0.71	0.38	0.69	0.72	0.60

Table 3.1. 2: Performance of the traditional feature selection approach performances with 12 features in solid tissue normal and primary tumor samples of STAD obtained from 320 full feature results

FS Model	ML Model	Accuracy	SEN	SPE	F1	Precision	AUC
CMIM	LogitBoost	0.65	0.76	0.41	0.74	0.73	0.60
FCBF	Adaboost	0.63	0.64	0.63	0.70	0.78	0.61
IG	RF	0.73	0.78	0.62	0.79	0.81	0.83
SKB	RF	0.74	0.81	**0.61**	0.80	0.82	0.83
XGB	XGBoost	**0.8**	**0.92**	0.56	**0.86**	**0.82**	**0.84**
MRMR	SVM with opt.	0.73	0.79	0.61	0.79	0.81	0.71

After the initial traditional feature selection approach, potential species that could be significant across all cancer types were identified and added to the table. For the ESCA cancer type, which has a total of 314 features, this method, applied to STAD, resulted in the selection of 16 species based on the SKB, IG, and XGBoost models. During this

process, a threshold of 0.5 was employed. Model performances were reassessed for these 16 species. These species are listed in Appendix Table 36 and, as indicated in that table, both ESCA and STAD samples share two species: *Helicobacter pylori* and *Helicobacter acinonychis*. Besides, the other species are unique to each cancer type, and there are three shared genera between the two cancer types: *Streptococcus*, *Prevotella*, and *Helicobacter*. When analyzing at the genus level, there is an increase for each cancer type compared to the species level; for instance, *Prevotella* is a dominant genus with 13 different species found in ESCA, STAD, and HNSC cancer types. *Streptococcus*, featuring 9 distinct species, is also among the common genera present in these cancer types. As a result, the traditional FS approach is not enough for selecting biomarkers and it needs to extra feature selection process.

Colon cancer data with rectum colorectal cancer data were combined, and colorectal cancer data were obtained because colorectal cancer is the coexistence of colon and rectum cancer [51]. After organizing the colon cancer and CRC data, the common and unique biomarkers were identified when the steps applied to STAD and ESCA above were used again. Four common biomarker candidates for CRC and COAD have been identified. When examined at the genus level, the most common of these two cancer types is the *Bacteroides* genus, which has eight different species.

Table 3.1. 3: Performance of the traditional feature selection approach's performance in blood tissue normal and primary tumor results of STAD data with 216 feature

FS Model	ML Model	Accuracy	Sensitivity	Specifity	F-measure	Precision	AUC
XGB	RF	0.88	0.87	0.90	0.88	0.91	0.94
SKB	Adaboost	0.84	0.87	0.82	0.84	0.84	0.91
MRMR	RF	0.67	0.42	0.91	0.55	0.82	0.82
IG	LogitBoost	0.89	0.83	0.94	0.88	0.94	0.91
FCBF	DT	0.72	0.70	0.74	0.71	0.75	0.69
CMIM	LogitBoost	0.72	0.52	0.92	0.64	0.86	0.77

Table 3.1. 4: Performance of the traditional feature approach's performance in blood tissue normal and primary tumor results of STAD dataset with 22 best feature

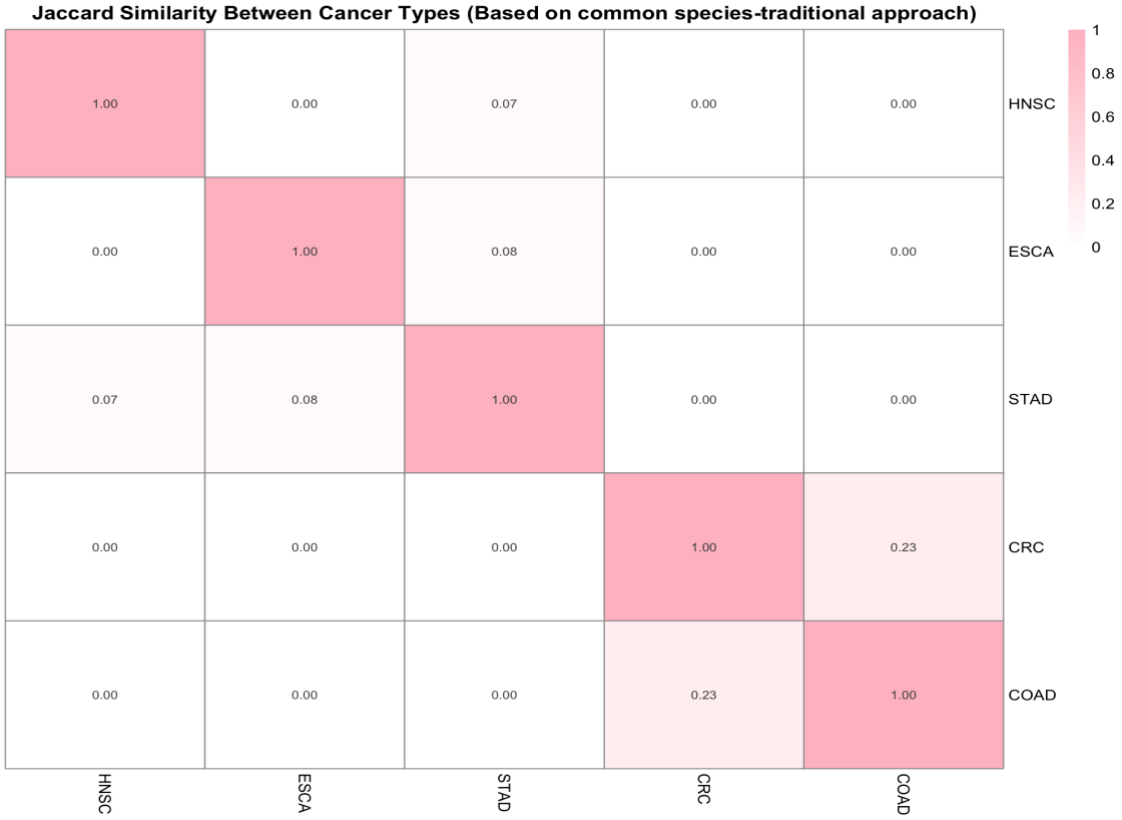
FS Model	ML Model	Accuracy	Sensitivity	Specifity	F-measure	Precision	AUC
XGB	Adaboost	0.84	0.80	0.88	0.83	0.86	0.86
SKB	RF	0.91	0.90	0.91	0.91	0.91	0.96
MRMR	Adaboost	0.84	0.84	0.84	0.84	0.85	0.89
IG	Adaboost	0.83	0.80	0.87	0.83	0.87	0.86
FCBF	RF	0.84	0.82	0.86	0.84	0.86	0.92
CMIM	RF	0.81	0.79	0.82	0.79	0.83	0.88

A traditional feature selection approach was utilized for the STAD dataset from blood-primary tumor samples, and all features (216) were used and evaluated without feature selection as in Table 3.1.3. Among the AUC values, the RF with XGB, Logitboost with IG, and Adaboost with SKB feature selection models have the highest values: 94%, 91%, and 91%, respectively. Then, the experiment was repeated by selecting a total of 22 features from the SKB, IG, and XGB feature selection models, with a threshold of 0.5, as seen in Table 3.1.4. All model performances, including CMIM feature selection model, reached the top point with 22 features. Among the AUC values of the prediction made with fewer features, the Adaboost with XGB, Adaboost with IG, and RF with SKB feature selection models again have the highest values, 86%, 86%, and 96%, respectively. The best AUC values belong to RF with SKB and RF with FCBF, 96% and 92% respectively. CMIM also has a high AUC value compared to the first run's result, like 77% to 88%. Almost all accuracy values are increased after selecting the best features from the first result. Controlling the number of features means less time and cost. In this respect, using only 22 features for STAD produced comparable evaluation criteria.

With the Jaccard similarity index, we can see the correlation in common possible biomarkers between these five types of cancers in solid normal tissue and blood normal tissue samples. In this traditional feature selection approach, we used species that have a greater than 0.5 p-value in each model output to get the model performance. This approach increased the model performance while decreasing the number of species used

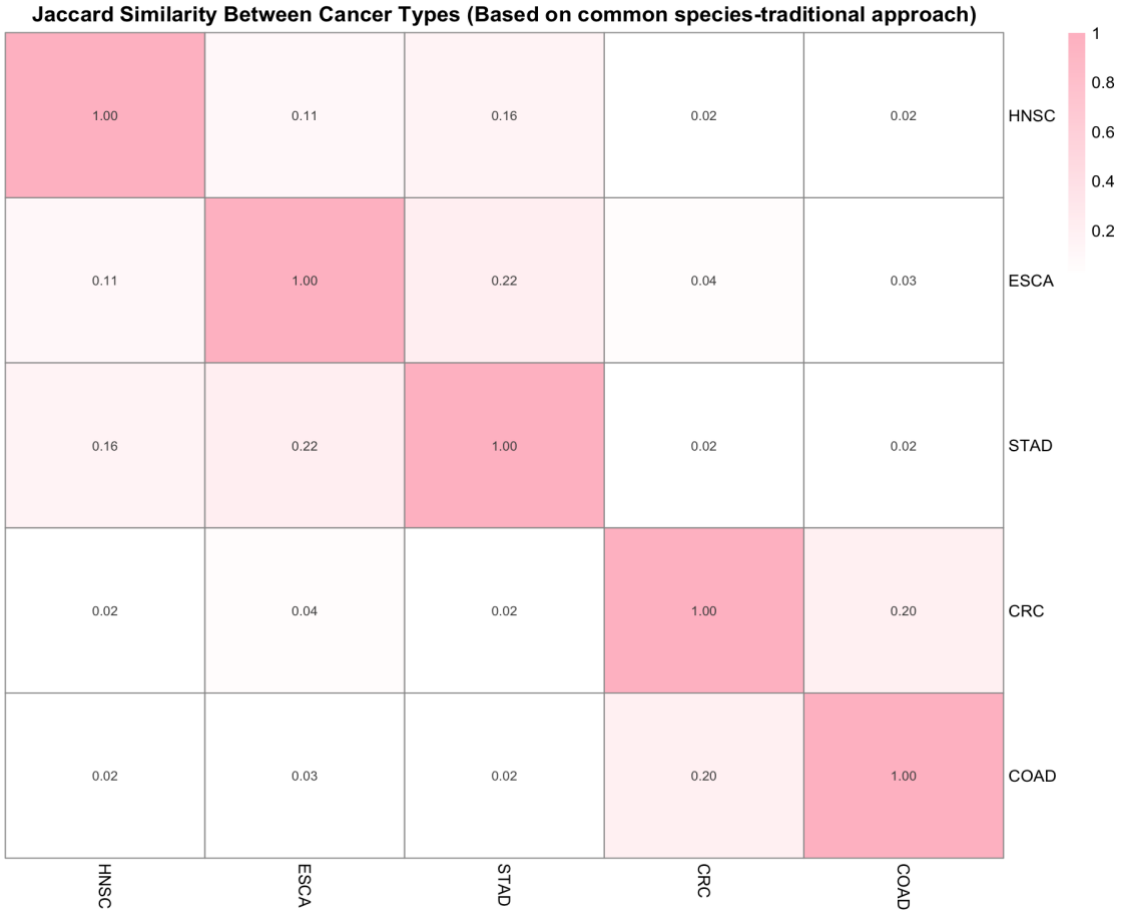
to 15-20. As the number of species decreased, the number of species that could be common also tended to decrease.

Table 3.1. 5: Jaccard Similarity index for the final best species of GI cancer patients with solid tissue normal and primary tumor after traditional feature selection approach



According to this TFSA from the solid tissue sample results in Table 3.1.5, there is no similarity between the ESCA, HNSC, and STAD triple group and the COAD and CRC binary group. When the similarities between ESCA, HNSC, and STAD cancers are examined, the similarity percentages are 8% at the highest. When the COAD and CRC pairs are examined, the similarity increases to 23%.

Table 3.1. 6: Jaccard Similarity index for the final best species of GI cancer patients with blood tissue normal and primary tumor after traditional feature selection approach



The similarities of these five types of cancers from the blood sample and primary tumor are slightly higher than the solid tissue normal results of TFSA, as seen in Table 3.1.6. When the similarities of ESCA, HNSC, and STAD cancers are examined, the highest similarity is between STAD and ESCA with 22%. The similarity between HNSC and ESCA is 11%, while the similarity between HNSC and STAD is 16%. There is a 4% similarity between ESCA and CRC cancer types in terms of biomarker candidates, while there is a 3% similarity between ESCA and COAD. The similarity between the STAD-HNSC and COAD-CRC binary groups is 2%. When the COAD and CRC binary groups are examined, the similarity reaches 20%.

3.2. MicrobiomeGSM Results

As the biological data-based approach, the MicrobiomeGSM model described above was used on KNIME 4.6. The Random Forest Classifier is applied using a standard setup, with 90% of the data used for training and 10% for testing in microbiomeGSM. To address the class imbalance in the datasets, a subsampling strategy is used, aiming for a 2:1 ratio between classes during model training. To evaluate the model's performance, Monte Carlo cross-validation is used. The results are averaged over 10 repetitions to reduce variation and improve reliability. This method helps provide a more accurate and objective assessment of the model; all data are included in the Appendix. The cancer type that gave the best result was COAD from blood and solid tissue normal samples. As seen in the Table 3.2.1, 3.2.2. and 3.2.3., the AUC values of the measurements made at the family-order-genus levels of blood samples are 95%, 94%, and 97%. In comparison, the specificity values are 92%, 97%, and 97%, and the sensitivity values are 87%, 88%, and 87%, respectively. The AUC values of the data obtained from the solid tissue normal samples of COAD are 83%, 88%, and 90%, while the specificity values are 70%, 65%, and 55%, and the sensitivity values are 98%, 96%, and 90%, respectively.

Table 3.2. 1: MicrobiomeGSM performance of blood-derived normal and primary tumor sample of COAD at the family level at 10 iterations

Clusters	Species	Accuracy	Sensitivity	Specificity	F1	AUC	Precision
10	191.7	0.95	0.94	0.96	0.95	0.99	0.96
9	185.1	0.95	0.94	0.96	0.95	0.99	0.96
8	177.4	0.95	0.93	0.96	0.94	0.98	0.96
7	165.8	0.95	0.94	0.96	0.95	0.99	0.96
6	157.8	0.96	0.95	0.96	0.95	0.99	0.96
5	144	0.95	0.94	0.96	0.95	0.98	0.96
4	127.2	0.95	0.93	0.96	0.94	0.98	0.96
3	109.7	0.95	0.93	0.96	0.94	0.98	0.96
2	81.4	0.94	0.91	0.96	0.93	0.98	0.96
1	43.2	0.92	0.87	0.97	0.91	0.95	0.97

Table 3.2. 2: MicrobiomeGSM performance of blood-derived normal and primary tumor sample of COAD at order level at 10 iterations

Clusters	Species	Accuracy	Sensitivity	Specificity	F1	AUC	Precision
10	260	0.94	0.92	0.95	0.93	0.94	0.95
9	257.1	0.94	0.92	0.95	0.93	0.95	0.95
8	254.2	0.93	0.91	0.94	0.92	0.94	0.95
7	248.6	0.94	0.92	0.95	0.93	0.94	0.95
6	242.1	0.93	0.92	0.93	0.93	0.95	0.94
5	235.6	0.92	0.90	0.93	0.91	0.95	0.94
4	230.5	0.92	0.91	0.93	0.92	0.95	0.94
3	216.2	0.92	0.90	0.94	0.92	0.95	0.95
2	198	0.91	0.89	0.93	0.91	0.95	0.94
1	112.8	0.90	0.88	0.92	0.90	0.94	0.92

Table 3.2. 3: MicrobiomeGSM performance of blood-derived normal and primary tumor sample of COAD at genus level at 10 iterations

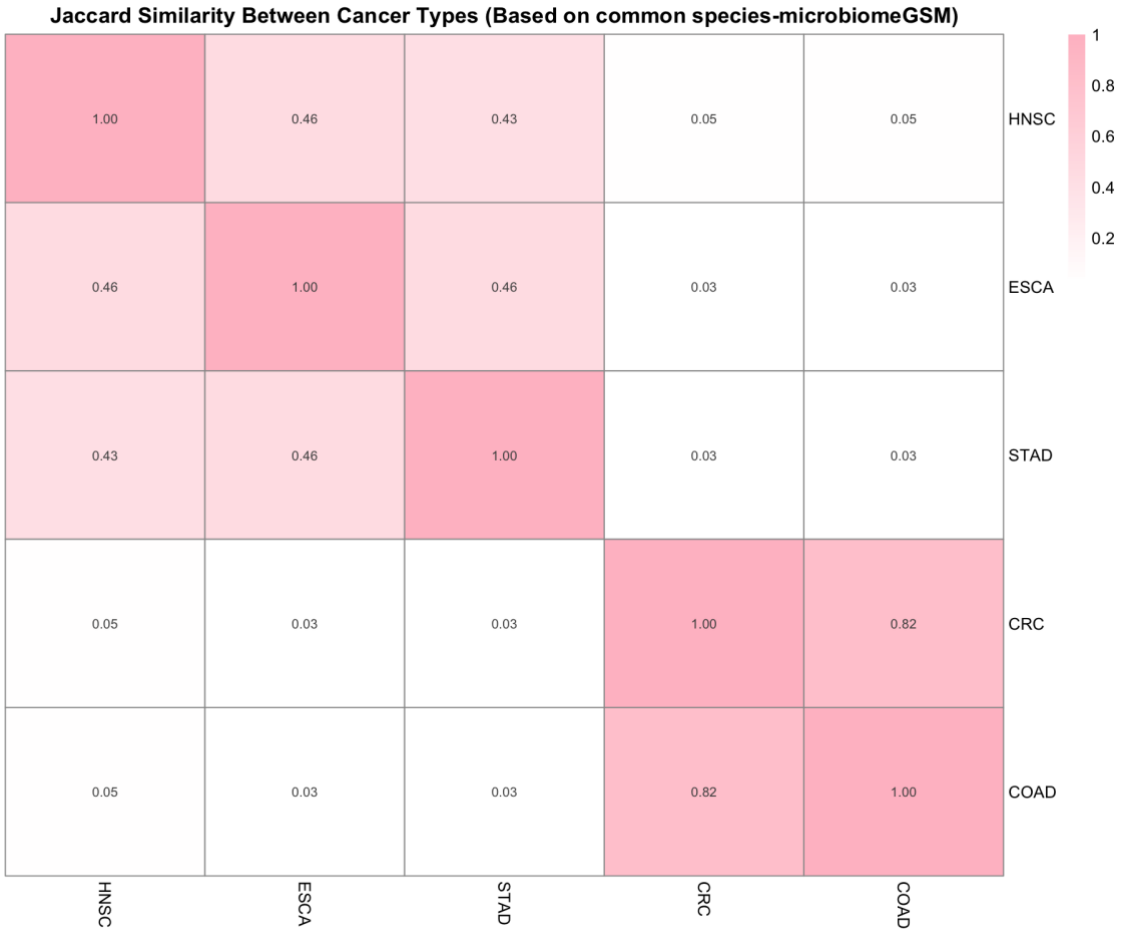
Clusters	Species	Accuracy	Sensitivity	Specificity	F1	AUC	Precision
10	88.9	0.92	0.86	0.98	0.91	0.97	0.98
9	77.3	0.94	0.88	0.99	0.93	0.97	0.99
8	73.4	0.94	0.88	0.99	0.93	0.97	0.99
7	69.6	0.93	0.88	0.98	0.92	0.97	0.98
6	62.4	0.93	0.88	0.98	0.92	0.97	0.98
5	57.1	0.91	0.85	0.97	0.90	0.97	0.97
4	54	0.93	0.87	0.98	0.92	0.97	0.98
3	47.1	0.93	0.87	0.99	0.92	0.97	0.99
2	41.3	0.92	0.87	0.97	0.91	0.97	0.97
1	30	0.92	0.87	0.97	0.91	0.97	0.97

Table 3.2. 4: Model comparison for COAD dataset with normal blood tissue and primary tumor

Approaches	Species	Accuracy	SN	SP	F1	AUC	Precision
Microbiome GSM-genus	30	0,92	0,87	0,97	0,91	0,97	0,97
TFSA-IG-XGB-(2. Run)	23	0.96	0.97	0.94	0.96	0.98	0.95
RF	682	0,95	0,94	0,97	0,95	0,97	0,97

This table 3.2.4 shows the differences between the used approaches for the colon cancer dataset with blood tissue as a negative control. Approaches are microbiomeGSM, Traditional feature selection approach with best FS (IG) and ML (XGB) in the second running, and Random Forest machine learning algorithm. Almost all performance metric results of all approach are more than 90%. MicrobiomeGSM used 30 different species, TFSA best FS (IG) and ML (XGB) in the second running used 23 species Random Forest doesn't use feature selection process to predict the taxonomic biomarker candidates in COAD dataset.

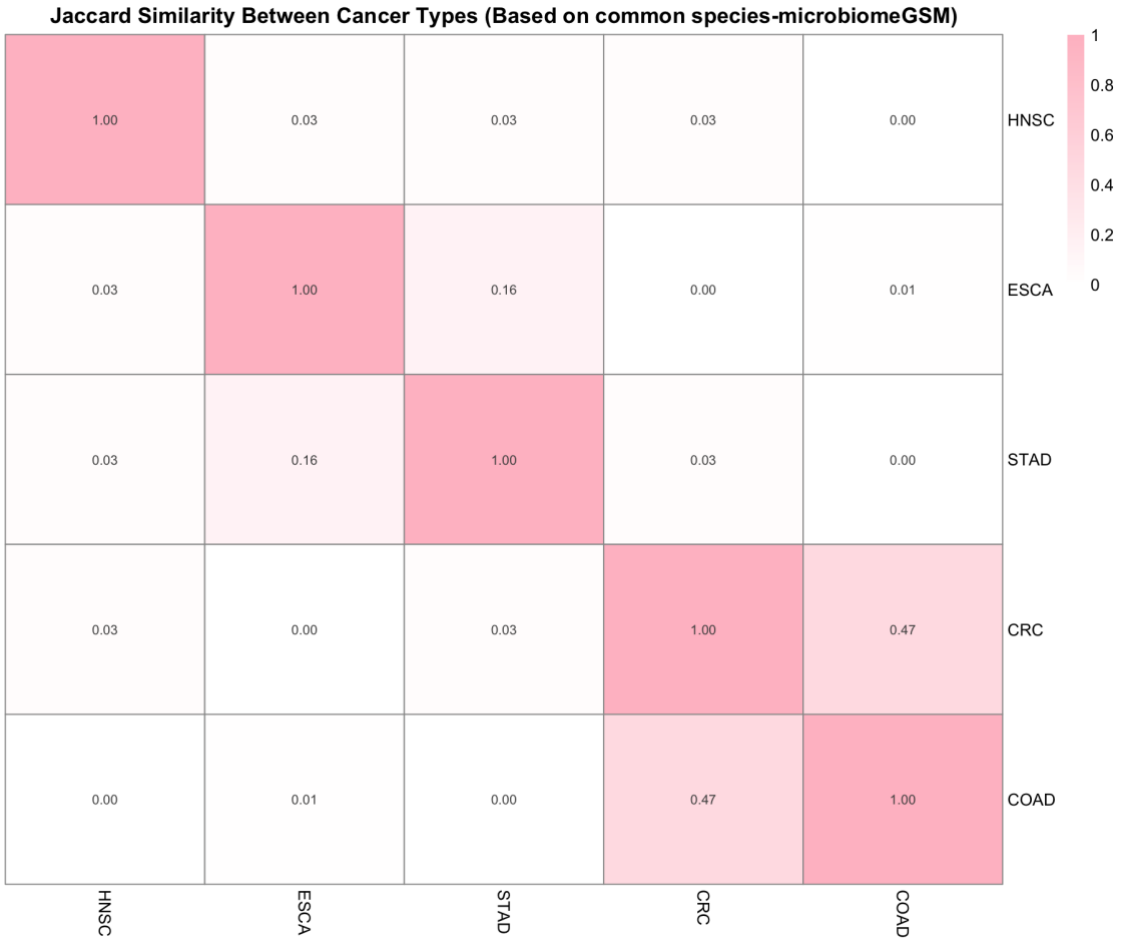
Table 3.2. 5: Jaccard similarity between GI cancer types with blood normal tissue and primary tumor from microbiomeGSM



The 15 most important genera and their top 5 species from blood-primary tumor samples were used in this similarity analysis. According to Table 3.2.5, the Jaccard similarity index between CRC and COAD is 0.82, indicating an 82% microbiome similarity between primary tumors when blood is used as a negative control. This suggests that CRC and COAD form a distinct group separate from other GI cancer types, with shared species comprising 81% of the key taxa. HNSC, ESCA, and STAD show moderate similarity, with Jaccard indices of 0.43 and 0.46, suggesting that nearly half of their key microbial species overlap. In contrast, the similarity between HNSC and CRC-COAD is only 5%, while ESCA and STAD share just 3% similarity with CRC-COAD.

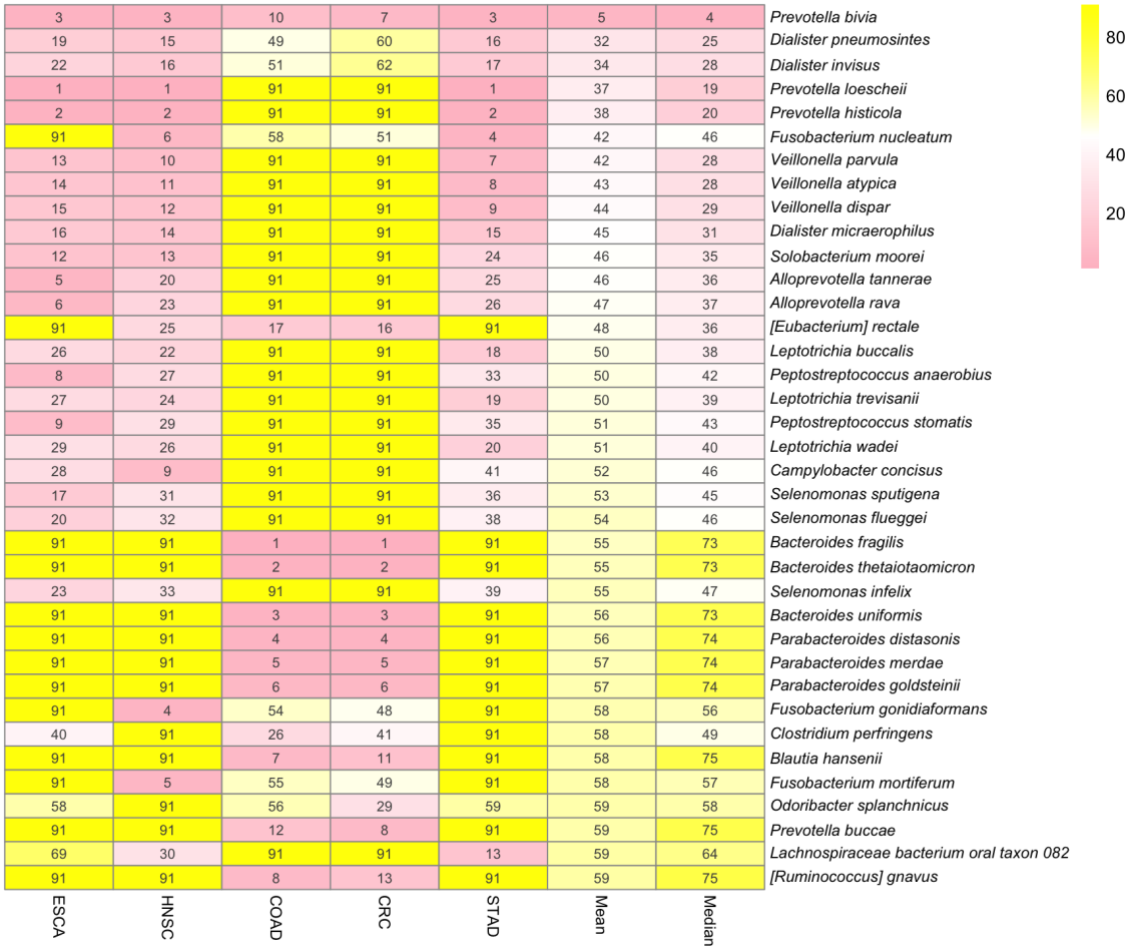
These findings indicate the presence of potential biomarker candidates that may be shared across GI cancers. While each group retains some common species, their overall microbiota compositions remain distinct.

Table 3.2. 6: Jaccard similarity between GI cancer types with solid normal tissue and primary tumor from microbiomeGSM



The 15 most important genera and their top 5 species from normal tissue-primary tumor samples were used in this similarity analysis. According to the Jaccard similarity index Table 3.2.6, CRC and COAD have 47% microbiome similarity between primary tumors when normal solid tissue of the patient is used as a negative control. While STAD and CRC cancer types share 3% common biomarker candidates, STAD and COAD have no common biomarker candidates. While STAD and ESCA have 16% same important microorganisms, STAD shares just 3% similarity with HSNC.

Table 3.2. 7: Represent the rank of taxonomic biomarker candidates which are common at least for two cancer types and are in the top 35 genus and their top 3 species from the blood tissue sample of patients.



This heatmap Table 3.2.7. compares the top 35 genus and their top 3 species taxonomic biomarker candidates across different GI cancer types based on p-values of species from the MicrobiomeGSM model from the blood tissue sample. This heatmap is ordered based on mean value so, we can see the more common and important biomarker for each cancer with one figure. Mean represent the average of the importance places of biomarker candidates for all cancer types. Median shows middle of the importance places of specific biomarker candidate for each cancer types. *Provetella bivia* is the third most important microorganism for ESCA, HNSC, and STAD, while the 10th most important microorganism for colon cancer and the 7th microorganism for CRC. *Dialister pneumosintes* is the 15th microorganism with the best p-value in head and neck cancer, 16th in stomach cancer, 19th in esophageal cancer, 49th in colon cancer, and 60th in colorectal cancer. *Dialister invisus* is common and important species for five cancer types as well. *Provetella loescheii* and *Provetella histicola* may be first two essential biomarker

candidates for ESCA, HNSC and STAD. 91st place of species means this species is not important among the first 35 genus and their top 3 species of cancers. *Fusobacterium nucleatum* is another most common biomarker candidate with 4 cancer types, HNSC, COAD, CRC and STAD. Another *Fusobacterium* species *Fusobacterium conidiaformans* and *Fusobacterium mortiferum* are shared species in HNSC, COAD and CRC. *Veillonella parvula, Veillonella atypica* and *Veillonella dispar* are shared species ESCA, HSNC and STAD from 7th to 15th place. While *Solobacterium moorei* is common for STAD, ESCA and HNSC, *Eubacterium rectale* are shared among HNSC, COAD and CRC.

Table 3.2. 8: Represent the rank of taxonomic biomarker candidates which are common at least for two cancer types and are in the top 40 species from the normal solid tissue sample of patients.

ESCA	STAD	CRC	HNSC	COAD	Mean	Median	
51	42	57	46	75	54	53	*Bulleidia extructa*
18	55	117	117	4	62	59	*Clostridium perfringens*
1	79	117	28	117	68	74	*Streptococcus anginosus*
59	3	47	117	117	69	64	*Aggregatibacter aphrophilus*
61	4	48	117	117	69	65	*Aggregatibacter segnis*
62	5	49	117	117	70	66	*Aggregatibacter sp. oral taxon 458*
2	77	117	38	117	70	74	*Streptococcus intermedius*
117	117	1	117	1	71	94	*Bacteroides fragilis*
117	117	2	117	2	71	94	*Bacteroides thetaiotaomicron*
117	117	3	117	3	71	94	*Bacteroides uniformis*
117	32	91	49	68	71	70	*Centipeda periodontii*
3	78	117	44	117	72	75	*Streptococcus constellatus*
117	117	7	117	5	73	95	*Clostridium paraputrificum*
7	6	117	117	117	73	95	*Helicobacter pylori*
117	117	8	117	6	73	95	*Clostridium sp. ASF356*
117	117	6	117	10	73	95	*Faecalitalea cylindroides*
9	7	117	117	117	73	95	*Helicobacter acinonychis*
4	14	117	117	117	74	95	*Streptococcus infantis*
10	9	117	117	117	74	96	*Helicobacter suis*
117	117	12	117	7	74	96	*Ruminococcus callidus*
117	117	13	117	8	74	96	*Ruminococcus lactaris*
117	117	14	117	9	75	96	*Ruminococcus sp. 5 1 39BFAA*
5	18	117	117	117	75	96	*Streptococcus sp. oral taxon 431*
117	117	53	13	77	75	76	*Fusobacterium gonidiaformans*
117	117	54	14	78	76	77	*Fusobacterium mortiferum*
117	117	20	68	60	76	72	*Roseburia hominis*
32	68	117	117	49	77	72	*Oribacterium sp. oral taxon 078*
117	117	32	1	117	77	97	*Prevotella bryantii*
117	117	17	117	16	77	97	*Subdoligranulum variabile*
33	69	117	117	50	77	73	*Oribacterium sp. oral taxon 108*
117	117	21	69	62	77	73	*Roseburia faecis*
117	117	18	117	17	77	97	*Subdoligranulum sp. 4 3 54A2FAA*

100
80
60
40
20

Table 3.2.8. compares the top 40 genus and their top 3 species taxonomic biomarker candidates across different GI cancer types based on p-values of species from the MicrobiomeGSM model from solid issue normal and primary tumor samples. These

candidates are common for at least two cancer types. All results are obtained from the microbiomeGSM model. Among these top biomarker candidates for each cancer type, HNSC has the fewest shared candidates with other cancer types. HNSC shares the *Prevotella byrantii* and *Fusobacterium* species with another cancer type, CRC. *Bulleidia extructa* has modarete important placed in disease but it is most common species for each cancer. *Bulleidia extructa* is the 42th microorganism with the best p-value in stomach cancer, 46th in head and neck cancer, 51th in esophageal cancer, 57th in colorectal cancer, and 75th in colon cancer.

3.3. Random Forest Results

Table 3.3. 1: The prediction results of the Random Forest model from ESCA with blood-derived normal tissue as a negative control.

row ID	Accuracy	Recall	Specificity	Precision	F1	AUC	Log Loss
Mean	0,86	0,93	0,79	0,84	0,88	0,87	0,48
Sd	0,07	0,09	0,13	0,09	0,06	0,08	0,12

The main idea of this RF experiment was to compare the performance of models. Microorganisms were not ranked and considered in this result section. Table 3.2.1 presents the results of the Random Forest prediction algorithm. The MicrobiomeGSM tool also employs a Random Forest model, and this table specifically evaluates the performance of the RF model alone. The results demonstrate the high performance of this model, indicating that MicrobiomeGSM was developed using the optimal model. All performance metrics reached to almost 80% or above for the ESCA dataset with blood tissue.

Chapter 4

Discussion

A review of the literature reveals no prior studies that have simultaneously investigated biomarker identification across five types of gastrointestinal cancers—stomach/gastric, esophageal, colon, colorectal, and head and neck cancer. This thesis aims to identify both unique and shared biomarker candidates present in the microbiota of these cancers using tissue and blood samples. These biomarkers may play a significant role in cancer progression, exerting either beneficial or detrimental effects on human health [52]. Additionally, this thesis seeks to determine whether common biomarker candidates among these cancer types could serve as indicators of metastasis. In this way, the doctor knows that cancer may develop in the patient's other tissue and can give the drugs accordingly by keeping the patient under observation. In short, the doctor can detect a possible metastasis early. The possibility of metastasis can be investigated in medicine with other methods, but a detection at the microorganism level is not possible. Biomarkers that exhibit carcinogenic properties can settle in the relevant environment and affect the microbiota of the location, triggering the development of cancer.

The MicrobiomeGSM model was developed to identify biomarkers from biological data using the Grouping-Scoring-Modeling approach, providing species-level insights. This model was compared with the traditional feature selection method, consisting of feature selection models such as Fast Correlation Based Filter, Select K Best, Extreme Gradient Boosting, Conditional Mutual Information Maximization, Maximum Likelihood and Minimum Redundancy, and Information Gain. The Random Forest model was also used alone and evaluated in terms of performance with the other two approaches. Since Random Forest, which contains multiple decision trees, generally gives good results, it was used as the main algorithm in the microbiomeGSM model, and when the results are considered, it shows good performance on the dataset, as in Table 3.3.1. Because good results may have several reasons; microbiome datas contain several

thousands features and RF which is ensemble learning approach can handle them without overfitting, microbiome data generally contain outlier and noise, RF can robust to them and reduce the overfitting chance, RF can estimate and provide the feature importance value of all species which shows the relationship between species and diseases, RF doesn't require a specific data distribution and it is flexible for various microbiome data, RF can handle the dataset with missing values [45].

After the RF, while Adaboost in the second run gave an acceptable performance with fewer features for blood samples, IG with Logitboost in the first run showed better performance than Adaboost. In the solid sample performance, we used stacking Logitboost with KNN, which is mentioned by Marcos-Zambrano [45], and it provides the highest SPE value among other models for STAD dataset with 320 features. In the stacking approaches, Logitboost provides class possibilities, KNN gives class prediction using the result from Logitboost, which is why it is more powerful than one model. While the remaining good results of STAD dataset with 320 features are provided by SKB with XGB approaches, STAD with 12 features are predicted by XGB with XGB approaches. XGBoost demonstrated moderate to high performance on solid tissue data from STAD patients using a reduced feature set. In the blood sample analysis of the STAD dataset with all 216 features, the combination of XGBoost and Random Forest following feature selection yielded the best classification performance. After an initial run, 22 features were selected from the original 216 using the SKB method, and the Random Forest model achieved the highest performance with accuracy: 91%, sensitivity: 90%, specificity: 91%, F-measure: 91%, precision: 91%, and AUC: 96%.

In multiple feature selection and multiple classifier-traditional feature selection workflow, Random Forest, LogitBoost (occasionally stacked with KNN), SVM, and AdaBoost have consistently demonstrated high classification performance in both blood and solid tissue samples. Regarding feature selection, Information Gain, XGBoost, and SelectKBest have frequently yielded superior results. The consistent success of these methods can be attributed to their robustness against high-dimensional data, ability to handle non-linear relationships, and effectiveness in identifying relevant microbial features. RF is robust against overfitting and can handle high-dimensional, sparse data typical of microbiome datasets. Logitboost provides high performance on the dataset with fewer features, as like to the study in the literature [52]. This boosting algorithm focuses

on difficult-to-classify samples by iteratively adjusting weights, enhancing prediction accuracy in noisy and imbalanced datasets, such as those found in cancer microbiome studies [52]. SVMs are effective in high-dimensional spaces and can model complex boundaries using kernel functions. They are particularly useful when the number of features exceeds the number of samples, a common scenario in microbiome data [53]. By combining multiple weak learners, AdaBoost creates a strong classifier that emphasizes misclassified instances, improving performance in datasets with subtle patterns, such as microbial signatures associated with cancer [52, 54]. IG Ranks features based on their ability to reduce entropy, which helps in identifying microbiome features with high discriminative power [52]. XGB, as a gradient boosting method, naturally ranks features by importance while training, enabling selection of the most predictive microbial taxa [54, 55]. SKB is a straightforward method that selects the top k features based on univariate statistical tests and is effective when combined with strong classifiers, especially in high-dimensional microbiome data.

The microbiomeGSM model has several important advantages, such as the opportunity to examine microorganisms in the target disease at taxonomic classes, unlike traditional approaches. It identifies microorganisms not only at the species level but also at the family order and genus levels by grouping their importance ranks, and this allows us to analyze the differences between various taxonomic classes. Moreover, the performance evaluation across distinct diseases, separate samples such as blood and tissue, and taxonomic classes (genus, family, order), and the comparative evaluation with various feature selection algorithms demonstrate the reliability of the MicrobiomeGSM tool. The MicrobiomeGSM model aims to distinguish the entire communities of species, not isolated single microbial entities, and can provide deep biological interpretations. By emphasizing groups of bacteria and viruses instead of single entities, the tool provides a holistic perspective and identifies microbial communities potentially involved in specific diseases.

The model performances obtained from blood samples of all cancer patients are superior to those obtained from normal solid tissue samples, as shown in the results. In this context, detecting potential biomarkers from the blood of cancer patients appears to be more feasible and effective than from solid tissue samples during model training. This

may be attributed to several factors; the foremost being the relatively small sample size of solid tissues. Blood samples generally provide a larger dataset compared to any type of solid tissue, which is crucial in machine learning, as models require a sufficient number of samples to accurately learn underlying patterns. Consequently, solid tissue samples may often be inadequate for robust model training.

Second reason may be that normal tissue from cancer patients could still possess an unbalanced microbiome, and all doctors may take the normal tissue samples in different areas of the organs with cancer, without specific guidance. So, the microbiome may change depending on the area. Maybe this is why our machine learning approach struggled to detect patterns in the dataset using a normal tissue sample as a negative control. According to the literature, when tumors (which are abnormal growths of cells) begin to form, the types of bacteria in the gut can change and this change usually begins near the area where the cells begin to become abnormal, known as dysplastic epithelium. Over time, these changes can spread to other parts of the gut, called the luminal compartment [56].

Also, traditional feature selection approaches and random forest algorithms had similar poor results with the normal solid tissue of the patient. Therefore, our hypothesis which is that an unbalanced microbiome found in the cancerous area may also affect the microbiome in other healthy areas, can be supported. Its mean model could not easily separate the positive and negative controls in the first try, if normal tissue samples are used. For example, feature selection approaches could achieve moderate performances in the second run with features that have high feature importance but, in the first try, all feature selection models failed. However, feature selection approaches included models that performed better than microbiomeGSM in the second run.

Additionally, although blood has traditionally been viewed as sterile, recent research shows that microbial signatures—including whole bacteria and cell-free DNA—can be present, especially in individuals with inflammation or dysbiosis [57]. These microbes likely originate from the gut or oral cavity and translocate into the bloodstream when epithelial barriers are compromised. Most microbial DNA in blood is found in the buffy coat, where immune cells like leukocytes may internalize bacterial components

[57]. Upon examining the results, it is evident that the highest performance results for each cancer type are obtained from blood samples used as a negative control.

When solid tissue samples were used as the negative control, fewer important common biomarker candidates were identified across different cancer types, whereas using blood samples led to the detection of more shared biomarker candidates which have high importance value. Since these negative controls belong to the same person as the positive controls, all factors such as diet, IBM, age, history of disease, and smoking are constant, which can provides experimental reliability. In short, the only thing that changes is the change in the number of microorganisms in different condition.

According to the ranked heatmaps in Tables 3.2.7 and 3.2.8, which list biomarker candidates, COAD and CRC do not share 100% of the same microorganisms; there are notable differences. For example, in Table 3.2.7, *Fusobacterium gonidiaformans* ranks 54th in the COAD dataset but 48th in the CRC dataset. Similarly, in Table 3.2.8 *Fusobacterium gonidiaformans* ranks 77th in COAD but 53th in CRC. In Tables 3.2.7 (blood as a negative control) *Prevotella buccae* ranks 12th in COAD but 8th in CRC. This variation is not specific to the *Fusobacterium gonidiaformans* and *Prevotella buccae*, as other species also show ranking differences. The CRC dataset was created by merging colon cancer and rectal cancer datasets, as they are often not separated in the literature. However, these results suggest that some microbial species might be specific to particular anatomical regions, such as the rectum, indicating that CRC and COAD may not be entirely identical.

It may be suggested that the grouping logic of the model requires further refinement at the species level. Although the model consistently identifies groups, it is currently challenging to find literature that fully supports these groups at the species level. Since most interpretations in the literature are at the genus level, this discussion is mainly supported by genus-level evidence rather than species-level.

[56] The researchers explained that the microbiota composition of early- and late-stage cancers can be different in humans, but this distinction is not available in the dataset

used. So it may affect the type and number of biomarker candidates which were obtained from machine learning models. Also, age-related changes in microbiota are possible in diseases according to the literature [58]. This age distinction is not visible for this study. Additionally, microbiome differences between stool and tissue samples may explain the variation between species found in this study and those reported in stool-based studies. With a larger dataset, it may also become possible to separate cancer types more clearly based on their anatomical location.

4.1 Literature Results of Common Biomarkers

Liang, W., and other researchers [59] conducted a wet lab experiment to investigate the presence of the *Prevotella* genus in stomach cancer tissue. Their positive results support their thesis that *Prevotella* could serve as a potential biomarker for gastric cancer. The number of Campylobacter decreases in stomach cancer, which supports our model result where blood samples containing the top 35 biomarker candidates are used as negative samples [60]. Also, Castañeda-Corzo, G., and coworkers [61] conducted a case-control study about *Prevotella* intermedia, which is in the *Prevotella* genus group and frequently found in oropharyngeal cancer, that is one of the head and neck cancer types. In addition to this study, Moreira's article [23] highlighted that *Prevotella* sp is significantly increased on cancer tissue in the esophagus compared to healthy individual samples. Our results supported that this genus has a chance to be a possible biomarker of the ESCA HNSC and STAD, like in the above research. *P. buccae and P. buccalis* play a role in the generally head and neck infections as well according to this research [62]. Microbiota contents may change depending on age; for example, it has been experimentally supported that *Prevotella bivia* species is associated with age in CRC patients ($P < 0.05$) [63]. In this thesis study, the data were not separated according to age, so this may be another reason for the differences between the results obtained and the results in the literature.

Administration of *Parabacteroides distasonis* alleviates CRC metastasis, with its metabolite 4-HPAA playing a crucial role [64]. This species ranks 4th in CRC and COAD according to result of blood-primary tumor samples. At the genus level, *Parabacteroides* and *Ruminococcus* are abundantly found in samples of stage 3 and 4 colorectal cancer

patients. For example, the *Parabacteroides* genus was observed to be more abundant in cancer tissue than in non-cancerous tissue, while the abundance of this genus was observed to indicate poor prognosis in cancer patients. [65]

Fusobacterium and *Clostridium* genera are the good-modarate ranked in the both result tables. The study [66] results showed that *Clostridium* and *Fusobacterium* genera were more abundant in patients with gastric cancer, but less abundant in healthy individuals. The article suggests that these may be biomarkers for gastric cancer because it has been observed that *Clostridium* and *Fusobacterium* genera in particular colonize and become quite prevalent in the gastric microenvironment of cancer patients. According to [67], head and neck cancer patients can host *Fusobacterium* at a higher rate compared to healthy individuals, with this microorganism showing greater abundance in cancer lesions. Our results showed that *Clostridium* and *Fusobacterium* genera are highly associated with stomach cancer and HNSC, especially for the Fusobacterium genus.

According to the literature, one of the evolved bacterial lineages in the early evolutionary process is *Bacteroidetes*, which affects the development process of colorectal cancer [67]. More than 90% of the microbiota diversity in the colon consists of *Bacteroidetes* and *Actinobacteria*. Animal experiments have shown that some species of *Bacteroides* bacteria disrupt the normal functioning of reactions in the colon, causing abnormal crypt foci and inducing the formation of CRC [68]. As we have seen in Table 3.2.7-8, *Bacteroides fragilis,* which belongs to the *Bacteroides* genus, has the potential to be a common biomarker candidate in colon and colorectal cancer. *Bacteroides fragilis* can cause inflammation in the intestines by producing toxin, and this inflammation can trigger colorectal cancer [69]. Enterotoxigenic form (ETBF) of *Bacteroides fragilis* linked to chronic colitis and colitis-associated colorectal cancer through its toxin BFT. BFT triggers inflammation, activates the STAT3 pathway, promotes Th17 responses, and elevates IL-17 levels, all contributing to tumor development. Clinical and animal studies have shown higher ETBF presence in CRC patients. In contrast, non-toxigenic *B. fragilis* appears to have anti-inflammatory and anti-tumor effects via polysaccharide A acting through TLR2, which suppresses CRC cell proliferation, migration, and epithelial-mesenchymal transition [70].

Dialister has been reported to be associated with several types of cancer, including esophageal, gastric, and head and neck cancers. *Dialister pneumosintes, Dialister micraerophilus,* and *Dialister invisus* are one of the *Dialister* species which is in the cellular content. In esophageal cancer, results are inconsistent: [71] reported a decrease in *Dialister* abundance in saliva samples from cancer patients, whereas [72] observed an increase in tissue samples. For gastric cancer, six studies have investigated this association. Except for Liang et al—which used stool samples and found a decrease in *Dialister*—all other studies using tissue samples reported a higher abundance of *Dialister* in cancer patients [73]. Notably, Liang et al. also observed that *Dialister* increased in postoperative samples compared to preoperative ones. In head and neck cancers, including laryngeal carcinoma and head and neck squamous cell carcinoma, eight studies consistently reported a significant increase in *Dialister* abundance in cancer patients across various sample types, such as saliva, oral swabs, and tissue biopsies. Yang et al. specifically identified *Dialister pneumosintes* in association with head and neck cancer [74].

The potential mechanisms through which *Dialister* may contribute to cancer development include its metabolic byproducts and immunomodulatory effects. *Dialister* produces acetate and lactate**,** which are known to support tumor growth and metastasis**,** while propionate may have anti-tumor properties. Furthermore, *Dialister* has been found to be more active in tumor sites [75]. It is also positively correlated with FoxP3+ regulatory T cells (Tregs)**,** which are known to suppress anti-tumor immune responses and are associated with poor clinical outcomes. Both acetate and lactate have been shown to induce the development of these Tregs, suggesting that *Dialister* may promote cancer progression through metabolic support and immune suppression

Solobacterium moorei has been implicated in two distinct cancer-related contexts: colorectal and gastric. In colorectal tissue, High levels of *Solobacterium moorei* have been identified in tissues from colorectal adenomatous polyps, with a significant positive correlation to local inflammatory responses [76]. This inflammation appears to play a role in weakening the intestinal barrier and accelerating the development of polyps, which are known precursors to colorectal cancer [76].

Experimental findings show that *S. moorei* adheres to HT-29 colorectal cancer cells and enhances their growth [76]. This effect is mediated by a Cna B-type domain-

containing surface protein, which allows the bacterium to interact with integrin α2 and β1 on tumor cells. The binding activates a cascade involving phospho-FAK and the PI3K-AKT-mTOR-C-myc signaling pathway, promoting tumor cell survival and suppressing apoptosis [77]. When integrin α2 and β1 genes were silenced using small interfering RNA, the cancer-promoting influence of *S. moorei* was significantly reduced. In both cell culture and animal models, blocking integrin α2/β1 effectively neutralized its tumor-enhancing activity [77].

In contrast, studies investigating gastric cancer progression observed a reduction in *S. moorei* abundance in patients with precancerous gastric lesions, such as intestinal metaplasia and dysplasia [78]. This depletion pattern resembles the paradoxical behavior of *Helicobacter pylori,* which initiates gastric carcinogenesis but declines as the disease advances [79]. Although *S. moorei* is typically known for its association with halitosis and oral opportunistic infections [76–80], its diminished presence in gastric lesions suggests a potentially complex role in early gastric tumorigenesis. Further research is necessary to determine whether this decrease is protective, incidental, or part of a broader microbial shift during precancerous transformation.

Solobacterium moorei has been linked to a reduced risk of esophageal adenocarcinoma, according to findings by Peters et al. [81]. In contrast, its presence has also been documented in advanced laryngeal squamous cell carcinoma which is part of head and neck cancer types. In a clinical case reported by Sarvari [82], *S. moorei* was isolated in high abundance from the surgical wound of a 43-year-old male patient with stage II laryngeal cancer, who had a history of heavy smoking and alcohol use. Despite undergoing multiple treatments, including radiotherapy and total laryngectomy, the bacterium was found along with other anaerobes such as *S. moorei, Fusobacterium nucleatum, Prevotella nanceiensis,* and *Prevotella buccae*, suggesting a potential role in the microbial environment of head and neck cancers.

Clostridium perfringens enterotoxin (CPE) exhibits contrasting effects in colorectal cancer and gastric cancer, depending on cellular context, receptor expression, and dosage [83]. In colorectal cancer, recent findings suggest that low-dose CPE promotes tumor progression, particularly in sessile serrated adenomas/polyps with dysplasia through its enterotoxin [83]. This toxin impairs tight junctions, leading to activation and nuclear translocation of the YAP protein, a driver of tumor progression especially in

BRAFV600E-mutant cells. These findings suggest a tumor-promoting role for *C. perfringens* in colorectal cancer and underline the importance of microbiota regulation [83].

In contrast, a study by Liang et al. [84] demonstrated that CPE has strong cytotoxic effects on gastric cancer cells that overexpress Claudin-4 (CL4). In SGC7901 cells and mouse xenograft models, CPE induced significant tumor cell death by binding to CL4 and forming pores in the cell membrane. Suppression of CL4 expression reduced this cytotoxicity, confirming the specificity of the interaction. These findings position CPE as a potential therapeutic agent in gastric cancer with high CL4 expression [84].

Streptococcus anginosus has been detected in both gastric cancer and esophageal cancer patients, suggesting its potential role in carcinogenesis [85]. Studies have shown that *S. anginosus* abundance increases with age, possibly due to a heightened risk of infections such as endocarditis and upper gastrointestinal malignancies [86, 87]. In GC cases, *S. anginosus* has been isolated from gastric mucosa and shown to induce a spectrum of gastric tissue damage in mice — including inflammation, atrophy of parietal cells, and early precancerous changes such as metaplasia and dysplasia [88]. Regarding EC, a recent case highlighted a pericardial-esophageal fistula as an early sign of advanced EC, with S. anginosus identified in the resulting abscess [89]. These findings suggest that this bacterium may promote esophageal carcinogenesis through inflammation-driven mechanisms, and that its removal could help reduce cancer risk [87].

Recent studies have highlighted distinct microbial shifts in the oral cavity of gastric cancer patients compared to healthy individuals [60]. In particular, the genera *Veillonella*, *Prevotella*, and *Aggregatibacter* are found in elevated abundance in both saliva and dental plaque samples from GC patients, suggesting a potential role in gastric carcinogenesis through inflammatory or immune-mediated mechanisms [90, 91].

Further research has underscored the significance of *Fusobacterium nucleatum* in GC development. According to literature review, this species exerts strong proinflammatory effects by releasing outer membrane vesicles that activate Toll-like receptors (TLRs), especially TLR4, promoting the secretion of inflammatory cytokines like TNF and IL-8 [92]. These responses contribute to the establishment of a proinflammatory microenvironment that evolves into a tumor-promoting niche [92]. .

Additionally, according to same literature review, *F. nucleatum*'s FadA adhesin protein interacts with E-cadherin on host cells to activate the Wnt/β-catenin signaling pathway, further promoting epithelial inflammation and carcinogenesis [92]. *Campylobacter* species, often co-detected with *F. nucleatum* in tumors, also play a proinflammatory role. Their enrichment—particularly in patients with gastritis—triggers the production of various cytokines and chemokines (e.g., TNF, IL-1β, IL-10, CXCL1/2/9/10), and activates key inflammatory pathways such as NF-κB, STAT, CREB1, and IRF signaling. These responses further contribute to mucosal damage and tumorigenesis [92].

Chapter 5

Conclusions and Future Prospects

5.1 Conclusions

In the last two decades, advances in technology have led to an increase in studies analyzing the microbiomes of patients using disease-specific samples such as tissue, body fluids, and stool. Understanding the relationship between microbiome composition and disease can play a critical role in improving disease diagnosis and guiding treatment strategies. The presence of taxonomic biomarkers may serve as key indicators for disease detection, and machine learning can help uncover these biomarkers from complex data. This thesis aims to explore metagenomic data from gastrointestinal cancers to identify existing patterns and to detect potential common and unique biomarker candidates among these cancers using the MicrobiomeGSM-based approach. While the primary method is the MicrobiomeGSM model, its performance is also compared with traditional feature selection approaches. The model is designed to group microorganisms at different taxonomic levels—family, order, and genus—to discover potential biomarkers by leveraging biological structure. This thesis aims to support future cancer–microbiome studies and highlights the need for more careful wet-lab work to produce disease-related data. More data will help to better understand relationship between the host, microbiome, and diseases.

5.2 Societal Impact and Contribution to Global Sustainability

The MicrobiomeGSM model uses the Grouping-Scoring-Modeling framework to reduce the number of microorganism groups, making the selection process faster and

more efficient. This efficiency speeds up the research process and helps generate quicker insights from the results. The model's approach allows large-scale data to be processed effectively, supports its use in medical settings, and enables personalized predictions by identifying biomarker candidates specific to each individual. These personalized findings can help improve the success of treatment strategies. This thesis supports the “Good Health and Well-being” goal of the United Nations Sustainable Development Goals.

After the removal of the primary tumor in cancer patients, there remains a risk of undetected malignant metastases, which can lead to unexpected cancer recurrence—a factor of critical clinical importance. A patient with one type of gastrointestinal cancer may develop metastasis in another region within the GI tract. Shared biomarker candidates specific to GI cancers could play a vital role in early detection. Since metastatic cancer cells cannot be medically visualized before forming a detectable mass, identifying common biomarkers may help draw attention to potential metastasis sites and guide treatment strategies accordingly. The objectives of this thesis align with the global sustainability goals, specifically supporting the 'Good Health and Well-Being' part because cancer is the most important disease in this world and need to be cured as soon as possible. At the same time, new therapeutic targets can be developed based on identified biomarker candidates or with different approaches.

5.3 Future Prospects

The MicrobiomeGSM model could include more statistical steps when selecting species-level features. Adding extra checks after the p-value at the species level might make the results more reliable. Making this final step more detailed can increase the biological accuracy of the model. Also, a simple recommendation system could be added to suggest antibiotics based on the detected biomarker candidates. The common biomarker candidates found in this study are expected to be tested in wet-lab experiments and used to help guide treatment choices. Specific and true antibiotic selection could also be guided by the identified biomarker candidates.

BIBLIOGRAPHY

[1] N. Wu et al., “Dysbiosis Signature of Fecal Microbiota in Colorectal Cancer Patients,” Microb. Ecol., vol. 66, no. 2, pp. 462–470, Aug. 2013, doi: 10.1007/s00248-013-0245-9.

[2] J. Zhang et al., “Expansion of Colorectal Cancer Biomarkers Based on Gut Bacteria and Viruses,” Cancers, vol. 14, no. 19, p. 4662, Sep. 2022, doi: 10.3390/cancers14194662.

[3] Z. Zhou et al., “Human Gut Microbiome-Based Knowledgebase as a Biomarker Screening Tool to Improve the Predicted Probability for Colorectal Cancer,” Front. Microbiol., vol. 11, 2020, Accessed: Jul. 27, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2020.596027

[4] 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.

[5] 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.

[6] W. Zhu, J.-Z. Wang, Z. Liu, and J.-F. Wei, “The bacteria inside human cancer cells: Mainly as cancer promoters,” Front. Oncol., vol. 12, p. 897330, Aug. 2022, doi: 10.3389/fonc.2022.897330.

[7] 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, Nov. 2023, doi: 10.3389/fmicb.2023.1264941.

[8] S. Kayali, E. Marabotto, and E. Giannini, “Gastrointestinal Tract Cancers, an Increasing Burden of the Modern Era: Epidemiology and Prevention,” Cancers, vol. 15, no. 18, p. 4634, Sep. 2023, doi: 10.3390/cancers15184634.

[9] S. R. Jardim, L. M. P. de Souza, and H. S. P. de Souza, “The Rise of Gastrointestinal Cancers as a Global Phenomenon: Unhealthy Behavior or Progress?,” Int. J. Environ. Res. Public. Health, vol. 20, no. 4, p. 3640, Feb. 2023, doi: 10.3390/ijerph20043640.

[10] S. Paschke et al., “Are Colon and Rectal Cancer Two Different Tumor Entities? A Proposal to Abandon the Term Colorectal Cancer,” Int. J. Mol. Sci., vol. 19, no. 9, p. 2577, Aug. 2018, doi: 10.3390/ijms19092577.

[11] T. I. A. for R. on Cancer (IARC), “Global Cancer Observatory.” Accessed: Jan. 11, 2024. [Online]. Available: https://gco.iarc.fr/

[12] M. Rebersek, “Gut microbiome and its role in colorectal cancer,” BMC Cancer, vol. 21, no. 1, p. 1325, Dec. 2021, doi: 10.1186/s12885-021-09054-2.

[13] Y. Cheng, Z. Ling, and L. Li, “The Intestinal Microbiota and Colorectal Cancer,” Front. Immunol., vol. 11, 2020, Accessed: Aug. 01, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fimmu.2020.615056

[14] C. C. Wong and J. Yu, “Gut microbiota in colorectal cancer development and therapy,” Nat. Rev. Clin. Oncol., vol. 20, no. 7, Art. no. 7, Jul. 2023, doi: 10.1038/s41571-023-00766-x.

[15] J. Yang, X. Zhou, X. Liu, Z. Ling, and F. Ji, “Role of the Gastric Microbiome in

Gastric Cancer: From Carcinogenesis to Treatment," Front. Microbiol., vol. 12, 2021, Accessed: Jan. 11, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2021.641322
[16] B. Zilberstein et al., "DIGESTIVE TRACT MICROBIOTA IN HEALTHY VOLUNTEERS," Clinics, vol. 62, no. 1, pp. 47–56, Jan. 2007, doi: 10.1590/S1807-59322007000100008.
[17] G. Nardone, D. Compare, and A. Rocco, "A microbiota-centric view of diseases of the upper gastrointestinal tract," Lancet Gastroenterol. Hepatol., vol. 2, no. 4, pp. 298–312, Apr. 2017, doi: 10.1016/S2468-1253(16)30108-X.
[18] J. Y. Park et al., "Dysbiotic change in gastric microbiome and its functional implication in gastric carcinogenesis," Sci. Rep., vol. 12, no. 1, Art. no. 1, Mar. 2022, doi: 10.1038/s41598-022-08288-9.
[19] S. Liu et al., "Intestinal bacteria are potential biomarkers and therapeutic targets for gastric cancer," Microb. Pathog., vol. 151, p. 104747, Feb. 2021, doi: 10.1016/j.micpath.2021.104747.
[20] H. Nishi et al., "Head and neck cancer patients show poor oral health as compared to those with other types of cancer," BMC Oral Health, vol. 23, no. 1, p. 647, Sep. 2023, doi: 10.1186/s12903-023-03356-6.
[21] K. Dorobisz, T. Dorobisz, and T. Zatoński, "The Microbiome's Influence on Head and Neck Cancers," Curr. Oncol. Rep., vol. 25, no. 3, pp. 163–171, 2023, doi: 10.1007/s11912-022-01352-7.
[22] D. N. Frank et al., "A Dysbiotic Microbiome Promotes Head and Neck Squamous Cell Carcinoma," Oncogene, vol. 41, no. 9, pp. 1269–1280, Feb. 2022, doi: 10.1038/s41388-021-02137-1.
[23] C. Moreira, C. Figueiredo, and R. M. Ferreira, "The Role of the Microbiota in Esophageal Cancer," Cancers, vol. 15, no. 9, Art. no. 9, Jan. 2023, doi: 10.3390/cancers15092576.
[24] J. Zhou et al., "Gut Microbiota for Esophageal Cancer: Role in Carcinogenesis and Clinical Implications," Front. Oncol., vol. 11, p. 717242, Oct. 2021, doi: 10.3389/fonc.2021.717242.
[25] Z. Lin et al., "Characteristics and interplay of esophageal microbiota in esophageal squamous cell carcinoma," BMC Cancer, vol. 22, no. 1, p. 696, Jun. 2022, doi: 10.1186/s12885-022-09771-2.
[26] G. Menon and B. Cagir, "Colon Cancer," in StatPearls, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: Mar. 14, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK470380/
[27] 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), Sep. 2022, pp. 1–6. doi: 10.1109/ASYU56188.2022.9925551.
[28] C. Zhang et al., "Combined Non-Invasive Prediction and New Biomarkers of Oral and Fecal Microbiota in Patients With Gastric and Colorectal Cancer," Front. Cell. Infect. Microbiol., vol. 12, 2022, Accessed: Oct. 24, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcimb.2022.830684
[29] P. Freitas et al., "Machine learning-based approaches for cancer prediction using microbiome data," Sci. Rep., vol. 13, no. 1, Art. no. 1, Jul. 2023, doi: 10.1038/s41598-023-38670-0.
[30] Y. Wu et al., "Identification of microbial markers across populations in early detection of colorectal cancer," Nat. Commun., vol. 12, no. 1, Art. no. 1, May 2021, doi: 10.1038/s41467-021-23265-y.

[31] M.-X. Li et al., “Using a machine learning approach to identify key prognostic molecules for esophageal squamous cell carcinoma,” BMC Cancer, vol. 21, no. 1, p. 906, Aug. 2021, doi: 10.1186/s12885-021-08647-1.
[32] “The Cancer Genome Atlas Program (TCGA) - NCI.” Accessed: Apr. 10, 2025. [Online]. Available: https://www.cancer.gov/ccg/research/genome-sequencing/tcga
[33] A. B. Dohlman et al., “The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants,” Cell Host Microbe, vol. 29, no. 2, pp. 281-298.e5, Feb. 2021, doi: 10.1016/j.chom.2020.12.001.
[34] B. Bakir-Gungor, M. Temiz, B. Canakcimaksutoglu, and M. Yousef, “Prediction of colorectal cancer based on taxonomic levels of microorganisms and discovery of taxonomic biomarkers using the Grouping-Scoring-Modeling (G-S-M) approach,” Comput. Biol. Med., vol. 187, p. 109813, Mar. 2025, doi: 10.1016/j.compbiomed.2025.109813.
[35] 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.
[36] 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.
[37] R. Tibshirani, “Regression Shrinkage and Selection Via the Lasso,” J. R. Stat. Soc. Ser. B Methodol., vol. 58, no. 1, pp. 267–288, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.
[38] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.
[39] 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, p. 1784, Dec. 2017, doi: 10.1038/s41467-017-01973-8.
[40] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, “Conditional likelihood maximisation: a unifying framework for information theoretic feature selection,” J Mach Learn Res, vol. 13, no. null, pp. 27–66, Ocak 2012.
[41] François Fleuret, “Fast binary feature selection with conditional mutu-al information.,” J. Mach. Learn. Res., vol. 5, no. 1531–1555, 2004.
[42] 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.
[43] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Mach. Learn. PYTHON.
[44] 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, Apr. 2022, doi: 10.7717/peerj.13205.
[45] 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,” Front. Microbiol., vol. 12, p. 634511, Feb. 2021, doi: 10.3389/fmicb.2021.634511.
[46] D. D’Elia et al., “Advancing microbiome research with machine learning: key findings from the ML4Microbiome COST action,” Front. Microbiol., vol. 14, Sep.

2023, doi: 10.3389/fmicb.2023.1257002.
[47] Q.-S. Xu and Y.-Z. Liang, “Monte Carlo cross validation,” Chemom. Intell. Lab. Syst., vol. 56, no. 1, pp. 1–11, Apr. 2001, doi: 10.1016/S0169-7439(00)00122-2.
[48] M. R. Berthold et al., “KNIME: The Konstanz Information Miner,” in Data Analysis, Machine Learning and Applications, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Eds., Berlin, Heidelberg: Springer, 2008, pp. 319–326. doi: 10.1007/978-3-540-78246-9_38.
[49] M. Madhikermi, S. Kubler, J. Robert, A. Buda, and K. Främling, “Data quality assessment of maintenance reporting procedures,” Expert Syst. Appl., vol. 63, pp. 145–164, Nov. 2016, doi: 10.1016/j.eswa.2016.06.043.
[50] I. Rizwan I Haque and J. Neubert, “Deep learning approaches to biomedical image segmentation,” Inform. Med. Unlocked, vol. 18, p. 100297, Jan. 2020, doi: 10.1016/j.imu.2020.100297.
[51] R. Wei, M. Okocha, and S. Lotfollahzadeh, “Rectal Cancer Microsurgery,” in StatPearls, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: Mar. 01, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK585048/
[52] S. Bang, D. Yoo, S.-J. Kim, S. Jhang, S. Cho, and H. Kim, “Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data,” Sci. Rep., vol. 9, no. 1, p. 10189, Jul. 2019, doi: 10.1038/s41598-019-46249-x.
[53] M. Teixeira, F. Silva, R. M. Ferreira, T. Pereira, C. Figueiredo, and H. P. Oliveira, “A review of machine learning methods for cancer characterization from microbiome data,” NPJ Precis. Oncol., vol. 8, p. 123, May 2024, doi: 10.1038/s41698-024-00617-7.
[54] Z. Karwowska et al., “Effects of data transformation and model selection on feature importance in microbiome classification data,” Microbiome, vol. 13, no. 1, p. 2, Jan. 2025, doi: 10.1186/s40168-024-01996-6.
[55] F. Bahrambanan et al., “The development of an efficient artificial intelligence-based classification approach for colorectal cancer response to radiochemotherapy: deep learning vs. machine learning,” Sci. Rep., vol. 15, p. 62, Jan. 2025, doi: 10.1038/s41598-024-84023-w.
[56] E. A. Ivleva and S. I. Grivennikov, “Microbiota-driven mechanisms at different stages of cancer development,” Neoplasia, vol. 32, p. 100829, Oct. 2022, doi: 10.1016/j.neo.2022.100829.
[57] G. Park, Oh ,Suji, Kim ,Minjeong, Jeong ,Yunsun, and G. and Kim, “Common microbial signatures in blood and their amplification in clinical disorders,” Gut Microbes Rep., vol. 2, no. 1, p. 2473450, Dec. 2025, doi: 10.1080/29933935.2025.2473450.
[58] Y. Wu, J. Xu, X. Rong, F. Wang, H. Wang, and C. Zhao, “Gut microbiota alterations and health status in aging adults: From correlation to causation,” Aging Med., vol. 4, no. 3, pp. 206–213, Jun. 2021, doi: 10.1002/agm2.12167.
[59] W. Liang et al., “Tumor-derived Prevotella intermedia aggravates gastric cancer by enhancing Perilipin 3 expression,” Cancer Sci., vol. 115, no. 4, p. 1141, Jan. 2024, doi: 10.1111/cas.16080.
[60] R. Xia et al., “Oral microbiota and gastric cancer: recent highlights and knowledge gaps,” J. Oral Microbiol., vol. 16, no. 1, p. 2391640, doi: 10.1080/20002297.2024.2391640.
[61] G.-J. Castañeda-Corzo, L.-F. Infante-Rodríguez, J.-C. Villamil-Poveda, J. Bustillo, A. Cid-Arregui, and D.-A. García-Robayo, “Association of Prevotella intermedia with oropharyngeal cancer: A patient-control study,” Heliyon, vol. 9, no. 3, p.

e14293, Mar. 2023, doi: 10.1016/j.heliyon.2023.e14293.
[62] E. Könönen, D. Fteita, U. K. Gursoy, and M. Gursoy, "Prevotella species as oral residents and infectious agents with potential impact on systemic conditions," J. Oral Microbiol., vol. 14, no. 1, p. 2079814, doi: 10.1080/20002297.2022.2079814.
[63] Y. Qin et al., "Consistent signatures in the human gut microbiome of old- and young-onset colorectal cancer," Nat. Commun., vol. 15, no. 1, p. 3396, Apr. 2024, doi: 10.1038/s41467-024-47523-x.
[64] J. Yu et al., "Interleukin-10 deficiency suppresses colorectal cancer metastasis by enriching gut Parabacteroides distasonis," J. Adv. Res., Nov. 2024, doi: 10.1016/j.jare.2024.11.024.
[65] J. Liu et al., "Identification of colorectal cancer progression-associated intestinal microbiome and predictive signature construction," J. Transl. Med., vol. 21, no. 1, p. 373, Jun. 2023, doi: 10.1186/s12967-023-04119-1.
[66] Y.-Y. Hsieh et al., "Increased Abundance of Clostridium and Fusobacterium in Gastric Microbiota of Patients with Gastric Cancer in Taiwan," Sci. Rep., vol. 8, no. 1, p. 158, Jan. 2018, doi: 10.1038/s41598-017-18596-0.
[67] J. D. Bronzato, R. A. Bomfim, D. H. Edwards, D. Crouch, M. P. Hector, and B. P. F. A. Gomes, "Detection of Fusobacterium in oral and head and neck cancer samples: A systematic review and meta-analysis," Arch. Oral Biol., vol. 112, p. 104669, Apr. 2020, doi: 10.1016/j.archoralbio.2020.104669.
[68] Y. Cao et al., "Changes in Bacteroides and the microbiota in patients with obstructed colorectal cancer: retrospective cohort study," BJS Open, vol. 7, no. 6, p. zrad105, Nov. 2023, doi: 10.1093/bjsopen/zrad105.
[69] W. T. Cheng, H. K. Kantilal, and F. Davamani, "The Mechanism of Bacteroides fragilis Toxin Contributes to Colon Cancer Formation," Malays. J. Med. Sci. MJMS, vol. 27, no. 4, pp. 9–21, Jul. 2020, doi: 10.21315/mjms2020.27.4.2.
[70] Y. Matsumiya et al., "Clinical significance of Bacteroides fragilis as a potential prognostic factor in colorectal cancer," Anaerobe, vol. 84, p. 102784, Dec. 2023, doi: 10.1016/j.anaerobe.2023.102784.
[71] X. Chen et al., "Oral Microbiota and Risk for Esophageal Squamous Cell Carcinoma in a High-Risk Area of China," PLoS ONE, vol. 10, no. 12, p. e0143603, Dec. 2015, doi: 10.1371/journal.pone.0143603.
[72] D. R. F. Elliott, A. W. Walker, M. O'Donovan, J. Parkhill, and R. C. Fitzgerald, "A non-endoscopic device to sample the oesophageal microbiota: a case-control study," Lancet Gastroenterol. Hepatol., vol. 2, no. 1, pp. 32–42, Jan. 2017, doi: 10.1016/S2468-1253(16)30086-3.
[73] W. Liang et al., "Gut microbiota shifts in patients with gastric cancer in perioperative period," Medicine (Baltimore), vol. 98, no. 35, p. e16626, Aug. 2019, doi: 10.1097/MD.0000000000016626.
[74] C.-Y. Yang et al., "Oral Microbiota Community Dynamics Associated With Oral Squamous Cell Carcinoma Staging," Front. Microbiol., vol. 9, May 2018, doi: 10.3389/fmicb.2018.00862.
[75] S. Yost et al., "Increased virulence of the oral microbiome in oral squamous cell carcinoma revealed by metatranscriptome analyses," Int. J. Oral Sci., vol. 10, no. 4, pp. 1–10, Nov. 2018, doi: 10.1038/s41368-018-0037-7.
[76] S. Yu et al., "Solobacterium moorei promotes the progression of adenomatous polyps by causing inflammation and disrupting the intestinal barrier," J. Transl. Med., vol. 22, p. 169, Feb. 2024, doi: 10.1186/s12967-024-04977-3.
[77] Y. Chen et al., "Solobacterium moorei promotes tumor progression via the Integrin α2/β1-PI3K-AKT-mTOR-C-myc signaling pathway in colorectal cancer," Int. J.

Biol. Sci., vol. 21, no. 4, pp. 1497–1512, Jan. 2025, doi: 10.7150/ijbs.102742.
[78] I. Barrak, A. Stájer, M. Gajdács, and E. Urbán, “Small, but smelly: the importance of Solobacterium moorei in halitosis and other human infections,” Heliyon, vol. 6, no. 10, p. e05371, Oct. 2020, doi: 10.1016/j.heliyon.2020.e05371.
[79] R. M. Ferreira et al., “Gastric microbial community profiling reveals a dysbiotic cancer-associated microbiota,” Gut, vol. 67, no. 2, pp. 226–236, Feb. 2018, doi: 10.1136/gutjnl-2017-314205.
[80] C. Alauzet et al., “A New Look at the Genus Solobacterium: A Retrospective Analysis of Twenty-Seven Cases of Infection Involving S. moorei and a Review of Sequence Databases and the Literature,” Microorganisms, vol. 9, no. 6, Art. no. 6, Jun. 2021, doi: 10.3390/microorganisms9061229.
[81] B. A. Peters et al., “Oral Microbiome Composition Reflects Prospective Risk for Esophageal Cancers,” Cancer Res., vol. 77, no. 23, pp. 6777–6787, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-1296.
[82] K. P. Sárvári et al., “Six cases of Solobacterium moorei isolated alone or in mixed culture in Hungary and comparison with previously published cases,” Anaerobe, vol. 65, p. 102241, Oct. 2020, doi: 10.1016/j.anaerobe.2020.102241.
[83] R. Fujiwara-Tani et al., “Role of Clostridium perfringens Enterotoxin on YAP Activation in Colonic Sessile Serrated Adenoma/Polyps with Dysplasia,” Int. J. Mol. Sci., vol. 21, no. 11, Art. no. 11, Jan. 2020, doi: 10.3390/ijms21113840.
[84] Z.-Y. Liang, X. Kang, H. Chen, M. Wang, and W.-X. Guan, “Effect of Clostridium perfringens enterotoxin on gastric cancer cells SGC7901 which highly expressed claudin-4 protein,” World J. Gastrointest. Oncol., vol. 9, no. 4, pp. 153–159, Apr. 2017, doi: 10.4251/wjgo.v9.i4.153.
[85] K. Xia, Y. Zhou, W. Wang, and Y. Cai, “Streptococcus anginosus: the potential role in the progression of gastric cancer,” J. Cancer Res. Clin. Oncol., vol. 151, no. 4, p. 143, Apr. 2025, doi: 10.1007/s00432-025-06201-1.
[86] E. Morita et al., “Predominant presence of Streptococcus anginosus in the saliva of alcoholics,” Oral Microbiol. Immunol., vol. 20, no. 6, pp. 362–365, Dec. 2005, doi: 10.1111/j.1399-302X.2005.00242.x.
[87] M. Narikiyo et al., “Frequent and preferential infection of Treponema denticola, Streptococcus mitis, and Streptococcus anginosus in esophageal cancers,” Cancer Sci., vol. 95, no. 7, pp. 569–574, Jul. 2004, doi: 10.1111/j.1349-7006.2004.tb02488.x.
[88] V. Isakov, “Autoimmune gastritis studies and gastric cancer: True renaissance or bibliometric illusion,” World J. Gastroenterol., vol. 30, no. 32, pp. 3783–3790, Aug. 2024, doi: 10.3748/wjg.v30.i32.3783.
[89] C. M. Peixoto Vieira Viseu Ferreira, S. Souaf, A. Martínez Monzonís, M. M. El-Diasty, and A. L. Fernandez, “Purulent Pericarditis as the First Manifestation of Esophageal Carcinoma,” Port. J. Card. Thorac. Vasc. Surg., vol. 31, no. 1, pp. 59–62, May 2024, doi: 10.48729/pjctvs.423.
[90] J. Shu et al., “Role of salivary glycopatterns for oral microbiota associated with gastric cancer,” Int. J. Biol. Macromol., vol. 209, no. Pt A, pp. 1368–1378, Jun. 2022, doi: 10.1016/j.ijbiomac.2022.04.133.
[91] S. Kageyama et al., “Characteristics of the Salivary Microbiota in Patients With Various Digestive Tract Cancers,” Front. Microbiol., vol. 10, p. 1780, 2019, doi: 10.3389/fmicb.2019.01780.
[92] R. Xia et al., “Oral microbiota and gastric cancer: recent highlights and knowledge gaps,” J. Oral Microbiol., vol. 16, no. 1, p. 2391640, 2024, doi: 10.1080/20002297.2024.2391640.

APPENDIX

Additional Results

Table A 1. COAD-READ (CRC) dataset at family level in blood tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	181.1	0.89	0.84	0.95	0.89	0.93	0.95
9	170.6	0.90	0.85	0.96	0.90	0.94	0.96
8	162.9	0.90	0.84	0.96	0.89	0.93	0.96
7	155.3	0.90	0.85	0.95	0.89	0.94	0.95
6	148	0.89	0.84	0.95	0.89	0.93	0.95
5	125.6	0.90	0.85	0.96	0.90	0.94	0.96
4	113.9	0.90	0.84	0.96	0.89	0.93	0.96
3	87.1	0.90	0.85	0.96	0.90	0.93	0.96
2	70.8	0.90	0.85	0.95	0.89	0.93	0.95
1	38.3	0.90	0.85	0.96	0.90	0.93	0.96

Table A 2. COAD-READ (CRC) dataset at order level in blood tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	258.1	0.93	0.89	0.96	0.93	0.96	0.97
9	254.7	0.92	0.89	0.95	0.92	0.96	0.95
8	250.5	0.93	0.90	0.96	0.93	0.96	0.97
7	246	0.93	0.91	0.96	0.93	0.96	0.96
6	240.2	0.93	0.92	0.95	0.94	0.97	0.95
5	236.4	0.93	0.90	0.96	0.93	0.97	0.96
4	229.9	0.93	0.90	0.96	0.93	0.96	0.96
3	226	0.93	0.90	0.96	0.93	0.97	0.96
2	198	0.92	0.90	0.95	0.92	0.96	0.95
1	114.4	0.93	0.93	0.94	0.94	0.96	0.95

Table A 3. COAD-READ (CRC) dataset at genus level in blood tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specifity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	95.5	0.93	0.92	0.94	0.93	0.96	0.95
9	88.2	0.92	0.90	0.94	0.92	0.96	0.95
8	85.2	0.92	0.89	0.95	0.92	0.96	0.96
7	79.5	0.92	0.90	0.94	0.92	0.97	0.94
6	70.6	0.93	0.91	0.95	0.93	0.96	0.96
5	62.8	0.93	0.90	0.96	0.93	0.97	0.96
4	53.2	0.93	0.91	0.95	0.93	0.96	0.96
3	46.5	0.93	0.91	0.96	0.93	0.96	0.96
2	41.8	0.93	0.91	0.96	0.93	0.97	0.97
1	33	0.93	0.89	0.96	0.93	0.96	0.97

Table A 4. ESCA dataset at family level in blood tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	89.7	0.72	0.66	0.80	0.72	0.84	0.82
9	84.5	0.76	0.70	0.83	0.76	0.86	0.84
8	79.5	0.77	0.72	0.83	0.77	0.85	0.84
7	74.7	0.74	0.70	0.80	0.74	0.82	0.82
6	66	0.71	0.64	0.80	0.70	0.83	0.80
5	60.7	0.74	0.68	0.83	0.74	0.84	0.84
4	55.6	0.72	0.68	0.78	0.72	0.84	0.80
3	50.4	0.71	0.60	0.85	0.67	0.81	0.81
2	45.2	0.71	0.60	0.85	0.67	0.83	0.81
1	28.8	0.73	0.62	0.88	0.70	0.80	0.88

Table A 5. ESCA dataset at order level in blood tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	121	0.80	0.80	0.80	0.81	0.82	0.84
9	117.1	0.74	0.70	0.80	0.74	0.83	0.81
8	108.5	0.74	0.72	0.78	0.74	0.81	0.80
7	103.9	0.78	0.76	0.80	0.78	0.82	0.83
6	94.2	0.74	0.72	0.78	0.74	0.84	0.79
5	84.2	0.74	0.70	0.80	0.74	0.83	0.81
4	79.1	0.79	0.74	0.85	0.78	0.85	0.88
3	65.6	0.76	0.72	0.80	0.76	0.83	0.82
2	48.1	0.74	0.70	0.80	0.74	0.81	0.81
1	37.2	0.73	0.68	0.80	0.72	0.79	0.82

Table A 6. ESCA dataset at genus level in blood tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	62.6	0.74	0.70	0.80	0.75	0.85	0.83
9	59.6	0.71	0.66	0.78	0.71	0.83	0.81
8	57.4	0.74	0.72	0.78	0.75	0.86	0.81
7	55.1	0.76	0.72	0.80	0.76	0.85	0.85
6	52.8	0.76	0.70	0.83	0.75	0.85	0.86
5	48.1	0.73	0.72	0.75	0.75	0.82	0.80
4	45.2	0.68	0.62	0.75	0.68	0.84	0.77
3	35.8	0.62	0.52	0.75	0.60	0.77	0.75
2	25.1	0.71	0.62	0.83	0.70	0.78	0.87
1	23.1	0.63	0.52	0.78	0.60	0.77	0.79

Table A 7. HNSC dataset at family level in blood tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	133.3	0.87	0.86	0.88	0.88	0.96	0.90
9	124.5	0.89	0.89	0.88	0.89	0.95	0.90
8	116	0.87	0.86	0.88	0.87	0.95	0.89
7	108.8	0.87	0.86	0.88	0.87	0.96	0.90
6	98.3	0.89	0.89	0.88	0.89	0.96	0.90
5	87.9	0.87	0.87	0.88	0.88	0.96	0.89
4	79	0.87	0.86	0.88	0.87	0.96	0.89
3	68.7	0.87	0.86	0.88	0.87	0.96	0.89
2	52.4	0.87	0.85	0.89	0.87	0.94	0.90
1	33.4	0.87	0.84	0.90	0.87	0.92	0.91

Table A 8. HNSC dataset at order level in blood tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	193.6	0.89	0.84	0.95	0.89	0.95	0.94
9	185	0.90	0.86	0.94	0.89	0.94	0.94
8	176.3	0.90	0.85	0.95	0.89	0.94	0.95
7	168.1	0.89	0.84	0.95	0.89	0.94	0.95
6	158.6	0.90	0.86	0.95	0.90	0.94	0.95
5	148.3	0.89	0.85	0.93	0.89	0.94	0.93
4	135.7	0.89	0.85	0.94	0.89	0.93	0.94
3	120	0.89	0.84	0.94	0.88	0.94	0.94
2	106	0.88	0.84	0.92	0.87	0.94	0.92
1	75.5	0.87	0.83	0.92	0.87	0.94	0.92

Table A 9. HNSC dataset at genus level in blood tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	83.3	0.87	0.89	0.86	0.88	0.95	0.88
9	79.8	0.87	0.89	0.85	0.88	0.95	0.87
8	73.2	0.85	0.86	0.85	0.86	0.94	0.86
7	68.5	0.87	0.88	0.86	0.88	0.94	0.88
6	64.1	0.86	0.89	0.84	0.87	0.95	0.86
5	60.3	0.87	0.89	0.85	0.88	0.94	0.87
4	56.2	0.87	0.89	0.85	0.87	0.95	0.87
3	50.1	0.86	0.87	0.85	0.86	0.95	0.86
2	47	0.86	0.89	0.82	0.86	0.94	0.84
1	38.7	0.84	0.85	0.82	0.84	0.94	0.85

Table A 10. STAD dataset at family level in blood tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	100	0.82	0.81	0.82	0.81	0.90	0.84
9	94.5	0.82	0.83	0.81	0.82	0.90	0.84
8	87.4	0.80	0.77	0.83	0.79	0.88	0.84
7	81.7	0.81	0.80	0.82	0.81	0.87	0.83
6	76.7	0.79	0.77	0.82	0.79	0.87	0.84
5	70.2	0.80	0.76	0.84	0.78	0.86	0.84
4	63.6	0.81	0.77	0.86	0.80	0.86	0.86
3	58.1	0.78	0.72	0.83	0.76	0.85	0.83
2	50.8	0.75	0.69	0.81	0.72	0.84	0.79
1	32.5	0.76	0.68	0.84	0.72	0.83	0.83

Table A 11. STAD dataset at order level in blood tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	137.2	0.83	0.84	0.81	0.83	0.92	0.82
9	132.4	0.86	0.89	0.82	0.86	0.92	0.84
8	124.3	0.83	0.87	0.80	0.84	0.93	0.81
7	120.4	0.82	0.84	0.79	0.82	0.92	0.80
6	112.8	0.81	0.83	0.79	0.81	0.91	0.80
5	97	0.83	0.84	0.82	0.83	0.89	0.83
4	86.9	0.82	0.82	0.82	0.82	0.88	0.82
3	76.1	0.81	0.79	0.82	0.80	0.87	0.82
2	46	0.78	0.76	0.81	0.78	0.85	0.80
1	32.8	0.78	0.70	0.86	0.76	0.82	0.84

Table A 12. STAD dataset at genus level in blood tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	77.6	0.83	0.79	0.87	0.81	0.87	0.87
9	73.1	0.81	0.78	0.83	0.80	0.86	0.84
8	68.8	0.81	0.76	0.87	0.79	0.85	0.86
7	63.6	0.80	0.73	0.87	0.78	0.84	0.85
6	59.7	0.78	0.71	0.86	0.76	0.83	0.83
5	56.2	0.78	0.71	0.84	0.76	0.81	0.83
4	53.7	0.79	0.74	0.84	0.78	0.81	0.83
3	49.2	0.77	0.71	0.83	0.76	0.80	0.82
2	44.7	0.77	0.72	0.82	0.76	0.79	0.80
1	36.7	0.77	0.68	0.86	0.73	0.77	0.82

Table A 13. COAD dataset at family level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	308.2	0.84	0.96	0.55	0.90	0.91	0.86
9	286.7	0.87	0.96	0.65	0.91	0.92	0.89
8	263.1	0.87	0.98	0.60	0.92	0.92	0.87
7	251.9	0.86	0.98	0.55	0.91	0.90	0.86
6	218.9	0.86	0.96	0.60	0.90	0.90	0.87
5	202.2	0.86	0.96	0.60	0.91	0.93	0.88
4	180.1	0.87	0.96	0.65	0.91	0.92	0.89
3	150	0.87	0.96	0.65	0.92	0.91	0.88
2	98	0.86	0.94	0.65	0.90	0.90	0.88
1	47.9	0.90	0.98	0.70	0.94	0.83	0.90

Table A 14. COAD dataset at order level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	490.1	0.90	1.00	0.65	0.94	0.93	0.89
9	477.3	0.91	1.00	0.70	0.95	0.92	0.90
8	464.8	0.91	1.00	0.70	0.95	0.94	0.90
7	457.4	0.91	1.00	0.70	0.95	0.93	0.90
6	425.2	0.90	1.00	0.65	0.94	0.94	0.89
5	417.4	0.91	1.00	0.70	0.95	0.93	0.91
4	404.1	0.90	1.00	0.65	0.94	0.95	0.89
3	373.8	0.90	1.00	0.65	0.94	0.94	0.89
2	320.2	0.94	1.00	0.80	0.96	0.92	0.93
1	178.8	0.87	0.96	0.65	0.91	0.88	0.88

Table A 15. COAD dataset at genus level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	133	0.90	1.00	0.65	0.94	0.87	0.89
9	127.9	0.91	1.00	0.70	0.95	0.88	0.90
8	113.1	0.91	1.00	0.70	0.95	0.87	0.90
7	98.3	0.93	1.00	0.75	0.95	0.84	0.92
6	95.5	0.91	1.00	0.70	0.95	0.86	0.90
5	84.7	0.89	0.98	0.65	0.93	0.82	0.88
4	81.4	0.86	0.96	0.60	0.91	0.81	0.86
3	74.2	0.86	0.98	0.55	0.91	0.83	0.86
2	49.8	0.81	0.94	0.50	0.88	0.75	0.84
1	27.2	0.80	0.90	0.55	0.87	0.79	0.84

Table A 16. COAD-READ (CRC) dataset at family level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	292	0.80	0.98	0.50	0.86	0.88	0.77
9	264.5	0.80	0.98	0.50	0.86	0.80	0.77
8	251.7	0.80	0.96	0.53	0.86	0.84	0.78
7	230.6	0.80	0.96	0.53	0.86	0.86	0.79
6	217.9	0.80	0.98	0.50	0.86	0.83	0.77
5	191.7	0.81	0.96	0.57	0.86	0.85	0.79
4	159.3	0.79	0.92	0.57	0.84	0.84	0.79
3	136.4	0.78	0.92	0.53	0.84	0.84	0.77
2	96.6	0.80	0.96	0.53	0.86	0.81	0.78
1	51.2	0.71	0.90	0.40	0.79	0.80	0.72

Table A 17. COAD-READ (CRC) dataset at order level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	486.4	0.85	0.94	0.70	0.89	0.92	0.86
9	475.5	0.83	0.96	0.60	0.88	0.93	0.82
8	466	0.89	0.98	0.73	0.92	0.94	0.88
7	461.8	0.88	0.98	0.70	0.91	0.92	0.86
6	447.9	0.88	0.96	0.73	0.91	0.92	0.88
5	430.8	0.86	0.96	0.70	0.90	0.92	0.87
4	414.8	0.89	0.98	0.73	0.92	0.92	0.87
3	406	0.88	0.98	0.70	0.91	0.93	0.86
2	380.2	0.88	0.96	0.73	0.91	0.95	0.88
1	190.6	0.88	0.98	0.70	0.91	0.93	0.85

Table A 18. COAD-READ (CRC) dataset at genus level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	142.2	0.74	0.86	0.53	0.80	0.86	0.76
9	135.2	0.78	0.90	0.57	0.83	0.87	0.78
8	126.3	0.74	0.88	0.50	0.80	0.83	0.75
7	118.9	0.75	0.88	0.53	0.81	0.84	0.76
6	113.6	0.75	0.86	0.57	0.81	0.86	0.77
5	109.1	0.76	0.90	0.53	0.82	0.84	0.77
4	100.9	0.73	0.86	0.50	0.79	0.83	0.74
3	90.3	0.75	0.86	0.57	0.81	0.82	0.77
2	70.5	0.76	0.86	0.60	0.82	0.85	0.79
1	51.9	0.70	0.84	0.47	0.77	0.81	0.73

Table A 19. ESCA dataset at family level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	125.7	0.73	0.92	0.25	0.83	0.54	0.76
9	117.2	0.71	0.90	0.25	0.82	0.64	0.76
8	106.9	0.74	0.92	0.30	0.84	0.54	0.78
7	95.1	0.73	0.90	0.30	0.83	0.59	0.78
6	83.7	0.70	0.88	0.25	0.81	0.59	0.75
5	77.4	0.71	0.88	0.30	0.81	0.55	0.76
4	59.1	0.74	0.92	0.30	0.83	0.63	0.77
3	47.1	0.73	0.92	0.25	0.83	0.67	0.76
2	31.4	0.74	0.94	0.25	0.84	0.60	0.76
1	20.5	0.74	0.96	0.20	0.84	0.60	0.75

Table A 20. ESCA dataset at order level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	157.7	0.71	0.92	0.20	0.82	0.56	0.75
9	154.8	0.77	1.00	0.20	0.86	0.53	0.76
8	151.1	0.73	0.98	0.10	0.84	0.56	0.73
7	125.4	0.73	0.94	0.20	0.83	0.53	0.75
6	110.8	0.70	0.90	0.20	0.81	0.59	0.74
5	84	0.66	0.84	0.20	0.78	0.54	0.73
4	68.5	0.66	0.86	0.15	0.78	0.58	0.72
3	34.2	0.66	0.88	0.10	0.78	0.52	0.71
2	22	0.63	0.82	0.15	0.76	0.57	0.71
1	12.7	0.69	0.88	0.20	0.80	0.49	0.74

Table A 21. ESCA dataset at genus level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	81.3	0.66	0.88	0.10	0.78	0.54	0.71
9	75.4	0.67	0.88	0.15	0.79	0.55	0.72
8	67.4	0.66	0.88	0.10	0.78	0.54	0.71
7	60.4	0.66	0.88	0.10	0.78	0.48	0.71
6	46.7	0.61	0.82	0.10	0.74	0.51	0.69
5	42.1	0.64	0.86	0.10	0.77	0.53	0.71
4	28.7	0.64	0.88	0.05	0.77	0.49	0.70
3	26.1	0.64	0.88	0.05	0.77	0.48	0.70
2	23.4	0.66	0.90	0.05	0.79	0.47	0.70
1	8.9	0.67	0.92	0.05	0.80	0.47	0.71

Table A 22. HNSC dataset at family level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	293.3	0.96	1.00	0.85	0.97	0.99	0.95
9	261.5	0.93	0.98	0.80	0.95	0.97	0.93
8	247.8	0.91	0.98	0.75	0.95	0.96	0.92
7	227.8	0.94	1.00	0.80	0.96	0.98	0.93
6	211.2	0.94	1.00	0.80	0.96	0.96	0.93
5	194.1	0.94	0.98	0.85	0.96	0.96	0.95
4	158.4	0.94	0.98	0.85	0.96	0.98	0.95
3	107.3	0.91	0.96	0.80	0.94	0.95	0.93
2	72.9	0.90	0.94	0.80	0.93	0.91	0.93
1	39	0.77	0.86	0.55	0.83	0.76	0.84

Table A 23. HNSC dataset at order level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	403.1	0.87	0.98	0.60	0.92	0.94	0.87
9	391.6	0.94	1.00	0.80	0.96	0.95	0.93
8	371.7	0.94	1.00	0.80	0.97	0.96	0.94
7	338.3	0.89	0.98	0.65	0.93	0.96	0.89
6	315.1	0.93	0.98	0.80	0.95	0.98	0.94
5	281.7	0.94	1.00	0.80	0.97	0.97	0.94
4	238.7	0.93	0.98	0.80	0.95	0.93	0.94
3	195.9	0.91	1.00	0.70	0.95	0.93	0.91
2	156.7	0.91	0.98	0.75	0.94	0.94	0.92
1	96.5	0.86	0.94	0.65	0.91	0.87	0.89

Table A 24. HNSC dataset at genus level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	168.3	0.91	0.98	0.75	0.95	0.92	0.92
9	160.7	0.90	0.98	0.70	0.94	0.92	0.90
8	146.9	0.87	0.96	0.65	0.92	0.92	0.89
7	139.1	0.89	0.98	0.65	0.93	0.93	0.89
6	127.3	0.90	1.00	0.65	0.94	0.94	0.89
5	122.3	0.89	0.98	0.65	0.93	0.90	0.89
4	90.9	0.90	0.98	0.70	0.94	0.92	0.90
3	80.7	0.84	0.96	0.55	0.90	0.93	0.85
2	54.8	0.84	0.98	0.50	0.90	0.86	0.85
1	37.8	0.83	0.96	0.50	0.89	0.83	0.85

Table A 25. STAD dataset at family level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	99.6	0.73	0.90	0.38	0.81	0.71	0.75
9	95.2	0.72	0.89	0.38	0.81	0.76	0.75
8	80.5	0.68	0.85	0.35	0.78	0.70	0.73
7	69.7	0.70	0.88	0.35	0.79	0.71	0.74
6	60.6	0.71	0.89	0.35	0.80	0.74	0.74
5	43.7	0.73	0.91	0.38	0.82	0.71	0.76
4	37.3	0.73	0.94	0.33	0.82	0.71	0.74
3	26.7	0.70	0.91	0.28	0.80	0.66	0.72
2	9.9	0.71	0.95	0.23	0.81	0.64	0.72
1	4.7	0.70	0.98	0.15	0.81	0.65	0.70

Table A 26. STAD dataset at order level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	91.7	0.67	0.83	0.35	0.77	0.72	0.72
9	88.5	0.70	0.85	0.40	0.79	0.74	0.75
8	81.4	0.69	0.86	0.35	0.79	0.73	0.73
7	74.1	0.72	0.86	0.43	0.80	0.76	0.75
6	61.3	0.70	0.86	0.38	0.79	0.73	0.73
5	45.7	0.73	0.86	0.45	0.80	0.77	0.76
4	32.1	0.75	0.86	0.53	0.82	0.76	0.80
3	25.9	0.73	0.91	0.38	0.82	0.74	0.75
2	20.2	0.74	0.94	0.35	0.83	0.73	0.75
1	5.6	0.73	0.98	0.25	0.83	0.66	0.73

Table A 27. STAD dataset at genus level in solid normal tissue as negative control from MicrobiomeGSM

#Clusters	#Species (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Precision (Mean)
10	94.8	0.71	0.88	0.38	0.80	0.65	0.74
9	81.2	0.70	0.88	0.35	0.80	0.67	0.74
8	71.4	0.69	0.88	0.33	0.79	0.68	0.74
7	69	0.70	0.88	0.35	0.80	0.67	0.74
6	53	0.70	0.89	0.33	0.80	0.65	0.74
5	49.3	0.68	0.88	0.28	0.78	0.60	0.72
4	45.2	0.70	0.90	0.30	0.80	0.61	0.72
3	35.3	0.71	0.91	0.30	0.81	0.60	0.73
2	25.5	0.70	0.93	0.25	0.80	0.58	0.71
1	13	0.69	0.95	0.18	0.80	0.60	0.70

Table A 28. Performance of traditional feature selection approach on COAD dataset at species level in normal solid sample with 684 features

FS Model	ML Model	Accuracy	SEN	SPE	F1	Precision	AUC
XGB	Adaboost	0.83	0.89	0.68	0.88	0.89	0.84
SKB	RF	0.89	0.99	0.66	0.93	0.89	0.89
MRMR	SVM_opt	0.86	0.99	0.53	0.91	0.85	0.85
IG	Adaboost	0.82	0.90	0.62	0.88	0.87	0.79
FCBF	SVM_opt	0.9	1.0	0.7	0.9	0.9	0.9
CMIM	SVM_opt	0.84	1.00	0.44	0.90	0.83	0.86

Table A 29. Performance of traditional feature selection approach on COAD dataset at species level in normal solid sample with 16 union features

FS Model	ML Model	Accuracy	SEN	SPE	F1	Precision	AUC
XGB	Stack_SVM_KNN	0.79	0.88	0.55	0.85	0.84	0.77
SKB	SVM_opt	0.88	0.92	0.76	0.91	0.92	0.90
MRMR	SVM_opt	0.9	0.9	0.8	0.9	0.9	0.9
IG	Stack_Logitboost_KNN	0.86	0.88	0.81	0.90	0.93	0.90
FCBF	SVM_opt	0.9	0.9	0.9	0.9	1.0	1.0
CMIM	SVM_opt	0.8	0.9	0.8	0.9	0.9	0.9

Table A 30. Performance of traditional feature selection approach on COAD-READ dataset at species level in normal solid sample with 685 union features

FS Model	ML Model	Accuracy	SEN	SPE	F1	Precision	AUC
XGB	Stack_Logitboost_KNN	0.75	0.82	0.63	0.80	0.82	0.77
SKB	XGBoost	0.78	0.88	0.60	0.83	0.79	0.80
MRMR	SVM_opt	0.84	1.00	0.57	0.89	0.82	0.81
IG	XGBoost	0.73	0.80	0.60	0.77	0.79	0.84
FCBF	SVM_opt	0.86	0.92	0.77	0.89	0.88	0.85
CMIM	Stack_Logitboost_KNN	0.73	0.76	0.67	0.78	0.82	0.78

Table A 31. Performance of traditional feature selection approach on COAD-READ dataset at species level in normal solid sample with 17 union features

FS Model	ML Model	Accuracy	SEN	SPE	F1	Precision	AUC
XGB	SVM_opt	0.86	0.94	0.73	0.90	0.86	0.89
SKB	XGBoost	0.90	0.92	0.87	0.92	0.93	0.92
MRMR	XGBoost	0.69	0.74	0.60	0.74	0.79	0.78
IG	Stack_Logitboost_KNN	0.89	0.98	0.73	0.92	0.88	0.85
FCBF	Adaboost	0.78	0.86	0.63	0.83	0.82	0.75
CMIM	XGBoost	0.75	0.80	0.67	0.79	0.81	0.86

Table A 32. Performance of traditional feature selection approach on ESCA dataset at species level in normal solid sample with 314 features

FS Model	ML Model	Accuracy	SEN	SPE	F1	Precision	AUC
CMIM	Adaboost	0.61	0.72	0.33	0.72	0.73	0.51
FCBF	Stack_Logitboost_KNN	0.6	0.6	0.6	0.6	0.8	0.7
IG	DT	0.68	0.76	0.47	0.76	0.79	0.65
MRMR	Stack_Logitboost_KNN	0.59	0.65	0.46	0.69	0.75	0.57
SKB	DT	0.7	0.77	0.525	0.77	0.81	0.68
XGB	DT	0.72	0.83	0.45	0.80	0.79	0.67

Table A 33. Performance of traditional feature selection approach on ESCA dataset at species level in normal solid sample with 16 union features

FS Model	ML Model	Accuracy	SEN	SPE	F1	Precision	AUC
CMIM	LogitBoost	0.65	0.64	0.66	0.71	0.84	0.71
FCBF	Adaboost	0.74	0.8	0.6	0.81	0.85	0.70
IG	RF	0.82	0.92	0.55	0.88	0.85	0.86
MRMR	LogitBoost	0.68	0.75	0.51	0.76	0.80	0.69
SKB	Adaboost	0.82	0.87	0.705	0.87	0.89	0.86
XGB	Adaboost	0.66	0.75	0.45	0.75	0.78	0.66

Table A 34. Performance of traditional feature selection approach on STAD dataset at species level in normal solid sample with 320 feature

FS Model	ML Model	Accuracy	SEN	SPE	F1	Precision	AUC
CMIM	Adaboost	0.67	0.80	0.40	0.76	0.73	0.55
FCBF	Adaboost	0.58	0.68	0.4	0.68	0.69	0.59
IG	DT	0.69	0.81	0.47	0.77	0.76	0.64
SKB	XGBoost	0.74	0.88	0.46	0.82	0.78	0.78
XGB	Stack_Logitboost _KNN	0.67	0.74	0.52	0.74	0.76	0.68
MRMR	Stack_Logitboost _KNN	0.60	0.71	0.38	0.69	0.72	0.60

Table A 35. Performance of traditional feature selection approach on STAD dataset at species level in normal solid sample with 12 features

FS Model	ML Model	Accuracy	SEN	SPE	F1	Precision	AUC
CMIM	LogitBoost	0.65	0.76	0.41	0.74	0.73	0.60
FCBF	Adaboost	0.63	0.64	0.63	0.70	0.78	0.61
IG	RF	0.73	0.78	0.62	0.79	0.81	0.83
SKB	RF	0.74	0.81	0.61	0.80	0.82	0.83
XGB	XGBoost	0.8	0.92	0.56	0.86	0.82	0.84
MRMR	SVM_opt	0.73	0.79	0.61	0.79	0.81	0.71

Table A 36. Performance of traditional feature selection approach on HNSC dataset at species level in normal solid sample with 697 features

FS Model	ML Model	Accuracy	SEN	SPE	F1	Precision	AUC
CMIM	SVM_opt	0.84	1.00	0.47	0.90	0.83	0.96
IG	Adaboost	0.87	0.93	0.71	0.91	0.90	0.87
SKB	Adaboost	0.82	0.89	0.65	0.87	0.87	0.82
FCBF	DT	0.74	0.92	0.30	0.84	0.78	0.61
XGB	SVM_opt	0.86	0.91	0.73	0.90	0.91	0.87
MRMR	SVM_opt	0.85	1.00	0.47	0.91	0.84	0.94

Table A 37. Performance of traditional feature selection approach on HNSC dataset at species level in normal solid sample with 19 features

FS Model	ML Model	Accuracy	SEN	SPE	F1	Precision	AUC
CMIM	SVM_opt	0.91	0.96	0.80	0.94	0.94	0.94
IG	LogitBoost	0.90	0.94	0.80	0.93	0.93	0.92
SKB	SVM_opt	0.89	0.92	0.80	0.92	0.93	0.92
FCBF	SVM_opt	0.87	0.90	0.80	0.91	0.93	0.95
XGB	SVM_opt	0.91	0.98	0.75	0.94	0.92	0.81
MRMR	SVM_opt	0.90	0.96	0.75	0.93	0.91	0.94

Table A 38. Species which have high importance value for solid-primary tumor sample from feature selection approach

STAD	ESCA	HNSC	COAD	COAD-READ
Helicobacter pylori	Helicobacter pylori	Streptococcus mutans	Odoribacter splanchnicus	Bacteroides ovatus
Helicobacter acinonychis	Helicobacter acinonychis	Eikenella sp. HMSC061C02	Porphyromonas asaccharolytica	Coprococcus eutactus
Helicobacter cetorum	Prevotella loescheii	Neisseria sp. HMSC064F03	Gemella morbillorum	Faecalitalea cylindroides
Lactobacillus vaginalis	Prevotella histicola	Rothia sp. HMSC061D12	[Eubacterium] rectale	Bacteroides coprophilus
Streptococcus infantis	Prevotella sp. oral taxon 317	Tannerella sp. oral taxon HOT-286	Bacteroides massiliensis	Bacteroides sp. 4_1_36
Aggregatibacter aphrophilus	Streptococcus sp. oral taxon 431	Actinomyces bouchesdurhonensis	Bacteroides finegoldii	Erysipelotrichaceae bacterium 2_2_44A
Prevotella pallens	Streptococcus sp. HSISS3	Prevotella veroralis	Clostridium sp. M62/1	Lachnospiraceae bacterium 3_1_57FAA_CT1
Prevotella enoeca	Streptococcus sp. HMSC056C01	Prevotella enoeca	Bacteroides sp. 4_1_36	Clostridiales bacterium VE202-06
Solobacterium moorei	Gemella haemolysans	Prevotella pleuritidis	Clostridium sp. ATCC BAA-442	Clostridiales bacterium VE202-21
Campylobacter sp. 10_1_50	Prevotella corporis	Veillonella sp. 3_1_44	Subdoligranulum sp. 4_3_54A2FAA	Bacteroides mediterraneensis
Streptococcus sp. HMSC077D04	Alloprevotella tannerae	Prevotella fusca	Ruminococcus sp. JC304	Holdemania sp. Marseille-P2844
Haemophilus sp. HMSC073C03	Bulleidia extructa	Alloprevotella rava	Clostridiales bacterium VE202-21	Campylobacter ureolyticus
	Prevotella salivae	Streptococcus sp. oral taxon 071	Bacteroides sp. UW	Parabacteroides merdae
	Prevotella sp. oral taxon 306	Selenomonas sp. CM52	Clostridium sp. FS41	Bacteroides massiliensis
	Prevotella sp. ICM33	Actinomyces sp. ICM58	Fournierella massiliensis	Bacteroides finegoldii
	Streptococcus sp. 263_SSPC	Actinomyces sp. ph3	Bacteroides mediterraneensis	Clostridium sp. FS41
		Haemophilus sp. HMSC073C03		
		Streptococcus sp. HMSC062D07		
		Prevotella phocaeensis		

Table A 39. Performance result of traditional feature selection method on COAD dataset at species level in blood tissue sample with 682 features

FS Model	ML Model	Accuracy	Sensitivity	Specifity	F-measure	Precision	AUC
XGB	RF	0.94	0.91	0.96	0.93	0.96	0.96
SKB	RF	0.93	0.86	0.99	0.92	0.99	0.96
MRMR	Stack_Logitboost_KNN	0.66	0.64	0.67	0.65	0.74	0.78
IG	RF	0.95	0.91	0.98	0.94	0.98	0.98
FCBF	SVM_opt	0.76	0.56	0.95	0.67	0.89	0.91
CMIM	Adaboost	0.8	0.75	0.93	0.82	0.92	0.92

Table A 40. Performance result of traditional feature selection method on COAD dataset at species level in blood tissue sample with 23 features

FS Model	ML Model	Accuracy	Sensitivity	Specifity	F-measure	Precision	AUC
XGB	Stack_SVM_KNN	0.91	0.83	0.99	0.90	0.99	0.96
SKB	Adaboost	0.94	0.93	0.95	0.94	0.95	0.98
MRMR	RF	0.895	0.82	0.97	0.89	0.97	0.94
IG	XGBoost	0.96	0.97	0.94	0.96	0.95	0.98
FCBF	LogitBoost	0.91	0.85	0.96	0.90	0.96	0.97
CMIM	XGBoost	0.92	0.94	0.9	0.92	0.91	0.96

Table A 41. Performance result of traditional feature selection method on COAD-READ dataset at species level in blood tissue sample with 683 features

FS Model	ML Model	Accuracy	Sensitivity	Specifity	F-measure	Precision	AUC
XGB	RF	0.93	0.91	0.96	0.93	0.96	0.96
SKB	RF	0.92	0.87	0.97	0.91	0.97	0.95
MRMR	Stack_Logitboost_KNN	0.72	0.54	0.91	0.65	0.88	0.83
IG	RF	0.96	0.96	0.96	0.96	0.96	0.97
FCBF	RF	0.77	0.58	0.97	0.70	0.96	0.94
CMIM	Stack_Logitboost_Kmenas	0.75	0.63	0.88	0.72	0.88	0.84

Table A 42. Performance result of traditional feature selection method on COAD-READ dataset at species level in blood tissue sample with 31 features

FS Model	ML Model	Accuracy	Sensitivity	Specifity	F-measure	Precision	AUC
CMIM	RF	0.95	0.95	0.95	0.95	0.96	0.98
XGB	RF	0.91	0.89	0.93	0.91	0.93	0.95
SKB	RF	0.95	0.92	0.98	0.95	0.98	0.97
MRMR	RF	0.90	0.82	0.99	0.89	0.98	0.96
IG	RF	0.94	0.93	0.96	0.94	0.96	0.96
FCBF	Adaboost	0.90	0.89	0.91	0.90	0.93	0.95

Table A 43. Performance result of traditional feature selection method on ESCA dataset at species level in blood tissue sample with 312 features

FS Model	ML Model	Accuracy	Sensitivity	Specifity	F-measure	Precision	AUC
XGB	Adaboost	0.76	0.78	0.725	0.78	0.81	0.79
SKB	SVM_opt	0.87	0.78	0.98	0.85	0.98	0.96
MRMR	SVM_opt	0.73	0.66	0.83	0.78	0.76	0.84
IG	Adaboost	0.86	0.86	0.85	0.86	0.89	0.92
FCBF	SVM_opt	0.71	0.66	0.78	0.71	0.80	0.85
CMIM	SVM_opt	0.7	0.58	0.85	0.66	0.87	0.89

Table A 44. Performance result of traditional feature selection method on ESCA dataset at species level in blood tissue sample with 18 features

FS Model	ML Model	Accuracy	Sensitivity	Specifity	F-measure	Precision	AUC
XGB	Adaboost	0.76	0.66	0.88	0.74	0.90	0.74
SKB	Stack_SVM_KNN	0.86	0.82	0.90	0.85	0.93	0.80
MRMR	DT	0.71	0.64	0.80	0.70	0.82	0.64
IG	Adaboost	0.82	0.84	0.80	0.84	0.86	0.82
FCBF	LogitBoost	0.8	0.8	0.8	0.81	0.84	0.83
CMIM	Stack_Logitboost_KNN	0.78	0.76	0.80	0.78	0.85	0.84

Table A 45. Performance result of traditional feature selection method on HNSC dataset at species level in blood tissue sample with 698 features

FS Model	ML Model	Accuracy	Sensitivity	Specifity	F-measure	Precision	AUC
XGB	XGBoost	0.90	0.93	0.86	0.90	0.88	0.95
SKB	XGBoost	0.87	0.84	0.90	0.86	0.90	0.93
MRMR	Adaboost	0.68	0.41	0.97	0.55	0.94	0.68
IG	LogitBoost	0.91	0.90	0.93	0.92	0.93	0.96
FCBF	Stack_Logitboost_KNN	0.59	0.31	0.88	0.43	0.75	0.67
CMIM	Stack_Logitboost_KNN	0.59	0.29	0.91	0.40	0.76	0.66

Table A 46. Performance result of traditional feature selection method on HNSC dataset at species level in blood tissue sample with 22 features

FS Model	ML Model	Accuracy	Sensitivity	Specifity	F-measure	Precision	AUC
XGB	XGBoost	0.89	0.86	0.92	0.88	0.92	0.92
SKB	RF	0.91	0.92	0.91	0.92	0.92	0.98
MRMR	Adaboost	0.90	0.89	0.92	0.90	0.93	0.94
IG	XGBoost	0.90	0.91	0.90	0.91	0.91	0.93
FCBF	LogitBoost	0.91	0.91	0.91	0.91	0.92	0.94
CMIM	XGBoost	0.90	0.90	0.91	0.91	0.92	0.96

Table A 47. Performance result of traditional feature selection method on STAD dataset at species level in blood tissue sample with 216 features

FS Model	ML Model	Accuracy	Sensitivity	Specifity	F-measure	Precision	AUC
XGB	RF	0.88	0.87	0.90	0.88	0.91	0.94
SKB	Adaboost	0.84	0.87	0.82	0.84	0.84	0.91
MRMR	RF	0.67	0.42	0.91	0.55	0.82	0.82
IG	LogitBoost	0.89	0.83	0.94	0.88	0.94	0.91
FCBF	DT	0.72	0.70	0.74	0.71	0.75	0.69
CMIM	LogitBoost	0.72	0.52	0.92	0.64	0.86	0.77

Table A 48. Performance result of traditional feature selection method on STAD dataset at species level in blood tissue sample with 22 features

FS Model	ML Model	Accuracy	Sensitivity	Specifity	F-measure	Precision	AUC
XGB	Adaboost	0.84	0.80	0.88	0.83	0.86	0.86
SKB	RF	0.91	0.90	0.91	0.91	0.91	0.96
MRMR	Adaboost	0.84	0.84	0.84	0.84	0.85	0.89
IG	Adaboost	0.83	0.80	0.87	0.83	0.87	0.86
FCBF	RF	0.84	0.82	0.86	0.84	0.86	0.92
CMIM	RF	0.81	0.79	0.82	0.79	0.83	0.88

Curriculum Vitae

EDUCATION

2022 - 2025 : Abdullah Gül University, Department of Bioengineering, Master of Science, TURKEY

2017 - 2022 : Abdullah Gül University, Department of Molecular Biology and Genetics, Bachelor of Science, TURKEY

EXPERIENCE

09/2024- 12/2024 : Barcelona Supercomputing Center, SPAIN**-** Computational Biology Intern

HONOURS & GRANTS

2023-2024: TUBITAK 2210-C Turkey Priority Areas Master's Scholarship

2021-2022: STAR-Intern Researcher Scholarship

2021: TEKNOFEST - International Biotechnology and Innovation Competition-Finalist

2020: TEKNOFEST - International Biotechnology and Innovation Competition-Finalist