IntegrIBS: Towards Building a Robust IBS Classifier with Integrated Microbiome Data

Dinakaran Palani[1], Nishad Bapatdhar[1], Renuka Manchiraju[1], Bipin Pradeep Kumar[1], Samik Ghosh[1], Sucheendra K. Palaniappan[1,2*]

[1]Iom Bioworks Private Limited, Bangalore, Karnataka, India.
[2]The Systems Biology Institute, Tokyo, Japan.

*Corresponding author(s). E-mail(s): sucheendra@sbi.jp;
Contributing authors: dinakaran@iombio.com; nishad@iombio.com; renuka@iombio.com; bipin@iombio.com; ghosh@iombio.com;

Abstract

Diagnosing Irritable Bowel Syndrome (IBS) is challenging due to its complex nature and overlapping symptoms with other conditions. In this study, we focus on developing a robust machine learning classifier to predict IBS conditions using microbiome data. By integrating multiple publicly available datasets, we addressed dataset-specific nuances and captured a broad spectrum of microbiome variations across different geographies and demographics. We evaluated several machine learning algorithms and observed that XGBoost model performed most robustly, achieving a mean accuracy of 0.75 with a standard deviation of 0.04 using 10-fold cross-validation. In contrast, models trained on individual datasets displayed significant variability and lower precision when tested on other datasets. Explainability analysis of the model identified 19 key bacterial taxa influencing predictions, which aligned with existing literature, supporting the model's biological relevance. This study demonstrated the effectiveness of dataset integration in creating reliable IBS diagnostic tools.

Keywords: Microbiome, IBS, Machine learning, Dataset

1 Introduction

Irritable Bowel Syndrome (IBS) is a common gastrointestinal disorder which is often characterized by chronic abdominal pain, flatulence, bloating, and altered bowel habits. It is a complex and heterogeneous condition, with overlapping symptoms with other gastrointestinal disorders and its pathophysiology is not fully understood. Due to this, diagnosing IBS is difficult. In fact, very often, diagnosing IBS is done by the principal of exclusion. Traditional diagnostic methods often miss the intricate interactions within the gut microbiome that have been increasingly understood to contribute to IBS [1]. Collins [2] showed the influence of gut microbiota on gastrointestinal function and its potential as a therapeutic target for IBS. Andrews et al. [3] stressed the importance of managing the microbiome in clinical practice for IBS patients, suggesting personalized treatment strategies based on microbiome profiling.

An ideal diagnostic tool for IBS based on microbiome profiling would involve asking a patient with symptoms resembling IBS to undergo a microbiome profile of the gut through a stool test. Once the microbial abundances are known, an algorithm would analyze these abundances to ascertain if the person has IBS and help the clinician start the treatment regimen. To this end, we focus on developing a machine learning classifier to robustly predict IBS using microbiome data. To add, the microbiome composition can vary significantly due to demographic, environmental, and societal factors. To address this variability, it is crucial to integrate multiple datasets from diverse conditions to identify the core set of causative microbes. Thus, we integrated multiple publicly available datasets to enhance model robustness. Ford et al. [4] discussed the need for comprehensive datasets to better understand IBS, which aligns with our approach.

Previous studies using microbiome data for IBS diagnosis have shown potential. Camilleri [5] reviewed diagnostic and treatment strategies for IBS, noting the potential of microbiome-based approaches. Wang and Liu [6] compared different classifiers for human microbiome data, showing the effectiveness of machine learning techniques. Fukui et al. [7] demonstrated the utility of machine learning-based gut microbiome analysis in identifying IBS patients. Tanaka et al. [8] also used a machine-learning approach to diagnose IBS subtypes, showing the utility of advanced algorithms. However, these studies often relied on single datasets, limiting their generalizability.

Our work aims to integrate diverse datasets aimed to capture a wide range of microbiome variations, in addition to performing dataset-specific processing steps which take in account nuances of data collection protocols of each study. This context specific processing and focus on being able to unify data cleanly would eventually help in enhancing the predictive accuracy and generalizability of our IBS classifier. This aligns with Carcy et al. [9], who conducted a large-scale meta-analysis of IBS, highlighting the importance of comprehensive data integration. Additionally, Bonetto [10] reviewed recent advances in IBS treatment, emphasizing the need for innovative diagnostic methods that leverage large-scale data. In summary, our study combines multiple datasets towards developing a more generalizable IBS diagnostic tool.

The paper is organized as follows: Section 2 describes the datasets and preprocessing methods. Section 3 details the bioinformatics analyses and dataset-specific processing, outlining the pipeline and framework for building the IBS classifier. Section

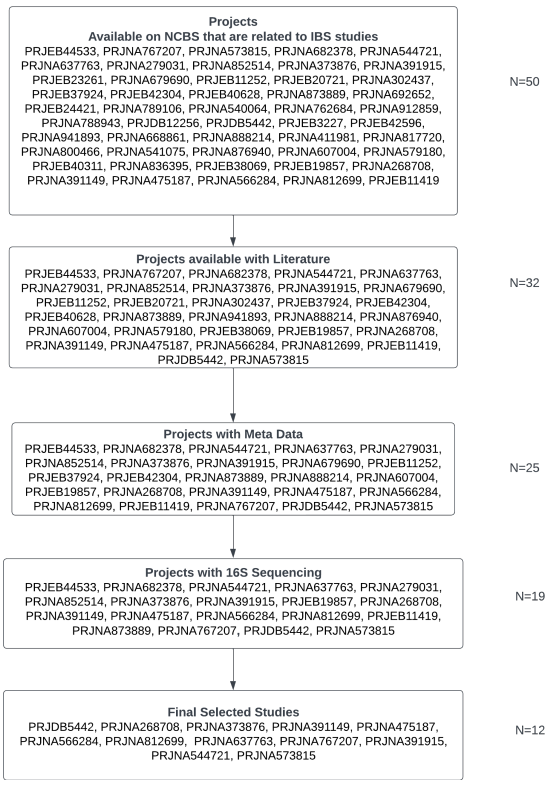

Fig. 1 Selection Criteria and Process for IBS Project Dataset Inclusion

4 presents the results, demonstrating the superiority of combining multiple datasets as we have done. We also provide, in Section 5, an explainable AI framework to understand the model's predictions beyond just a black box, which can further help in understanding the reasons for the model's predictions. Finally, Section 6 concludes the paper and suggests future research directions.

2 Data curation and pre-processing

A thorough search of the National Center for Biotechnology Information (NCBI) databases was conducted to identify studies related to Irritable Bowel Syndrome (IBS) and the gut microbiome. Most of the identified studies were downloaded; however, many lacked essential information such as primer details and metadata necessary to distinguish between IBS and healthy samples.

Although the uploaded data should ideally be raw data, it is often altered by various sequence trimming and merging techniques before submission. Many studies have not documented the alterations or trimming and merging techniques applied to their sequencer output. As a result, each study must be evaluated individually based on the quality of its data.

For this study, we began with 50 studies obtained from the NCBI website. Among these, 18 studies lacked associated literature but contained sequences and metadata, which were included in the analysis. Subsequently, of the remaining 32 studies, only 25 had metadata capable of distinguishing between IBS and control samples. Further refinement based on primer information reduced the selection to 19 relevant studies, after excluding those related to Internal Transcribed Spacer (ITS) and Whole Metagenome Shotgun Sequencing (WMGS). This process underscores the significance of documenting metadata and primer information in clinical trial research. The process is described in Fig. 1

Our analysis focused on gut microbiome data from the filtered 19 independent studies worldwide. Further, four studies were excluded due to challenges with $FastQC$ and $DADA2$ resulting in insufficient reads, boiling down in a final selection of 12 studies. At the author's request, one study data was made available[11]. Finally, these studies encompassed 1002 samples from individuals diagnosed with Irritable Bowel Syndrome (IBS) and 450 healthy control samples, with sample sizes per study ranging from 10 to 300. All datasets were downloaded from the European Nucleotide Archive (ENA)[12], and accompanying metadata were retrieved from the BioSamples database[13]. Table 1 summarizes the selected studies for this, listing their PubMed identifier (PMID), number of IBS and healthy control (HC) samples, data provenance, targeted regions, and the country of origin. This provides a clear snapshot of the dataset diversity and methodological details.

Table 1 Table of Final Studies Picked for Data Integration: Metadata and Provenance Details

PMID	# IBS	# HC	Data Provenance	Regions	Country
28628918[14]	10	10	PRJDB5442	V1-V2	Japan
26239401[15]	202	88	PRJNA268708	V4	Spain
28457228[16]	29	23	PRJNA373876	V3-V5	USA
33925672[17]	36	41	PRJNA391149	V1-V3	Italy
30090090[18]	17	9	PRJNA475187	V3-V4	China
31681624[19]	15	14	PRJNA566284	V4	China
36624530[20]	305	175	PRJNA812699	V4	USA
22180058[11]	37	20	PMID_22180058	V4	Sweden
32727141[7]	85	26	PRJNA637763	V1-V2	Japan
35704960[21]	60	-	PRJNA767207	V3-V4	Russia
29708822[22]	68	-	PRJNA391915	V6-V7	USA
32552668[23]	84	44	PRJNA544721	V3-V4	China
33168839[24]	54	-	PRJNA573815	V3–V4	China

2.1 Power of Combined Dataset

Despite having identical clinical objectives, individual studies can yield contradicting results regarding the same bacteria due to smaller sample sizes. This variability underscores the need for a larger, more diverse dataset to identify robust correlations between the microbiome and disease conditions. Additionally, gut microbiome diversity varies significantly by geographic location, suggesting that combining datasets

from different regions can capture a broader spectrum of variations. Previous research corroborates that microbial associations differ across populations, affecting the robustness and generalizability of microbiome studies [25, 26]. Our results also show that classifiers trained on single-study data perform poorly when tested on other datasets, highlighting the limitations of models developed from isolated datasets.

By integrating microbial profiles from multiple studies, we aim to leverage the increased sample size and diversity to develop a more robust microbial signature. This approach is expected to mitigate the limitations of small sample sizes and study-specific biases, enhancing the classifier's ability to generalize across different populations. We hypothesize that integrating multiple datasets will enhance the predictive accuracy and robustness of our IBS classifier. To test this hypothesis, we compare the performance of machine learning classifiers trained on the combined dataset against those trained on individual datasets. We show that models trained on the combined dataset exhibit higher accuracy.

3 Data specific analysis pipeline

This section outlines the preprocessing and analysis pipeline used to address the dataset-specific nuances of 16S rRNA sequencing data from the studies we identified in the previous section. Our goal is to integrate diverse datasets with the goal of improved predictability of IBS vs healthy. Key steps in the pipeline include sequencing quality control, read trimming, inference of amplicon sequence variants (ASVs), training the classifier based on primer sequences, and taxonomical classification. Each step is structured to standardize and optimize the data for reliable downstream analysis.

- **Sequencing Quality Control:** Quality control of the raw sequencing reads was performed using FastQC and MultiQC. FastQC provides comprehensive statistics and visualizations to assess the quality of the sequencing data, while MultiQC aggregates results from multiple tools to provide a summary report. Raw FASTQ files were analyzed for quality scores across all bases, GC content, sequence duplication levels, and the presence of adapter sequences. Datasets may be discarded due to incomplete information causing failures in quality control analyses. The output reports were manually inspected to ensure the data quality was sufficient for downstream analysis.
- **Trimming of Reads:** Adapter trimming and quality filtering of the raw reads were conducted using Cutadapt (version 3.4) [27]. Adapter sequences were identified and specified for removal. Reads were trimmed to remove low-quality bases, ensuring a minimum quality . Post trimming, the quality of the reads was reassessed using FastQC to confirm the removal of adapters and improvement in read quality.
- **Infer Amplicon Sequence Variants (ASVs):** Amplicon sequence variants (ASVs) were inferred using the $DADA2$ pipeline [28]. The workflow involved several steps:
  - Filtering and Trimming: Reads were filtered to remove low-quality regions and trimmed to a uniform length to reduce error rates.
  - Error Learning: Error rates were estimated from the data itself.

- Denoising: Sequences were denoised to infer true biological sequences, correcting for sequencing errors.
- Chimera Removal: Chimeric sequences were identified and removed.

- **Training a Classifier Based on Primer Sequence for sequence classification of bacteria:** A classifier was trained using a reference database tailored to the specific primer sequences used in this study. The reference sequences corresponding to the target region were extracted and formatted to train the classifier. This step ensured accurate taxonomic classification by aligning the reference database with the amplified region defined by the primers.
- **Taxonomical Classification:** Taxonomical classification of the inferred ASVs was performed using QIIME2's feature classifier plugin method with the trained reference database. The classifier was trained on the SILVA database[29], with reference reads extracted based on the primer sequences used in the studies.

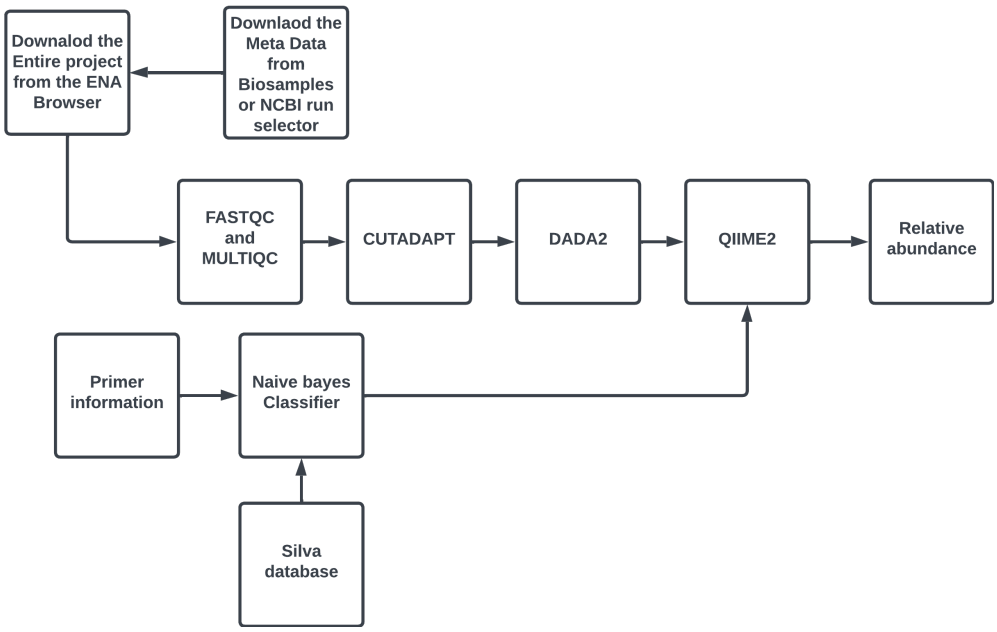

Fig. 2 Block Diagram of the Bioinformatics Pipeline for 16S rRNA Sequencing Data Processing

3.1 Pipeline Implementation

The curated datasets were processed using the nf-core/ampliseq pipeline, facilitated by the NF Tower platform. The nf-core/ampliseq pipeline is a robust and standardized workflow for processing amplicon sequencing data, particularly 16S rRNA gene sequences. The Azure Blob Storage and Azure Batch services were integrated to run

the nf-core/ampliseq pipeline efficiently. Each study thoroughly inspected the generated MultiQC reports, and appropriate quality cutoffs were determined to minimize data loss.

3.2 Trimming Strategy Considerations

Trimming involves the removal of low-quality bases from sequencing reads, which can significantly impact the accuracy of downstream analyses such as alignment and assembly. Two primary approaches to trimming include cutoff lengths and quality-based trimming that are used by the DADA2 pipeline

- **Cutoff lengths (truncLen)**:
  - *truncLen* specifies the exact read length at which reads should be truncated. This is useful when there is a specific target read length or consistent read length. However, it may lead to the loss of valuable information if read lengths vary significantly or if there are substantial quality drops toward the end of the reads.
- **Quality-based trimming** ($trunc_{qmin}$, $trunc_{qmax}$, max_{ee}, min_{len}):
  - $trunc_{qmin}$ and $trunc_q max$ set the minimum and maximum acceptable quality scores for read truncation, respectively.
  - max_{ee} sets the maximum number of expected errors allowed in a truncated read.
  - min_{len} sets the minimum length a read must retain after truncation to be retained for further analysis. This approach allows for more flexible trimming based on quality scores, potentially retaining more data if quality allows.

3.3 Microbiome Dataset specific processing

The datasets were processed according to their unique characteristics. Below is a summary of each study and the specific pre-processing steps applied.

- **PRJNA573815**: This clinical study investigates the use of laxatives and probiotics for the treatment of Irritable Bowel Syndrome (IBS). The microbiome data is accessible via the ENA browser. Samples with the sample alias and baseline samples were selected for analysis. Primers were already removed, and merged reads were uploaded to the ENA browser. Multiple analyses were conducted with different cutoff lengths, starting from 445 and decreasing in increments of 5. A cutoff length of 425 was selected as it allowed a higher number of reads to pass through the DADA2 pipeline.
- **PRJNA544721**: This study aimed to identify the IBS phenotypes and healthy controls. Using the metatag isolation source, we identified samples of IBS-D and healthy controls. The primers were removed, and multiple analyses were conducted with cutoff lengths starting from 450 to 420. A cutoff length of 430 was selected because it allowed more reads to pass through.
- **PRJNA391915**: This study focuses on the effects of rifaximin on IBS-D patients. Baseline samples with read lengths of 150 and 140 were selected. Samples were split and run individually due to poor quality, and quality trimming parameters were applied. The final analysis included 68 samples, which were added to the database.

- **PRJNA767207**: This study examined different therapeutics for IBS patients. Samples were identified using metadata by looking at patient ID and visit number. Primers were removed, and a single read length was used. Analyses started with different cutoff lengths ranging from 455 to 415, with 420 chosen as the final read length.
- **PRJNA637763**: This study used machine learning techniques to classify patients as IBS or healthy. Samples were identified based on the description metatag. Double-ended reads with primers were removed using Cutadapt. Multiple combinations of reads were analyzed, with final cutoff lengths of 235 and 230 selected.
- **PMID-22180058**: This study defined IBS subtypes by species-specific alterations. Data access was granted via personal request, and demultiplexing was done using a provided mapping file. Single-ended reads with primers removed were processed with multiple Qiime2 classifiers.
- **PRJNA812699**: This study identified multi-omics data for IBS subtypes. Samples were differentiated using the gastrointestinal tract disorder tag. Primers were already cut, and cutoff lengths of 247 for forward reads and 237 for reverse reads were used.
- **PRJNA566284**: This study had two different cutoff length ranges, 219 and 270. Samples were batched based on read lengths, with a forward cutoff of 260 and reverse cutoff of 220 for the 270 batch, and 212 for both forward and reverse reads in the 219 batch. Primers were removed before analysis.
- **PRJNA475187**: This study aimed to find gut microbiome alterations in IBS-D patients. Samples were differentiated using the experiment-alias metatag. Primers were removed using Cutadapt, and cutoff lengths of 272 for forward reads and 247 for reverse reads were chosen.
- **PRJNA391149**: This pilot study identified dietary connections in IBS patients. Samples were classified using the meta tag. Based on the percentage of input nonchimeric reads, samples were split into two groups. Primers were removed, and a cutoff length of 320 was used for single-ended samples.
- **PRJNA373876**: This correlation study examined the relationship between gut microbiome and brain volumes. Samples were identified using the host diseases tag, with values for IBS and healthy controls. Primers were removed, reads were merged into single-ended reads, and a cutoff length of 354 was used.
- **PRJNA268708**: This study observed reduced butyrate-producing bacteria in IBS patients. Samples were identified using the description tag to distinguish between IBS and healthy controls. Primers were removed, and a cutoff length of 310 was used.
- **PRJDB5442**: This study aimed to identify potential donors for fecal microbiota transplantation (FMT) in IBS patients. Samples were identified using the description tag. Primers were removed using Cutadapt, and a cutoff length of 320 was chosen.

3.4 From Taxonomical Classification to a Robust Classifier for IBS

The problem is often posed as a classification task to distinguish between IBS patients and healthy controls. The goal is to transform the high-dimensional taxonomic data into informative features for predictive modeling. The process begins with the taxonomic profiles generated by QIIME2 or DADA2, which include relative abundances, presence/absence of taxa, and diversity metrics for each sample. These relative abundances are then used as features that serve as inputs for training machine learning models (classifier). Finally, the model predicts whether a given sample belongs to an IBS patient or a healthy control.

4 Results

In this section, we present our findings of our various experiments. First, we demonstrate that models trained on individual datasets perform poorly when tested across other datasets, highlighting the limitations of isolated dataset approaches. Next, we show that the results of training our best-performing model using the combined dataset. This approach aims to leverage the increased sample size and diversity to improve predictive accuracy and robustness for IBS.

4.1 Classifier Training on individual datasets

Understanding the generalizability of classifiers trained on microbiome data from specific studies is crucial due to significant geographic variations in gut microbiome diversity. This experiment aims to evaluate the performance of a classifier trained on one specific dataset when tested on all other datasets, highlighting the limitations of models trained on isolated datasets and underscoring the necessity for a comprehensive dataset integration approach.

We implemented and compared four different machine learning models: XGBoost, Random Forest, a combination of XGBoost and Random Forest, and Naive Bayes. These models were chosen based on their performance in initial experiments and their use widely in the literature. For all the models we used model parameters as described in Table 3. In our initial experiments, XGBoost outperformed other models (as shown in the next subsection), prompting us to present the results for XGBoost here. Models were trained on individual datasets using taxonomic profiles generated by QIIME2 and tested across other datasets. The results are summarized in Table 2 for XGBoost (results for other models available on the supplementary website). The table shows the performance of XGBoost classifiers trained on one dataset and tested on others, including the number of healthy controls (HC) and IBS samples used for training and testing. For example, the first row of Table 2 shows the results for an XGBoost classifier trained on the dataset with PMID 33925672. This model was trained using 41 healthy control (HC) samples and 36 IBS samples. The trained model was then tested on a combined dataset consisting of 409 HC samples and 966 IBS samples from other studies.

Train PMID	Accuracy	Precision	Recall	F1 Score	AUROC	Train Data	Test Data
33925672	0.544	0.737728	0.544513	0.626563	0.547019	HC: 41, IBS: 36	HC: 409, IBS: 966
36624530	0.48251	0.747449	0.420373	0.538108	0.557579	HC: 175, IBS: 305	HC: 275, IBS: 697
31681624	0.421644	0.786713	0.227964	0.353496	0.544028	HC: 14, IBS: 15	HC: 436, IBS: 987
30090090	0.593268	0.706844	0.702538	0.704684	0.53832	HC: 9, IBS: 17	HC: 441, IBS: 985
28628918	0.670391	0.707006	0.895161	0.790036	0.579504	HC: 10, IBS: 10	HC: 440, IBS: 992
28457228	0.469286	0.759009	0.346351	0.475653	0.58459	HC: 23, IBS: 29	HC: 427, IBS: 973
26239401	0.648881	0.701232	0.85375	0.770011	0.540824	HC: 88, IBS: 202	HC: 362, IBS: 800
22180058	0.635842	0.71577	0.785492	0.749012	0.556972	HC: 20, IBS: 37	HC: 430, IBS: 965
32552668	0.620091	0.706879	0.772331	0.738157	0.574442	HC: 44, IBS: 84	HC: 406, IBS: 918
32727141	0.630127	0.694367	0.820065	0.752	0.530703	HC: 26, IBS: 85	HC: 424, IBS: 917

Table 2 Performance of XGBoost Classifiers Trained on Individual Datasets and Tested on Other Datasets

The results demonstrate that XGBoost classifiers trained on individual datasets perform inconsistently when tested on other datasets, reflecting the limitations of using isolated datasets for training. For instance, the classifier trained on PMID 33925672 achieved an accuracy of 0.544 when tested on the combined dataset of other studies, indicating poor generalizability. Similar trends were observed across other datasets, with accuracy and other performance metrics varying significantly.

4.2 Classifier Training on Combined Datasets

To improve the robustness and accuracy of our IBS classifier, we combined all available datasets into a single comprehensive dataset. This approach leverages the increased sample size and diversity, addressing the limitations observed with individual datasets. To evaluate the performance and robustness of the combined dataset, we employed a 10-fold cross-validation (CV) strategy to give us a good estimate of the model's generalizability.

XGBoost demonstrated the best performance, achieving a mean accuracy of 0.75 with a standard deviation of 0.04 in 10-fold cross-validation. Performance metrics included a precision of 0.79, recall of 0.88, F1 score of 0.83, and a ROC AUC of 0.81 (Table 4).

The combined XGBoost and Random Forest model showed a mean accuracy of 0.73, with precision and recall scores of 0.77 and 0.88, respectively. Random Forest alone achieved similar accuracy (0.73) and a higher recall (0.94). Logistic Regression and Naive Bayes performed less effectively, with accuracies of 0.67 and 0.50, respectively (Table 4).

Model	Parameter
XGBoost	`n_jobs=8`, `n_estimators=150`, `enable_categorical=True`
XGBoost+RF	`n_jobs=8`, `n_estimators=150`, `enable_categorical=True`
Random Forest	`n_estimators=150`, `random_state=1`
Logistic Regression	`max_iter=1000`, `random_state=1`
Naive Bayes	Default parameters

Table 3 Model Parameters Overview.
`n_jobs=8`: Utilizes all available CPU cores for computation, used in XGBoost and XGBoost+RF.
`n_estimators=150`: Number of boosting rounds (in XGBoost) or trees (in Random Forest).
`enable_categorical=True`: Supports categorical data directly in XGBoost.
`random_state=1`: Sets a seed for the random number generator for reproducibility.
`max_iter=1000`: Maximum iterations for solver convergence in Logistic Regression.
Default parameters: Naive Bayes uses default settings.

Table 4 Performance Metrics of Different Machine Learning Models on combined datasets

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
XGBoost	0.75 ± 0.04	0.79 ± 0.04	0.88 ± 0.03	0.83 ± 0.03	0.81
XGBoost+rf	0.73 ± 0.03	0.77 ± 0.02	0.88 ± 0.04	0.82 ± 0.02	0.78
Random Forest	0.73 ± 0.03	0.74 ± 0.02	0.94 ± 0.03	0.83 ± 0.02	0.79
Logistic Regression	0.67 ± 0.04	0.72 ± 0.02	0.83 ± 0.05	0.77 ± 0.03	0.62
Naive Bayes	0.50 ± 0.04	0.81 ± 0.06	0.36 ± 0.05	0.50 ± 0.05	0.62

The ROC curves as shown in Fig 3 illustrate the performance of various classifiers in distinguishing between IBS and healthy samples. XGBoost achieved the highest AUC of 0.81, followed by the combined XGBoost+RF (0.78) and Random Forest (0.79). Logistic Regression and Naive Bayes showed lower AUCs of 0.62.

5 Explainability analysis of the IBS classifier using SHAP

To gain insights into the model's decision-making process, we employed SHAP (SHapley Additive exPlanations) values to interpret the XGBoost model. SHAP values help identify the contribution of each feature, in this case, specific bacteria, to the model's predictions. Our analysis revealed 19 bacteria with significant impacts on the model's performance as shown in Fig 4.

5.1 Key Findings from SHAP Analysis

The SHAP analysis identified 19 key bacterial taxa influencing the IBS classifier's predictions. Notably, 16 of these bacteria have been previously mentioned in the literature, underscoring their relevance in IBS studies.

5.2 Summary of Bacterial Contributions

Table 5 summarizes the bacteria with positive and negative correlations to IBS, derived from SHAP values and supported by literature evidence. Positive correlations indicate

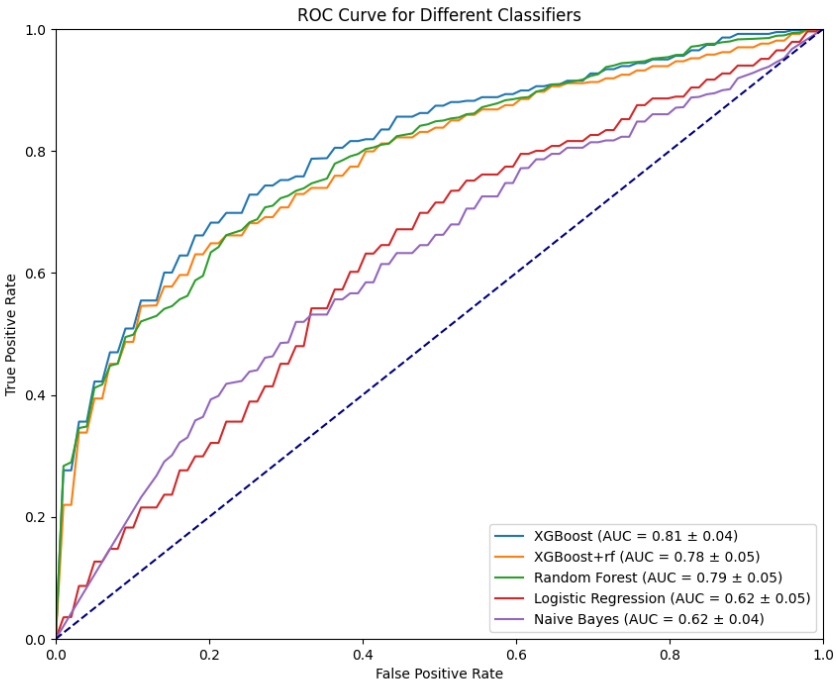

Fig. 3 Receiver Operating Characteristic (ROC) Curves for Different Classifiers

that increased levels of these bacteria are associated with a higher likelihood of IBS, while negative correlations suggest the opposite.

5.3 Detailed Insights from SHAP Values and Literature

5.3.1 Positively Correlated Bacteria

- Pseudomonas: High SHAP values indicate a strong association with IBS. Literature supports this, suggesting that increased levels of Pseudomonas are linked to a higher likelihood of IBS [30].
- Lachnoclostridium: Positive SHAP values imply a higher risk of IBS with elevated levels. This is corroborated by studies showing a positive correlation with IBS [19].

5.3.2 Negatively Correlated Bacteria

- Bacteroides: SHAP values reveal a negative association with IBS, with literature showing predominantly negative correlations, indicating lower levels reduce IBS risk [33][34][36][32][37][38][39][40].

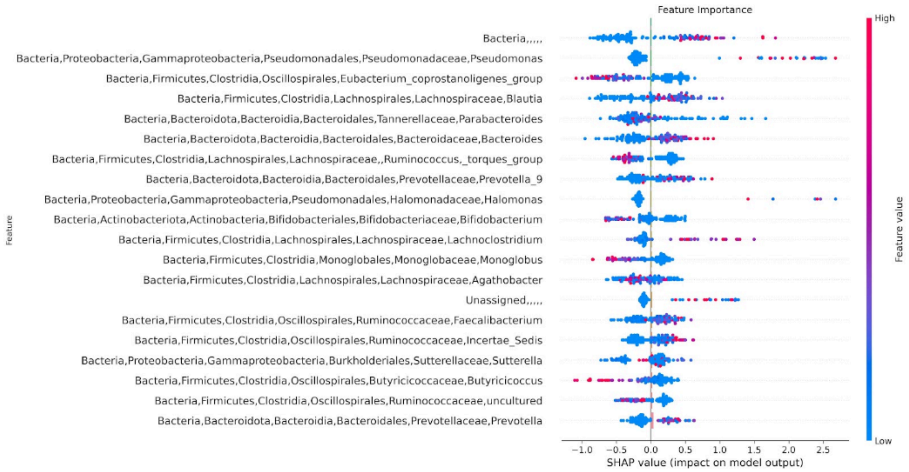

Fig. 4 SHAP Values Indicating Key Bacterial Contributions to IBS Predictions by the XGBoost Classifier

Table 5 Summary of Bacterial Correlations with IBS: Positive and Negative Associations

Bacteria	**Positive co-rel**	**References**	**Negative co-rel**	**References**
Pseudomonas	1	[30]	0	-
Bacteroides	2	[31, 32]	6	[33–39]
Butyricicoccus	0	-	1	[40]
Blautia	1	[41]	1	[32]
Sutterella	0	-	1	[42]
Lachnoclostridium	1	[19]	0	-
Parabacteroides	0	-	3	[36, 37, 43]
Bifidobacterium	0	-	6	[33, 39, 44–47]
Prevotella_9	1	[48]	0	-
Eubacterium_coprostanoligenes_group	0	-	1	[48]
Anaerostipes	0	-	1	[49]
Butyricimonas	0	-	1	[50]
Faecalibacterium	1	[34]	5	[32, 43, 45, 46, 49]
Ruminococcaceae DTU089	0	-	1	[50]
Subdoligranulum	1	[42]	0	-
Alistipes	0	-	1	[43]

- Sutterella: Negative SHAP values and supporting literature suggest that reduced levels of Sutterella are linked to a lower likelihood of IBS [42].
- Bifidobacterium: Consistently negative SHAP values indicate lower IBS risk, supported by multiple studies [44][33][19][45][38][19][39]
- Anaerostipes: Negative SHAP values suggest a reduction in IBS risk with higher levels, corroborated by literature [49].

5.3.3 Mixed Impact Bacteria

- Blautia: SHAP values show both positive and negative impacts, reflecting the variability within this genus. Literature reports mixed correlations, indicating that different species of Blautia have varying effects on IBS [32, 41].
- Faecalibacterium: Both positive and negative SHAP values align with mixed literature findings, suggesting species-specific impacts on IBS [34], [45][49][32][46][43].

This explainability analysis using SHAP values enhances our understanding of the model's predictions and also aligns with existing literature, validating the model's relevance in identifying robust microbial signatures associated with IBS.

6 Discussions and Conclusion

In this study, we demonstrate the benefit of integrating diverse datasets to improve the accuracy and robustness of an IBS classifier. Our initial experiments revealed when a classifier is trained only on a single data set, it displays significant variability and loss of precision when tested on different datasets, highlighting the need for larger datasets to enhance model generalizability. By combining several publicly available IBS datasets, which encompass different geographies and demographics, we accounted for a broader spectrum of microbiome variations, creating a comprehensive dataset for IBS classification. In our 10-fold cross-validation on the combined dataset, we found that models performed better overall with XGBoost achieving the highest mean accuracy. SHAP analysis provided insights into the model's decision-making, identifying key bacterial taxa influencing predictions. The identified bacteria aligned with existing literature.

In terms of future work, we believe that the model performance can be further improved by integrating clinical metadata alongside microbiome data. Combining clinical parameters such as dietary habits, medication usage, and symptom severity with microbiome data can provide a more holistic understanding of IBS and improve model predictions. Additionally, exploring advanced machine learning techniques such as deep learning and ensemble learning methods could yield better performance. Further, incorporating multi-omics data, including metagenomics, metabolomics, and transcriptomics, could provide a deeper understanding of the IBS-associated microbiome. Multi-omics integration can reveal interactions between different biological layers, offering a more comprehensive view.

Supplementary information. All data, code, models, and additional results of the paper are available online at https://bit.ly/IntegrIBS

Acknowledgements. The Authors would like to acknowledge colleagues, especially Marcus Mata and Gobika Krishnan, for doing the literature review and downloading the databases. Author Sucheendra K. Palaniappan would like to acknowledge and thank the ONRG Grant for the Nobel Turing Challenge to The Systems Biology Institute (Grant number: N62909-21-1-2032) for funding support for part of this study.

References

[1] Shaikh, S.D., Sun, N., Canakis, A., Park, W.Y., Weber, H.C.: Irritable bowel syndrome and the gut microbiome: a comprehensive review. Journal of Clinical Medicine **12**(7), 2558 (2023)

[2] Collins, S.M.: A role for the gut microbiota in IBS. Nature Reviews Gastroenterology & Hepatology **11**(8), 497–505 (2014) https://doi.org/10.1038/nrgastro.2014.40 . Publisher: Nature Publishing Group. Accessed 2024-06-21

[3] Andrews, C.N., Sidani, S., Marshall, J.K.: Clinical Management of the Microbiome in Irritable Bowel Syndrome. Journal of the Canadian Association of Gastroenterology **4**(1), 36–43 (2021) https://doi.org/10.1093/jcag/gwz037 . Accessed 2024-06-21

[4] Ford, A.C., Sperber, A.D., Corsetti, M., Camilleri, M.: Irritable bowel syndrome. The Lancet **396**(10263), 1675–1688 (2020) https://doi.org/10.1016/S0140-6736(20)31548-8 . Publisher: Elsevier. Accessed 2024-06-21

[5] Camilleri, M.: Diagnosis and Treatment of Irritable Bowel Syndrome: A Review. JAMA **325**(9), 865–877 (2021) https://doi.org/10.1001/jama.2020.22532 . Accessed 2024-06-21

[6] Wang, X.-W., Liu, Y.-Y.: Comparative study of classifiers for human microbiome data. Medicine in Microecology **4**, 100013 (2020) https://doi.org/10.1016/j.medmic.2020.100013 . Accessed 2024-06-21

[7] Fukui, H., Nishida, A., Matsuda, S., Kira, F., Watanabe, S., Kuriyama, M., Kawakami, K., Aikawa, Y., Oda, N., Arai, K., Matsunaga, A., Nonaka, M., Nakai, K., Shinmura, W., Matsumoto, M., Morishita, S., Takeda, A.K., Miwa, H.: Usefulness of Machine Learning-Based Gut Microbiome Analysis for Identifying Patients with Irritable Bowels Syndrome. Journal of Clinical Medicine **9**(8), 2403 (2020) https://doi.org/10.3390/jcm9082403 . Accessed 2024-06-21

[8] Tanaka, K., Tanigawa, N., Song, I., Komatsu, T., Kuriki, Y., Tanaka, Y., Fukudo, S., Urano, Y., Fukuda, S.: A protease activity-based machine-learning approach as a complementary tool for conventional diagnosis of diarrhea-predominant irritable bowel syndrome. Frontiers in Microbiology **14** (2023) https://doi.org/10.3389/fmicb.2023.1179534 . Publisher: Frontiers. Accessed 2024-06-21

[9] Carcy, S., Ostner, J., Tran, V., Menden, M., Müller, C.L.: MetaIBS - large-scale amplicon-based meta analysis of irritable bowel syndrome. bioRxiv. Pages: 2024.01.22.575775 Section: New Results (2024). https://doi.org/10.1101/2024.01.22.575775 . https://www.biorxiv.org/content/10.1101/2024.01.22.575775v1 Accessed 2024-06-21

[10] Recent advances in the treatment of irritable bowel syndrome. https://doi.org/

10.20452/pamw.16067 . https://www.mp.pl/paim/issue/article/16067 Accessed 2024-06-21

[11] Jeffery, I.B., O'Toole, P.W., Öhman, L., Claesson, M.J., Deane, J., Quigley, E.M.M., Simrén, M.: An irritable bowel syndrome subtype defined by species-specific alterations in faecal microbiota. Gut **61**(7), 997–1006 (2012) https://doi.org/10.1136/gutjnl-2011-301501 . Publisher: BMJ Publishing Group Section: Irritable bowel syndrome. Accessed 2024-06-19

[12] Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., *et al.*: The European nucleotide archive. Nucleic acids research **39**(suppl_1), 28–31 (2010). Publisher: Oxford University Press

[13] Gostev, M., Faulconbridge, A., Brandizi, M., Fernandez-Banet, J., Sarkans, U., Brazma, A., Parkinson, H.: The BioSample database (BioSD) at the european bioinformatics institute. Nucleic acids research **40**(D1), 64–70 (2012). Publisher: Oxford University Press

[14] Mizuno, S., Masaoka, T., Naganuma, M., Kishimoto, T., Kitazawa, M., Kurokawa, S., Nakashima, M., Takeshita, K., Suda, W., Mimura, M., Hattori, M., Kanai, T.: Bifidobacterium-Rich Fecal Donor May Be a Positive Predictor for Successful Fecal Microbiota Transplantation in Patients with Irritable Bowel Syndrome. Digestion **96**(1), 29–38 (2017) https://doi.org/10.1159/000471919 . Accessed 2024-06-19

[15] Pozuelo, M., Panda, S., Santiago, A., Mendez, S., Accarino, A., Santos, J., Guarner, F., Azpiroz, F., Manichanh, C.: Reduction of butyrate- and methane-producing microorganisms in patients with Irritable Bowel Syndrome. Scientific Reports **5**(1), 12693 (2015) https://doi.org/10.1038/srep12693 . Publisher: Nature Publishing Group. Accessed 2024-06-19

[16] Labus, J.S., Hollister, E.B., Jacobs, J., Kirbach, K., Oezguen, N., Gupta, A., Acosta, J., Luna, R.A., Aagaard, K., Versalovic, J., Savidge, T., Hsiao, E., Tillisch, K., Mayer, E.A.: Differences in gut microbial composition correlate with regional brain volumes in irritable bowel syndrome. Microbiome **5**(1), 49 (2017) https://doi.org/10.1186/s40168-017-0260-z . Accessed 2024-06-19

[17] Altomare, A., Del Chierico, F., Rocchi, G., Emerenziani, S., Nuglio, C., Putignani, L., Angeletti, S., Lo Presti, A., Ciccozzi, M., Russo, A., Cocca, S., Ribolsi, M., Muscaritoli, M., Cicala, M., Guarino, M.P.L.: Association between Dietary Habits and Fecal Microbiota Composition in Irritable Bowel Syndrome Patients: A Pilot Study. Nutrients **13**(5), 1479 (2021) https://doi.org/10.3390/nu13051479 . Number: 5 Publisher: Multidisciplinary Digital Publishing Institute. Accessed 2024-06-19

[18] Zhuang, X., Tian, Z., Li, L., Zeng, Z., Chen, M., Xiong, L.: Fecal Microbiota

Alterations Associated With Diarrhea-Predominant Irritable Bowel Syndrome. Frontiers in Microbiology **9** (2018) https://doi.org/10.3389/fmicb.2018.01600 . Publisher: Frontiers. Accessed 2024-06-19

[19] Zhu, S., Liu, S., Li, H., Zhang, Z., Zhang, Q., Chen, L., Zhao, Y., Chen, Y., Gu, J., Min, L., Zhang, S.: Identification of Gut Microbiota and Metabolites Signature in Patients With Irritable Bowel Syndrome. Frontiers in Cellular and Infection Microbiology **9** (2019) https://doi.org/10.3389/fcimb.2019.00346 . Publisher: Frontiers. Accessed 2024-06-19

[20] Jacobs, J.P., Lagishetty, V., Hauer, M.C., Labus, J.S., Dong, T.S., Toma, R., Vuyisich, M., Naliboff, B.D., Lackner, J.M., Gupta, A., Tillisch, K., Mayer, E.A.: Multi-omics profiles of the intestinal microbiome in irritable bowel syndrome and its bowel habit subtypes. Microbiome **11**(1), 5 (2023) https://doi.org/10.1186/s40168-022-01450-5 . Accessed 2024-06-15

[21] Ivashkin, V.T., Kudryavtseva, A.V., Krasnov, G.S., Poluektov, Y.M., Morozova, M.A., Shifrin, O.S., Beniashvili, A.G., Mamieva, Z.A., Kovaleva, A.L., Ulyanin, A.I., Trush, E.A., Erlykin, A.G., Poluektova, E.A.: Efficacy and safety of a food supplement with standardized menthol, limonene, and gingerol content in patients with irritable bowel syndrome: A double-blind, randomized, placebo-controlled trial. PLOS ONE **17**(6), 0263880 (2022) https://doi.org/10.1371/journal.pone.0263880 . Publisher: Public Library of Science. Accessed 2024-06-19

[22] Fodor, A.A., Pimentel, M., Chey, W.D., Lembo, A., Golden, P.L., Israel, R.J., Carroll, I.M.: Rifaximin is associated with modest, transient decreases in multiple taxa in the gut microbiota of patients with diarrhoea-predominant irritable bowel syndrome. Gut Microbes **10**(1), 22–33 (2019) https://doi.org/10.1080/19490976.2018.1460013 . Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19490976.2018.1460013. Accessed 2024-06-19

[23] Liu, T., Gu, X., Li, L.-X., Li, M., Li, B., Cui, X., Zuo, X.-l.: Microbial and metabolomic profiles in correlation with depression and anxiety co-morbidities in diarrhoea-predominant IBS patients. BMC Microbiology **20**(1), 168 (2020) https://doi.org/10.1186/s12866-020-01841-4 . Accessed 2024-06-19

[24] Li, M., Xu, R., Li, Y.-Q.: Sequential laxative-probiotic usage for treatment of irritable bowel syndrome: a novel method inspired by mathematical modelling of the microbiome. Scientific Reports **10**(1), 19291 (2020) https://doi.org/10.1038/s41598-020-75225-z

[25] Suzuki, T.A., Worobey, M.: Geographical variation of human gut microbial composition. Biology letters **10**(2), 20131037 (2014)

[26] Loftus, M., Hassouneh, S.A.-D., Yooseph, S.: Bacterial associations in the healthy human gut microbiome across populations. Scientific reports **11**(1), 2828 (2021)

[27] Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal **17**(1), 10–12 (2011)

[28] Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P.: Dada2: High-resolution sample inference from illumina amplicon data. Nature methods **13**(7), 581–583 (2016)

[29] Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., Glöckner, F.O.: Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. Nucleic acids research **35**(21), 7188–7196 (2007)

[30] Wu, B.-Y., Xu, P., Cheng, L., Wang, Q.-Q., Qiu, H.-Y., Yan, X.-J., Chen, S.-L.: Mucosa-Associated Microbiota Dysbiosis in the Terminal Ileum Correlates With Bowel Symptoms in Diarrhea-Predominant Irritable Bowel Syndrome. Clinical and Translational Gastroenterology **15**(2), 00658 (2024) https://doi.org/10.14309/ctg.0000000000000658 . Accessed 2024-06-20

[31] Liu, Y., Zhang, L., Wang, X., Wang, Z., Zhang, J., Jiang, R., Wang, X., Wang, K., Liu, Z., Xia, Z., Xu, Z., Nie, Y., Lv, X., Wu, X., Zhu, H., Duan, L.: Similar Fecal Microbiota Signatures in Patients With Diarrhea-Predominant Irritable Bowel Syndrome and Patients With Depression. Clinical Gastroenterology and Hepatology **14**(11), 1602–16115 (2016) https://doi.org/10.1016/j.cgh.2016.05.033 . Publisher: Elsevier. Accessed 2024-06-21

[32] Tap, J., Derrien, M., Törnblom, H., Brazeilles, R., Cools-Portier, S., Doré, J., Störsrud, S., Le Nevé, B., Öhman, L., Simrén, M.: Identification of an Intestinal Microbiota Signature Associated With Severity of Irritable Bowel Syndrome. Gastroenterology **152**(1), 111–1238 (2017) https://doi.org/10.1053/j.gastro.2016.09.049 . Publisher: Elsevier. Accessed 2024-06-20

[33] Kim, S.-E., Choi, S.C., Park, K.S., Park, M.I., Shin, J.E., Lee, T.H., Jung, K.W., Koo, H.S., Myung, S.-J.: Change of Fecal Flora and Effectiveness of the Short-term VSL#3 Probiotic Treatment in Patients With Functional Constipation. Journal of Neurogastroenterology and Motility **21**(1), 111–120 (2015) https://doi.org/10.5056/jnm14048 . Accessed 2024-06-20

[34] Mancabelli, L., Milani, C., Lugli, G.A., Turroni, F., Mangifesta, M., Viappiani, A., Ticinesi, A., Nouvenne, A., Meschi, T., Sinderen, D., Ventura, M.: Unveiling the gut microbiota composition and functionality associated with constipation through metagenomic analyses. Scientific Reports **7**, 9879 (2017) https://doi.org/10.1038/s41598-017-10663-w . Accessed 2024-06-20

[35] Swidsinski, A., Loening–Baucke, V., Verstraelen, H., Osowska, S., Doerffel, Y.: Biostructure of Fecal Microbiota in Healthy Subjects and Patients With Chronic Idiopathic Diarrhea. Gastroenterology **135**(2), 568–5792 (2008) https://doi.org/10.1053/j.gastro.2008.04.017 . Publisher: Elsevier. Accessed 2024-06-21

[36] Vervier, K., Moss, S., Kumar, N., Adoum, A., Barne, M., Browne, H., Kaser, A., Kiely, C.J., Neville, B.A., Powell, N., Raine, T., Stares, M.D., Zhu, A., Negro, J.D.L.R., Lawley, T.D., Parkes, M.: Two microbiota subtypes identified in irritable bowel syndrome with distinct responses to the low FODMAP diet. Gut **71**(9), 1821–1830 (2022) https://doi.org/10.1136/gutjnl-2021-325177 . Publisher: BMJ Publishing Group Section: Irritable bowel syndrome. Accessed 2024-06-20

[37] Noor, S.O., Ridgway, K., Scovell, L., Kemsley, E.K., Lund, E.K., Jamieson, C., Johnson, I.T., Narbad, A.: Ulcerative colitis and irritable bowel patients exhibit distinct abnormalities of the gut microbiota. BMC Gastroenterology **10**(1), 134 (2010) https://doi.org/10.1186/1471-230X-10-134 . Accessed 2024-06-20

[38] Zhao, L., Yang, W., Chen, Y., Huang, F., Lu, L., Lin, C., Huang, T., Ning, Z., Zhai, L., Zhong, L.L.D., Lam, W., Yang, Z., Zhang, X., Cheng, C., Han, L., Qiu, Q., Shang, X., Huang, R., Xiao, H., Ren, Z., Chen, D., Sun, S., El-Nezami, H., Cai, Z., Lu, A., Fang, X., Jia, W., Bian, Z.: A *Clostridia*-rich microbiota enhances bile acid excretion in diarrhea-predominant irritable bowel syndrome. The Journal of Clinical Investigation **130**(1), 438–450 (2024) https://doi.org/10.1172/JCI130976 . Publisher: American Society for Clinical Investigation. Accessed 2024-06-20

[39] Ponnusamy, K., Choi, J.N., Kim, J., Lee, S.-Y., Lee, C.H.: Microbial community and metabolomic comparison of irritable bowel syndrome faeces. Journal of Medical Microbiology **60**(6), 817–827 (2011) https://doi.org/10.1099/jmm.0.028126-0 . Publisher: Microbiology Society,. Accessed 2024-06-20

[40] Napolitano, M., Fasulo, E., Ungaro, F., Massimino, L., Sinagra, E., Danese, S., Mandarino, F.V.: Gut Dysbiosis in Irritable Bowel Syndrome: A Narrative Review on Correlation with Disease Subtypes and Novel Therapeutic Implications. Microorganisms **11**(10), 2369 (2023) https://doi.org/10.3390/microorganisms11102369 . Accessed 2024-06-20

[41] Zhang, S., Wang, R., Li, D., Zhao, L., Zhu, L.: Role of gut microbiota in functional constipation. Gastroenterology Report **9**(5), 392–401 (2021) https://doi.org/10.1093/gastro/goab035 . Accessed 2024-06-20

[42] Su, Q., Tun, H.M., Liu, Q., Yeoh, Y.K., Mak, J.W.Y., Chan, F.K., Ng, S.C.: Gut microbiome signatures reflect different subtypes of irritable bowel syndrome. Gut Microbes **15**(1), 2157697 https://doi.org/10.1080/19490976.2022.2157697 . Accessed 2024-06-20

[43] Rajilić–Stojanović, M., Biagi, E., Heilig, H.G.H.J., Kajander, K., Kekkonen, R.A., Tims, S., Vos, W.M.d.: Global and Deep Molecular Analysis of Microbiota Signatures in Fecal Samples From Patients With Irritable Bowel Syndrome. Gastroenterology **141**(5), 1792–1801 (2011) https://doi.org/10.1053/j.gastro.2011.07.043 . Publisher: Elsevier. Accessed 2024-06-20

[44] Ohkusa, T., Koido, S., Nishikawa, Y., Sato, N.: Gut Microbiota and Chronic

Constipation: A Review and Update. Frontiers in Medicine **6** (2019) https://doi.org/10.3389/fmed.2019.00019 . Publisher: Frontiers. Accessed 2024-06-20

[45] Mei, L., Zhou, J., Su, Y., Mao, K., Wu, J., Zhu, C., He, L., Cui, Y.: Gut microbiota composition and functional prediction in diarrhea-predominant irritable bowel syndrome. BMC Gastroenterology **21**(1), 105 (2021) https://doi.org/10.1186/s12876-021-01693-w . Accessed 2024-06-20

[46] Rajilić-Stojanović, M., Jonkers, D.M., Salonen, A., Hanevik, K., Raes, J., Jalanka, J., Vos, W.M., Manichanh, C., Golic, N., Enck, P., Philippou, E., Iraqi, F.A., Clarke, G., Spiller, R.C., Penders, J.: Intestinal Microbiota And Diet in IBS: Causes, Consequences, or Epiphenomena? The American Journal of Gastroenterology **110**(2), 278–287 (2015) https://doi.org/10.1038/ajg.2014.427 . Accessed 2024-06-20

[47] Kerckhoffs, A.P., Samsom, M., Rest, M.E., Vogel, J., Knol, J., Ben-Amor, K., Akkermans, L.M.: Lower Bifidobacteria counts in both duodenal mucosa-associated and fecal microbiota in irritable bowel syndrome patients. World Journal of Gastroenterology : WJG **15**(23), 2887–2892 (2009) https://doi.org/10.3748/wjg.15.2887 . Accessed 2024-06-20

[48] Liu, Y., Li, W., Yang, H., Zhang, X., Wang, W., Jia, S., Xiang, B., Wang, Y., Miao, L., Zhang, H.: Leveraging 16S rRNA microbiome sequencing data to identify bacterial signatures for irritable bowel syndrome. Frontiers in cellular and infection microbiology **11**, 645951 (2021) https://doi.org/10.3389/fcimb.2021.645951 . Publisher: Frontiers Media SA. Accessed 2024-06-15

[49] Agnello, M., Carroll, L.N., Imam, N., Pino, R., Palmer, C., Varas, I., Greene, C., Hitschfeld, M., Gupta, S., Almonacid, D.E., Hoaglin, M.C.: Gut microbiome composition and risk factors in a large cross-sectional IBS cohort. BMJ Open Gastroenterology **7**(1), 000345 (2020) https://doi.org/10.1136/bmjgast-2019-000345 . Accessed 2024-06-20

[50] Smirnova, Y.D., Gryaznova, M.V., Burakova, I.Y., Syromyatnikov, M.Y., Sviridova, T.N., Lebedeva, O.P., Maslov, A.Y., Popov, V.N.: Study of microbiome aberrations in patients with irritable bowel syndrome with diarrhea by next-generation sequencing **9**(4), 446–460 (2023)