Metagenomics - AI powered prediction of Inflammatory Bowel's Disease and Probiotic Recommendation

Mathew Thomas[a]

Sastha Kumar N[ab]

Sai Janakiram[a]

Dr. Sabarinath Subramaniam[a]

Dr. Nidheesh M[b]

[a]GenAI Research Labs,Sivasakthi Science Foundation ,Kowdiar, Thiruvananthapuram,695003, Kerala, India

[b]Associate Professor, School of Biotechnology, Amritapuri, Campus, Amritapuri, Clappana P.O., Kollam, 690525, Kerala, India

Corresponding author:

Dr. Sabarinath Subramaniam

Director,

Sivasakthi Science Foundation,

Kowdiar, Thiruvananthapuram,695003, Kerala, India

Phone: +1(704)771-0203

Email: shabari@sivasakthifoundation.org

Abstract

Background and Objective: The dysbiosis of human gut microbiome has been increasingly seen to have a relation in the development of autoimmune diseases, with specific microbial signatures having causative association with specific conditions. Inflammatory bowel disease (IBD) is one such autoimmune ailment. This paper proposes a predictive tool that can identify the IBD status of an individual based on the composition of the gut microbiome using machine learning and AI agents driven techniques. The technology can strengthen the suspicion of a potential IBD diagnosis a patient may have based on their gut microbiome profile. **Methods:** The tool processes patient gut metagenome using integrated Kneaddata and MetaPhlAn to generate taxonomic profiles. These are fed into an XGBoost classifier to predict IBD or healthy status. Dysbiotic taxa are identified via Z-score and fold change. CrewAI delivers personalized probiotic recommendations based on diagnosis and dysbiosis.

Results: The tuned XGBoost model achieved 86.6% accuracy. On validation using single ulcerative colitis sample, the tool correctly predicted IBD status but misclassified it as Crohn's disease(possibly due to overlapping microbial signatures), identifying Faecalibacterium and Flavonifractor as dysbiotic taxa.The probiotic recommended was Faecalibacterium prausnitzii ,backed with reasoning basedon scientific literature.

Conclusions: Despite limited validation sample size, the high accuracy , correct IBD detection ,dysbiosis analysis and elaborate probiotic recommendation suggest promising potential; further validation needed

Keywords: Metagenomics, Inflammatory Bowel Disease, Machine Learning, AI Agents, Gut Dysbiosis, Probiotics

1. Introduction

The human body consists of a vast community of trillions of microbes, collectively known as the human microbiome, with the gastrointestinal tract, particularly the colon, being the most densely populated habitat collectively called the human gut microbiome. This intricate community, composed of over 1000 bacterial species, maintains a crucial symbiotic relationship with the host, influencing vital processes such as digestion, vitamin production, and immune system modulation (Durack et al., 2018). While the taxonomic profiles of individual microbiomes vary significantly due to factors like geography, diet, and host genetics, their underlying functional potential via the metabolic pathways shows remarkable consistency across populations. However, any imbalance in the composition and function of this microbiome, which is consistent, is referred to as dysbiosis. Dysbiosis is strongly linked to the development and progression of chronic conditions, including metabolic, autoimmune, and neurological diseases. Currently, it is increasingly recognized as a key environmental factor that can shift the immune response towards pro-inflammatory pathways, thereby instigating autoimmune processes(Luca & Shoenfield,2018).

IBDs, encompassing Crohn's disease (CD) and Ulcerative Colitis (UC), are characterized by chronic gastrointestinal inflammation and exhibit altered gut microbiome compositions, including reduced bacterial diversity, a decrease in beneficial bacteria (like *Faecalibacterium prausnitzii*) (Belizário & Faintuch,2018), and increase in potentially harmful or opportunistic taxa (like Proteobacteria, *Escherichia-Shigella*, and *Enterococcus*)(Alam et al.,2020).

The mechanisms by which gut dysbiosis contributes to autoimmune diseases include the loss of intestinal barrier integrity, which is popularly referred to as “Leaky Gut,” and has been observed in various autoimmune diseases. Tight junctions, which include proteins like occludin, are crucial for maintaining the barrier function. Disruption of the barrier allows microbial products and bacteria to leak into the host tissues and bloodstream, potentially causing the immune system to respond(Fasano,2011; Paray et al., 2020). This allows microbial products to trigger systemic inflammation and molecular mimicry, where microbial antigens structurally similar to some host self-antigens result in the onset of an immune reaction against the self-antigens. Also, the modulation of the immune system by microbial metabolites like short-chain fatty acids (SCFAs) can lead to an increase in systemic inflammation(Sadeghpour Heravi,2024).

According to current research, diagnostic delays are frequent; on average, they occur between 2 months and 8 years after the onset of symptoms (Cross et al., 2023). The diagnosis delay for UC is usually 4 months, while for CD, it is roughly 9 months. Children and mild or vague presentation cases are more likely to have delays (Vavricka et al., 2011). Traditional IBD diagnostic methods face challenges due to symptom variability and invasiveness, which can hinder an accurate and timely diagnosis. The validity and reproducibility of evaluations are affected by the significant variability between observers that hinders endoscopic scoring systems as a whole, such as the Mayo score, which is utilized for assessing the progression of the disease (Javaid et al., 2022).

In response, machine learning (ML), combined with high-throughput sequencing technologies like shotgun metagenomics, has emerged as a powerful computational approach for analyzing complex microbiome data, enabling the accurate classification of IBD patients and the identification of microbial biomarkers for disease prediction (D’Urso & Broccolo , 2024).

Improvements in metagenomics, metatranscriptomics, and metabolomics have become essential in deciphering the intricate interactions between the gut microbiome and IBD. Integration of metagenomics data with host genomic data and clinical metadata is crucial to deciphering the intricate host-microbe interactions underlying IBD and designing microbiome-based diagnostics, prognostics, and therapeutic interventions.

Probiotics are live microorganisms that, when administered adequately, confer health benefits through various mechanisms, including direct antimicrobial activity, strengthening the intestinal barrier by enhancing tight junctions, and modulating the host immune system to reduce inflammation (Shan et al., 2022; Guo & Lv, 2023). While evidence for their effectiveness in Crohn's disease is often limited, substantial evidence supports use of specific probiotic strains, such as *Escherichia coli Nissle 1917* and the multi-strain preparation VSL#3, for inducing and maintaining remission in patients with mild to moderate Ulcerative Colitis (Derwa et al., 2017). While generally considered safe one must use only specific probiotic strains which are documented to have beneficial effects on mild or moderate inflammatory diseases, as wrong recommendations could result in side effects in immunocompromised patients. Crucially, the efficacy of probiotics is strain-specific and dose-dependent, hence the variability in research findings(Estevinho et al., 2024). This highlights the need for personalized insights on what probiotic strains could be useful for a particular individual (Westfall et al.,2021).This paper presents a Metagenomic-ML-based tool that predicts IBD via gut microbiome profile and uses CrewAI to deliver personalized probiotics recommendation.

2. Methodology

2.1 Study Design

The tool aims to fulfill the following purpose: predicting user's disease state and, if diseased, recommending suitable probiotic. The tool is structured into four distinct parts: generating a user's taxonomic profile, predicting the disease using a machine learning model that is trained on the taxonomic profiles of healthy individuals and individuals with IBD, identifying the dysbiotic taxa, and thereby providing personalized probiotic recommendations using CrewAI, if a disease state exists.

2.2 Data Collection

2.2.1 Dataset for IBD prediction model

The data utilized for this project were derived from shotgun metagenomic sequencing, which provides detailed taxonomic profiles of microbial communities(McIve et al., 2018). Two primary sources were integrated: Dataset 1 was retrieved from the Inflammatory Bowel Disease Multi'omics Database (IBDMDB), part of the Integrative Human Microbiome Project (HMP2), contributing 1638 samples with taxonomic classification up to the strain level. Dataset 2 originated from the results of the paper "Gut microbiome structure and metabolic activity in inflammatory bowel disease" (Franzosa et al., 2018), comprising 220 stool samples from two cohorts (155 from Massachusetts, 65 from the Netherlands). These samples were from individuals with Crohn's disease (CD), Ulcerative colitis (UC), and non-IBD controls, with taxonomic classification up to the species level. Both the dataset included metadata features Age, Immunosuppressant usage, Antibiotic usage, and Diagnosis.

2.2.2 Dataset for metagenomic tools

Tools like Knead Data and MetaPhlAn2 were used to process the users' metagenome, with KneadData using the GRCh38 human genome to remove host DNA and MetaPhlAn using the mpa_v31_CHOCOPhlAn_201901 database for taxonomic profiling based on over 17,000 microbial genomes. Both these tools were incorporated into our pipeline.

2.3 Data Preprocessing

A significant challenge in integrating these datasets was their inconsistency, specifically differing levels of taxonomic classification. To address this, a preprocessing approach was employed, involving two key steps: First, collapsing the organisms to a higher taxonomic level, specifically the genus level, by summing the relative abundances of species or strains within the same genus . This increased feature commonality between datasets. Second, the datasets were merged, ensuring that all features (organisms and metadata) from both datasets were retained, with uncommon taxa assigned a relative abundance of zero to maintain sum totals. The resulting merged dataset contained 1858 samples and a taxonomic profile of 195 genera, representing 485 healthy, 838 CD, and 535 UC samples. Other preprocessing steps included relative abundance normalization (renormalizing genus-level abundances to sum to 1), metadata encoding (e.g., Diagnosis: Healthy=0, CD=1, UC=2; Immunosuppressant/Antibiotic Usage: Yes=1, No=0), and min-max normalization of Age to a range between 0 and 1.

2.4 Tool Development and Evaluation

2.4.1 User Taxonomic Profile Generation

The first part of the methodology focuses on User Taxonomic Profile Generation. This begins with processing raw paired-end FASTQ files from the user's gut metagenome sample. Kneaddata (v0.12.0) is employed to meticulously remove host-contaminant DNA, specifically human DNA, and low-quality reads(https://github.com/biobakery/kneaddata). This step involves several integrated tools: Bowtie2 (v2.4.1) is used for alignment against the GRCh38 human genome (a pre-indexed database), and Trimmomatic (v0.39) handles adapter trimming and quality filtering(Han et al.,2024). Following host DNA removal, these processed paired-end reads are fed into MetaPhlAn3 (v3.0.14) to generate the crucial taxonomic profile. MetaPhlAn3 utilizes the mpa_v31_CHOCOPhlAn_201901 database for its profiling, specifically designed to output genus-level abundances by using the --tax_lev g flag(Beghini et al., 2021). The relative abundances of these genera are then normalized to sum to 1.0 per sample. Unclassified genera are removed, and the profile undergoes feature alignment for consistency with training dataset, making it a valid input for prediction(Lin et al.,2020).

2.4.2 Disease Prediction Using XGBoost

The second core component is Disease Prediction, which leverages an XGBoost machine learning model. XGBoost was strategically chosen due to its robust capabilities in handling high-dimensional data typical of gut microbiome profiles (such as 195 genera and 4 clinical metadata

features)(Chang et al., 2023). Its effectiveness stems from its ability to model the complex and nonlinear patterns and interactions among genera that determine an individual's IBD status(Ma et al., 2020).

The model's hyperparameters are carefully tuned to ensure optimal performance. To find the best combination of hyperparameters, RandomSearchCV is employed. This method randomly samples 20 combinations from the defined parameter space, evaluating each using 5-fold cross-validation with accuracy as the scoring metric. This approach efficiently identifies optimal settings, striking a balance between model complexity and generalization capabilities, thereby enhancing prediction accuracy (Bergstra & Bengio, 2012).

The XGBoost estimator is configured to use multi-class logarithmic loss to rank feature importance. Recursive feature elimination is used to iteratively remove the lowest-ranked features until a predefined number of 180 genera are selected. This feature selection is performed exclusively on the training dataset, which is an 80/20 split of the main dataset, to prevent information leakage into the test set. The selected genera are then consistently applied to the test dataset. The XGBoost model itself is optimized for multi-class logarithmic loss, which is ideal for the CD/UC/Healthy classification task.

2.4.3 Dysbiotic Taxa Prediction

The third critical component is the dysbiotic taxa identification.This tool identifies the top 2 dysbiotic taxa that are most altered compared to a healthy gut microbiome and are likely contributing to the inflammatory environment(Kim et al., 2024). To achieve this, a healthy cohort baseline dataset is established by filtering all healthy samples from the same dataset used to train the XGBoost model and then taking the average of the relative abundances of each genus

present in each healthy sample, thereby creating a baseline healthy taxonomic profile. This baseline is then used for comparison against the patient’s taxonomic profile.

Two statistical techniques are used for Dysbiosis Quantification: Z-score and Fold Change

Z-score measures how many standard deviations away a patient's genus abundance is from the mean abundance of that genus in the healthy baseline.The formulae is shown in equation (1) in Figure 1 where X is the observed abundance, μ is the mean healthy abundance, and σ is the standard deviation of healthy abundances. Genera with a Z-score greater than 1.5 are considered to have a significant deviation and are ranked in descending order.

Fold Change measures the relative difference between the patient's genus abundance and the mean healthy abundance .The formula is shown in equation (2) in Figure 1 where X is the patient's abundance, μ is the mean healthy abundance, and a small pseudocount (1e-6) is added to avoid division by zero.Genera with a fold change > 1.5 (enriched) or < 0.67 (depleted) are considered dysbiotic and are also ranked.

The top two dysbiotic taxa are selected primarily based on their Z-score ranking. Fold change serves as a secondary ranking criterion, acting as a tiebreaker when Z-scores are similar. Taxa with zero abundances in both healthy samples and the patient's profile are disregarded.

2.4.4 Probiotic Recommendation

The Probiotic Recommendation System utilizes a CrewAI setup(https://docs.crewai.com/), which is a multi-AI agent framework designed to automate personalized and precise probiotic recommendations

This setup features two main agents:

The Microbiome Research Scientist agent conducts in-depth literature reviews, leveraging current scientific articles and clinical guidelines related to probiotics based on the provided disease condition and dysbiotic taxa.

The Clinical Probiotic Specialist agent analyzes the findings from the research scientist agent and recommends the most biologically appropriate and clinically applicable probiotic for the patient, providing an explanation of its mechanism of action, particularly on the identified dysbiotic taxa

The interaction between these agents is coordinated by a Crew object.The system requires three specific inputs from the user: the Disease name (e.g., Crohn's Disease), Dysbiotic Taxa 1 genus (e.g., Klebsiella), and Dysbiotic Taxa 2 genus (e.g., Salmonella). Both agents are equipped with TavilySearchTool, an internet-based API integrated into the Crew pipeline, enabling them to access recent peer-reviewed literature, probiotic case studies, and clinical guidelines. This internet access is crucial for ensuring the recommendations are precise, accurate, and clinically relevant based on the most current research. The agents' reasoning and recommendation generation are powered by Gemma 3 (Gemma team, 2025), a large language model that is installed locally. The prompts for these agents are meticulously engineered, defining their specific Role and Goal, providing a Backstory to contextualize their domain, and offering Task Instructions with a clear schema for their expected output. This structured approach ensures the generated recommendations are clear, consistent, and easily interpretable.

2.5 Tool Interface integration

We use the Streamlit interface to integrate the two components, the dysbiotic taxa prediction pipeline and the probiotic recommendation pipeline, to form a user-friendly and intuitive tool

that users can use to enter their gut metagenome and receive the dysbiosis analysis and the probiotic recommendation.

3. Results

3.1 XGBoost Model Results

3.1.1 RandomSearchCV results:

Hyperparameter	Value
subsample	**0.8**
reg_lambda	**1**
n_estimators	**200**
max_depth	**10**
learning_rate	**0.2**
gamma	**0.1**
colsample_bytree	**1.0**

Table 1: Best hyperparameter obtained through randomizedsearchcv

RandomSearchCV was used to tune the hyperparameters of the model. The best hyperparameters obtained for the model include a subsample of 0.8, meaning 80% of the training data is used per tree, and a reg_lambda of 1, which applies moderate regularization to balance bias and variance. The model uses 200 trees (n_estimators) to capture complex patterns, with a max_depth of 10 allowing deep trees, though regularization and subsampling mitigate overfitting risks. A learning rate of 0.2 ensures significant contributions from each tree, supported by regularization to prevent overfitting. A gamma of 0.1 ensures splits only occur with a minimum loss reduction, avoiding unnecessary complexity, and a colsample_bytree of 1 indicates all features are used, suitable when all are relevant for prediction.

3.1.2 XGBoost Model Results

The XGBoost model got an overall Accuracy of 86.56%, showing strong performance in classifying samples as CD, UC, or Healthy, as shown in Figure 2. The model excelled in identifying UC, with a Precision of 0.9082, a Recall of 0.8990, and F1 Score of 0.9036, suggesting high reliability for UC diagnosis.

For CD, the model showed a high recall of 93.98 %, ensuring most CD cases were correctly identified, though its precision of 84.32% indicates some false positives, as the F1 Score is 88.89 %. However, performance on the Healthy class was weaker, with a Recall of 71.96 % and an F1 Score of 078.57 %.The confusion matrix shown in Figure 3 shows strong UC performance, mostly satisfactory CD (minor misclassifications), but notable Healthy to CD misclassification, yielding Recall: 71.96%, F1: 78.57%, and Precision: 86.52% for Healthy.

Given in Figure 4, we can see the top 10 features contributing the most to the prediction. The immunosuppressant metadata feature contributes the most to prediction. Inflammatory diseases, like IBD, are autoimmune in nature . Immunosuppressants reduce the strength of the immune system, thereby reducing autoimmune symptoms as well. Patient response to immunosuppressants is also used by physicians as a way to diagnose IBD (Nakase et al., 2021). This features overcontribution can explain the misclassification of "Healthy" groups (Non-IBD) as CD.

3.2 Validation

For validation, the sample CSM9X233 (human gut metagenome) from the Inflammatory Bowel Disease Multi-omics Database (IBDMDB), part of the longitudinal multi-omics study "Human microbiome in Inflammatory Bowel Disease" was selected . The sample was sequenced using the Illumina HiSeq 2000 platform with whole-genome shotgun metagenomics (paired-end), yielding run SRR5950521 (~9.65 million reads, 626.7 MB compressed file size). The data were submitted by Harvard T.H. Chan School of Public Health and made publicly available on 2017-08-22.

The sample is taken from a Crohn's disease(CD) patient aged 45 years old with antibiotic and immunosuppressant usage.The fastq files (forward and reverse reads)are uploaded to the

tool.The metagenomic processing gives us a taxonomic profile of the patient. This prepared profile is then used for XGBoost prediction.

3.2.1 XGBoost model Output

We feed the input to the XGBoost model. The user is instructed to enter the metadata as we mentioned before. We enter the following details for metadata.The model prediction shown in Figure 5 shows the diagnosis as Crohn's Disease , which is a misdiagnosis of the single Ulcerative Colitis patient sample.Since the validation sample size is very small the tool still shows potential as seen in performance metrics (Accuracy: 86.56%); further validation is required.

3.2.2 Dysbiosis analysis Output

Using Z score and fold change, the top 2 dysbiotic taxa or genera are found to be Faecalibacterium (Gallazo et al., 2019) and Flavonifractor.Table 2 shows that Faecalibacterium and Flavonifractor pass the criteria and are the top two in rank for being dysbiotic, considering both z-score and fold change.

Genus	Z-score	Fold Change	Dysbiotic	Z-score Rank	Fold Change Rank
Faecalibacterium	3.107	4.05	True	1	4
Flavonifractor	1.63	5.13	True	2	1

Table 2: The table shows the Z- and fold change of the top 2 dysbiotic taxa. Faecalibacterium and Flavonifractor are two taxa or genera that cause.

3.2.3 Crew AI Probiotic Recommendation Output

The dysbiotic taxa and the predicted disease information to Crew AI to get the result.The recommendation by Crew AI setup is Faecalibacterium prausnitzii with the mechanistic reasoning on why it is recommended as shown Figure 6 . This recommendation is for assistive purposes only. To give user an insight into their health, and should not be taken as medical advice

4. Discussion

The methodology and results of this paper provide a robust framework for predicting inflammatory bowel disease (IBD) states and recommending probiotics using a multi-step pipeline that integrates metagenomic data processing, machine learning, dysbiosis analysis, and AI agent-driven recommendations. The pipeline successfully achieves its dual goals of disease prediction and probiotic recommendation, but there are several key aspects to discuss, including

the strengths of the approach, limitations, clinical implications, and avenues for future improvement.

The computational pipeline excels in processing shotgun metagenomic data by integrating KneadData(for host DNA removal via GRCh38) and MetaPhlAn (genus-level profiling using mpa_v31_CHOCOPhlAn_201901) with machine learning. It resolves taxonomic inconsistencies across datasets by collapsing to genus level and imputing zero-abundance for missing taxa, yielding a unified training set of 1,858 samples with 195 genera and four metadata features. This ensures robust, consistent input for the XGBoost model.

The XGBoost model, tuned via RandomSearchCV, achieves 86.56% accuracy, with strong performance in UC (Precision: 0.91, Recall: 0.90, F1: 0.90) and CD (Recall: 0.94).It effectively handles high-dimensional taxonomic data (195 genera) and complex interactions, with tuned hyperparameters(max_depth=10,learning_rate=0.2,subsample=0.8,reg_lambda=1,gamma=0.1) balancing complexity and generalization to prevent overfitting.High recall for CD and UC minimizes false negatives, critical for timely clinical intervention

The dysbiosis analysis, using Z-scores and fold changes, identifies the top dysbiotic genera, providing a data-driven basis for probiotic recommendation, although this method is not standard but it is conceptually sound enough. The Crew AI setup, powered by the Gemma 3 LLM and Tavily search, adds a layer of interpretability and clinical relevance by recommending probiotics with explanations of their mechanisms of action. This integration of machine learning with AI-driven recommendation systems is a significant step toward personalized probiotics, as it tailors interventions to the patient’s specific disease state and dysbiotic taxa.

Despite its strengths, the pipeline has notable limitations. The pipeline's main limitation is the misdiagnosis of single validation Ulcerative Colitis sample as Crohn's Disease despite predicting the IBD status correctly, this could possibly be due to overlapping of microbial profile in UC and CD.The limited validation sample size suggests further validation is needed. Poor Healthy class performance (Recall: 0.72, F1: 0.79) due to class imbalance (485 Healthy vs. 839 CD, 536 UC). This imbalance likely contributes to the misclassification of Healthy samples as CD., This biases the model toward over-predicting CD/UC to improve overall accuracy, especially influenced by the high-weighted "Immunosuppressants" feature, causing false positives in Healthy patients on such drugs for other inflammatory conditions. Clinically, this risks unnecessary testing, treatment, anxiety, and costs.

The Crew AI setup depends on scientific literature accessed via Tavily. If recent studies or clinical guidelines on probiotics for specific taxa (e.g., Faecalibacterium, Flavonifractor) are limited, the recommendations may lack depth or accuracy. Nevertheless the tool still has potential to improve upon these limitations with more high quality training data and validation samples.

5. Conclusion

The pipeline's ability to predict IBD with high sensitivity (Recall for CD: 0.9398, UC: 0.8990) is promising for clinical applications, as it minimizes the risk of missing diseased patients who need timely intervention. However, the misclassification of Healthy samples as CD or UC (low Recall for Healthy: 0.7196) poses a challenge for clinical adoption, as overdiagnosis could lead to unnecessary procedures. Addressing class imbalance through techniques like SMOTE oversampling or adjusting class weights in XGBoost (e.g., increasing the weight for Healthy samples) could improve performance on the Healthy class, reducing false positives.

The identification of dysbiotic taxa like Faecalibacterium and Flavonifractor aligns with existing literature, as Faecalibacterium (specifically Faecalibacterium prausnitzii) is often depleted in IBD patients and associated with anti-inflammatory properties . The recommendation of Faecalibacterium prausnitzii by Crew AI is plausible, as this probiotic is a key anti-inflammatory commensal bacterium . However, clinical validation through trials is necessary to confirm the efficacy of these recommendations, as the tool's outputs are assistive and not a substitute for medical advice.

Future improvements could include incorporating species- or strain-level taxonomic profiling to enhance the precision of dysbiosis analysis and probiotic recommendations. Expanding the Healthy baseline dataset to include diverse populations would improve the generalizability of the dysbiosis analysis. Additionally, integrating more metadata features (e.g., dietary habits, disease severity) could improve the XGBoost model's performance and provide a more holistic view of the patient's condition. Finally, validating the Crew AI recommendations against clinical outcomes in IBD patients would strengthen the tool's utility in personalized medicine.

Acknowledgements

This work was carried out at GenAI Research Labs, an initiative of Sivasakthi Science Foundation. The authors gratefully acknowledge the guidance and supervision of Dr. Sabarinath Subramaniam, PhD, whose expertise, insightful feedback, and continuous support were instrumental throughout the project.

Ethics Statement

This study utilized anonymized, publicly available datasets. No ethical approval was required.

Declaration of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Funding Interest

This research was supported by the Sivasakthi Science Foundation. The authors acknowledge the collaboration and infrastructure support provided by the Amrita School of Biotechnology.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Notebooklm and Perplexity in order to gain insight from/summarize research papers and finding references respectively. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

Alam, M. T., Amos, G. C. A., Murphy, A. R. J., Murch, S., Wellington, E. M. H., & Arasaradnam, R. P. (2020). Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels. *Gut Pathogens, 12*(1), Article 1. https://doi.org/10.1186/s13099-019-0341-6

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., & Segata, N. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife, 10*, Article e65088. https://doi.org/10.7554/eLife.65088

Belizário, J. E., & Faintuch, J. (2018). Microbiome and gut dysbiosis. In R. Silvestre & E. Torrado (Eds.), *Metabolic interaction in infection* (Vol. 109, pp. 459–476). Springer International Publishing. https://doi.org/10.1007/978-3-319-74932-7_13

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*(1), 281–305

Chang, C.-C., Liu, T.-C., Lu, C.-J., Chiu, H.-C., & Lin, W.-N. (2023). Machine learning strategy for identifying altered gut microbiomes for diagnostic screening in myasthenia gravis. *Frontiers in Microbiology, 14*, Article 1227300. https://doi.org/10.3389/fmicb.2023.1227300

CrewAI. (2025). *CrewAI documentation*. https://www.crewai.com/

Cross, E., Saunders, B., Farmer, A. D., & Prior, J. A. (2023). Diagnostic delay in adult inflammatory bowel disease: A systematic review. *Indian Journal of Gastroenterology, 42*(1), 40–52. https://doi.org/10.1007/s12664-022-01303-x

De Luca, F., & Shoenfeld, Y. (2018). The microbiome in autoimmune diseases. *Clinical and Experimental Immunology, 195*(1), 74–85. https://doi.org/10.1111/cei.13158

Derwa, Y., Gracie, D. J., Hamlin, P. J., & Ford, A. C. (2017). Systematic review with meta-analysis: The efficacy of probiotics in inflammatory bowel disease. *Alimentary Pharmacology & Therapeutics, 46*(4), 389–400. https://doi.org/10.1111/apt.14203

D'Urso, F., & Broccolo, F. (2024). Applications of artificial intelligence in microbiome analysis and probiotic interventions—An overview and perspective based on the current state of the art. *Applied Sciences, 14*(19), Article 8627. https://doi.org/10.3390/app14198627

Durack, J., & Lynch, S. V. (2019). The gut microbiome: Relationships with disease and opportunities for therapy. *Journal of Experimental Medicine, 216*(1), 20–40. https://doi.org/10.1084/jem.20180448

Estevinho, M. M., Yuan, Y., Rodríguez-Lago, I., Sousa-Pimenta, M., Dias, C. C., Barreiro-de Acosta, M., Jairath, V., & Magro, F. (2024). Efficacy and safety of probiotics in IBD: An overview of systematic reviews and updated meta-analysis of randomized controlled trials. *United European Gastroenterology Journal, 12*(7), 960–981. https://doi.org/10.1002/ueg2.12636

Fasano, A. (2012). Leaky gut and autoimmune diseases. *Clinical Reviews in Allergy & Immunology, 42*(1), 71–78. https://doi.org/10.1007/s12016-011-8291-x

Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., Vatanen, T., Hall, A. B., Mallick, H., McIver, L. J., Sauk, J. S., Wilson, R. G., Stevens, B. W., Scott, J. M., Pierce, K., Deik, A. A., Bullock, K., Imhann, F., Porter, J. A., ... Xavier, R. J. (2018). Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology, 4*(2), 293–305. https://doi.org/10.1038/s41564-018-0306-4

Galazzo, G. L., Pierik, M. J., Jonkers, D. M. A. E., & Penders, J. (2019). Faecal microbiota dynamics and their relation to disease course in Crohn's disease. *Journal of Crohn's and Colitis, 13*(10), 1273–1282. https://doi.org/10.1093/ecco-jcc/jjz049

Gemma Team. (2025). *Gemma 3 technical report*. Google DeepMind. https://arxiv.org/abs/2503.19786

Guo, N., & Lv, L. (2023). Mechanistic insights into the role of probiotics in modulating immune cells in ulcerative colitis. *Immunity, Inflammation and Disease, 11*(10), Article e1045. https://doi.org/10.1002/iid3.1045

Han, L., Liu, X., Lan, Y., Hua, Y., Fan, Z., & Li, Y. (2024). Metagenomic analysis demonstrates distinct changes in the gut microbiome of Kawasaki diseases children. *Frontiers in Immunology, 15*, Article 1416185. https://doi.org/10.3389/fimmu.2024.1416185

Javaid, A., Shahab, O., Adorno, W., Fernandes, P., May, E., & Syed, S. (2022). Machine learning predictive outcomes modeling in inflammatory bowel diseases. *Inflammatory Bowel Diseases, 28*(6), 819–829. https://doi.org/10.1093/ibd/izab187

Kim, H. S., Oh, S. J., Kim, B. K., Kim, J. E., Kim, B.-H., Park, Y.-K., Yang, B.-G., Lee, J.-Y., Bae, J.-W., & Lee, C. K. (2024). Dysbiotic signatures and diagnostic potential of gut microbial markers for inflammatory bowel disease in Korean population. *Scientific Reports, 14*(1), Article 23701. https://doi.org/10.1038/s41598-024-74002-6

Lin, H., & Peddada, S. D. (2020). Analysis of microbial compositions: A review of normalization and differential abundance analysis. *npj Biofilms and Microbiomes*, *6*(1), Article 1. https://doi.org/10.1038/s41522-020-00160-w

Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2019). *The Inflammatory Bowel Disease Multi'omics Database (IBDMDDB)* [Data set]. https://ibdmdb.org/

Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., & Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in Biology and Medicine*, *127*, Article 103761. https://doi.org/10.1016/j.compbiomed.2020.103761

McIver, L. J., Abu-Ali, G., Franzosa, E. A., Schwager, R., Morgan, X. C., Waldron, L., Segata, N., & Huttenhower, C. (2018). bioBakery: A meta'omic analysis environment. *Bioinformatics*, *34*(7), 1235–1237. https://doi.org/10.1093/bioinformatics/btx754

Nakase, H., Uchino, M., Shinzaki, S., Matsuura, M., Matsuoka, K., Kobayashi, T., Saruta, M., Hirai, F., Hata, K., Hiraoka, S., Esaki, M., Sugimoto, K., Fuji, T., Watanabe, K., Nakamura, S., Inoue, N., Itoh, T., Naganuma, M., Hisamatsu, T., ... Koike, K. (2021). Evidence-based clinical practice guidelines for inflammatory bowel disease 2020. *Journal of Gastroenterology*, *56*(6), 489–526. https://doi.org/10.1007/s00535-021-01784-1

Paray, B. A., Albeshr, M. F., Jan, A. T., & Rather, I. A. (2020). Leaky gut and autoimmunity: An intricate balance in individuals health and the diseased state. *International Journal of Molecular Sciences*, *21*(24), Article 9770. https://doi.org/10.3390/ijms21249770

Sadeghpour Heravi, F. (2024). Gut microbiota and autoimmune diseases: Mechanisms, treatment, challenges, and future recommendations. *Current Clinical Microbiology Reports*, *11*(1), 18–33. https://doi.org/10.1007/s40588-023-00213-6

Shan, Y., Lee, M., & Chang, E. B. (2022). The gut microbiome and inflammatory bowel diseases. *Annual Review of Medicine*, *73*(1), 455–468. https://doi.org/10.1146/annurev-med-042320-021020

Vavricka, S. R., Spigaglia, S. M., Rogler, G., Pittet, V., Michetti, P., Felley, C., Mottet, C., Braegger, C. P., Rogler, D., Straumann, A., Bauerfeind, P., Fried, M., & Schoepfer, A. M. (2012). Systematic evaluation of risk factors for diagnostic delay in inflammatory bowel disease. *Inflammatory Bowel Diseases*, *18*(3), 496–505. https://doi.org/10.1002/ibd.21719

Westfall, S., Carracci, F., Estill, M., Zhao, D., Wu, Q., Shen, L., Simon, J., & Pasinetti, G. M. (2021). Optimization of probiotic therapeutics using machine learning in an artificial human gastrointestinal tract. *Scientific Reports*, *11*(1), Article 1067. https://doi.org/10.1038/s41598-020-79947-y

Figure Captions

Figure 1: Equations for calculating dysbiosis at the genus level.

Equation (1) Z-score - It quantifies how many standard deviations the genus abundance in a patient sample (X) deviates from the mean abundance of the same genus in healthy samples (μ), normalized by the standard deviation (σ).

Equation (2) Fold change: It expresses the ratio of the patient's genus abundance (X) to the mean abundance of the same genus in healthy samples (μ). A small pseudocount (10^{-6}) is added to the denominator to prevent division by zero when μ = 0.

Figure 2: The performance metrics of all three class labels, that is, UC, CD, and Healthy, and on the extreme right is the overall accuracy of the model.

Figure 3: The confusion matrix, which shows us where the model made misclassifications.We model, most of the misclassification happens when Healthy samples are predicted to have Crohn's disease. This is in accordance with what we saw earlier with the healthy class having a Recall of 71.96 % and an F1 Score of 078.57 %.

Figure 4: Immunosuppressant is shown as the most contributing or important feature; a suggested explanation for this could be that patients with chronic inflammatory conditions, even without proper diagnosis, are put on immunosuppressants and continue if responding well, thus people who continually take immunosuppressants are more likely to have an inflammatory condition like IBD.

Figure 5: The mode prediction shows Crohn’s Disease as the diagnosis.Misdiagnosis (CD vs UC) likely due to microbial profile overlap.Although IBD status, as opposed to Healthy status, is correctly predicted.

Figure 6: Faecalibacterium prausnitzii is recommended as a probiotic; its effect and mechanism of action on Crohn’s disease, and how it helps the dysbiotic taxa, are also mentioned in the recommendation

Abstract

Background and Objective: The dysbiosis of human gut microbiome has been increasingly seen to have a relation in the development of autoimmune diseases, with specific microbial signatures having causative association with specific conditions. Inflammatory bowel disease (IBD) is one such autoimmune ailment. This paper proposes a predictive tool that can identify the IBD status of an individual based on the composition of the gut microbiome using machine learning and AI agents driven techniques. The technology can strengthen the suspicion of a potential IBD diagnosis a patient may have based on their gut microbiome profile. **Methods:** The tool processes patient gut metagenome using integrated Kneaddata and MetaPhlAn to generate taxonomic profiles. These are fed into an XGBoost classifier to predict IBD or healthy status. Dysbiotic taxa are identified via Z-score and fold change. CrewAI delivers personalized probiotic recommendations based on diagnosis and dysbiosis.

Results: The tuned XGBoost model achieved 86.6% accuracy. On validation using single ulcerative colitis sample, the tool correctly predicted IBD status but misclassified it as Crohn's disease(possibly due to overlapping microbial signatures), identifying Faecalibacterium and Flavonifractor as dysbiotic taxa.The probiotic recommended was Faecalibacterium prausnitzii ,backed with reasoning basedon scientific literature.

Conclusions: Despite limited validation sample size, the high accuracy , correct IBD detection ,dysbiosis analysis and elaborate probiotic recommendation suggest promising potential; further validation needed

Keywords: Metagenomics, Inflammatory Bowel Disease, Machine Learning, AI Agents, Gut Dysbiosis, Probiotics

1. Introduction

The human body consists of a vast community of trillions of microbes, collectively known as the human microbiome, with the gastrointestinal tract, particularly the colon, being the most densely populated habitat collectively called the human gut microbiome. This intricate community, composed of over 1000 bacterial species, maintains a crucial symbiotic relationship with the host, influencing vital processes such as digestion, vitamin production, and immune system modulation [1].While the taxonomic profiles of individual microbiomes vary significantly due to factors like geography, diet, and host genetics, their underlying functional potential via the metabolic pathways shows remarkable consistency across populations. However, any imbalance in the composition and function of this microbiome, which is consistent, is referred to as dysbiosis. Dysbiosis is strongly linked to the development and progression of chronic conditions, including metabolic, autoimmune, and neurological diseases. Currently, it is increasingly recognized as a key environmental factor that can shift the immune response towards pro-inflammatory pathways, thereby instigating autoimmune processes[2].

IBDs, encompassing Crohn's disease (CD) and Ulcerative Colitis (UC), are characterized by chronic gastrointestinal inflammation and exhibit altered gut microbiome compositions, including reduced bacterial diversity, a decrease in beneficial bacteria (like *Faecalibacterium prausnitzii*) , [3]and increase in potentially harmful or opportunistic taxa (like Proteobacteria, *Escherichia-Shigella*, and *Enterococcus*)[4].

The mechanisms by which gut dysbiosis contributes to autoimmune diseases include the loss of intestinal barrier integrity, which is popularly referred to as “Leaky Gut,” and has been observed in various autoimmune diseases. Tight junctions, which include proteins like occludin, are

crucial for maintaining the barrier function. Disruption of the barrier allows microbial products and

bacteria to leak into the host tissues and bloodstream, potentially causing the immune system to respond [5,6]. This allows microbial products to trigger systemic inflammation and molecular mimicry, where microbial antigens structurally similar to some host self-antigens result in the onset of an immune reaction against the self-antigens. Also, the modulation of the immune system by microbial metabolites like short-chain fatty acids (SCFAs) can lead to an increase in systemic inflammation[7].

According to current research, diagnostic delays are frequent; on average, they occur between 2 months and 8 years after the onset of symptoms [8].The diagnosis delay for UC is usually 4 months, while for CD, it is roughly 9 months. Children and mild or vague presentation cases are more likely to have delays [9].Traditional IBD diagnostic methods face challenges due to symptom variability and invasiveness, which can hinder an accurate and timely diagnosis. The validity and reproducibility of evaluations are affected by the significant variability between observers that hinders endoscopic scoring systems as a whole, such as the Mayo score, which is utilized for assessing the progression of the disease[10].

In response, machine learning (ML), combined with high-throughput sequencing technologies like shotgun metagenomics, has emerged as a powerful computational approach for analyzing complex microbiome data, enabling the accurate classification of IBD patients and the identification of microbial biomarkers for disease prediction [11].

Improvements in metagenomics, metatranscriptomics, and metabolomics have become essential in deciphering the intricate interactions between the gut microbiome and IBD. Integration of metagenomics data with host genomic data and clinical metadata is crucial to deciphering the intricate host-microbe interactions underlying IBD and designing microbiome-based diagnostics, prognostics, and therapeutic interventions.

Probiotics are live microorganisms that, when administered adequately, confer health benefits through various mechanisms, including direct antimicrobial activity, strengthening the intestinal barrier by enhancing tight junctions, and modulating the host immune system to reduce inflammation[12,13]. While evidence for their effectiveness in Crohn's disease is often limited, substantial evidence supports use of specific probiotic strains, such as *Escherichia coli Nissle 1917* and the multi-strain preparation VSL#3, for inducing and maintaining remission in patients with mild to moderate Ulcerative Colitis [14]. While generally considered safe one must use only specific probiotic strains which are documented to have beneficial effects on mild or moderate inflammatory diseases, as wrong recommendations could result in side effects in immunocompromised patients. Crucially, the efficacy of probiotics is strain-specific and dose-dependent, hence the variability in research findings [15].This highlights the need for personalized insights on what probiotic strains could be useful for a particular individual [16].This paper presents a Metagenomic-ML-based tool that predicts IBD via gut microbiome profile and uses CrewAI to deliver personalized probiotics recommendation.

2. Methodology

2.1 Study Design

The tool aims to fulfill the following purpose: predicting user's disease state and, if diseased, recommending suitable probiotic. The tool is structured into four distinct parts: generating a user's taxonomic profile, predicting the disease using a machine learning model that is trained on the taxonomic profiles of healthy individuals and individuals with IBD, identifying the dysbiotic taxa, and thereby providing personalized probiotic recommendations using CrewAI, if a disease state exists.

2.2 Data Collection

2.2.1 Dataset for IBD prediction model

The data utilized for this project were derived from shotgun metagenomic sequencing, which provides detailed taxonomic profiles of microbial communities [17].Two primary sources were integrated: Dataset 1 was retrieved from the Inflammatory Bowel Disease Multi'omics Database (Dataset 1), part of the Integrative Human Microbiome Project (HMP2), contributing 1638 samples with taxonomic classification up to the strain level. Dataset 2 originated from the results of the paper "Gut microbiome structure and metabolic activity in inflammatory bowel disease" [18], comprising 220 stool samples from two cohorts (155 from Massachusetts, 65 from the Netherlands). These samples were from individuals with Crohn's disease (CD), Ulcerative colitis (UC), and non-IBD controls, with taxonomic classification up to the species level (Dataset 2). Both the dataset included metadata features Age, Immunosuppressant usage, Antibiotic usage, and Diagnosis.

2.2.2 Dataset for metagenomic tools

Tools like Knead Data and MetaPhlAn2 were used to process the users' metagenome, with KneadData using the GRCh38 human genome to remove host DNA and MetaPhlAn using the mpa_v31_CHOCOPhlAn_201901 database for taxonomic profiling based on over 17,000 microbial genomes. Both these tools were incorporated into our pipeline.

2.3 Data Preprocessing

A significant challenge in integrating these datasets was their inconsistency, specifically differing levels of taxonomic classification. To address this, a preprocessing approach was employed, involving two key steps: First, collapsing the organisms to a higher taxonomic level, specifically the genus level, by summing the relative abundances of species or strains within the same genus . This increased feature commonality between datasets. Second, the datasets were merged, ensuring that all features (organisms and metadata) from both datasets were retained, with uncommon taxa assigned a relative abundance of zero to maintain sum totals. The resulting merged dataset contained 1858 samples and a taxonomic profile of 195 genera, representing 485 healthy, 838 CD, and 535 UC samples. Other preprocessing steps included relative abundance normalization (renormalizing genus-level abundances to sum to 1), metadata encoding (e.g., Diagnosis: Healthy=0, CD=1, UC=2; Immunosuppressant/Antibiotic Usage: Yes=1, No=0), and min-max normalization of Age to a range between 0 and 1.

2.4 Tool Development and Evaluation

2.4.1 User Taxonomic Profile Generation

The first part of the methodology focuses on User Taxonomic Profile Generation. This begins with processing raw paired-end FASTQ files from the user's gut metagenome sample. Kneaddata (v0.12.0) is employed to meticulously remove host-contaminant DNA, specifically human DNA, and low-quality reads . This step involves several integrated tools: Bowtie2 (v2.4.1) is used for alignment against the GRCh38 human genome (a pre-indexed database), and Trimmomatic (v0.39) handles adapter trimming and quality filtering [19].Following host DNA removal, these processed paired-end reads are fed into MetaPhlAn3 (v3.0.14) to generate the crucial taxonomic profile. MetaPhlAn3 utilizes the mpa_v31_CHOCOPhlAn_201901 database for its profiling, specifically designed to output genus-level abundances by using the --tax_lev g flag [20]. The relative abundances of these genera are then normalized to sum to 1.0 per sample. Unclassified genera are removed, and the profile undergoes feature alignment for consistency with training dataset, making it a valid input for prediction [21].

2.4.2 Disease Prediction Using XGBoost

The second core component is Disease Prediction, which leverages an XGBoost machine learning model. XGBoost was strategically chosen due to its robust capabilities in handling high-dimensional data typical of gut microbiome profiles (such as 195 genera and 4 clinical metadata features) [22]. Its effectiveness stems from its ability to model the complex and nonlinear patterns and interactions among genera that determine an individual's IBD status [23].

The model's hyperparameters are carefully tuned to ensure optimal performance. To find the best combination of hyperparameters, RandomSearchCV is employed. This method randomly

samples 20 combinations from the defined parameter space, evaluating each using 5-fold cross-validation with accuracy as the scoring metric. This approach efficiently identifies optimal settings, striking a balance between model complexity and generalization capabilities, thereby enhancing prediction accuracy [24].

The XGBoost estimator is configured to use multi-class logarithmic loss to rank feature importance. Recursive feature elimination is used to iteratively remove the lowest-ranked features until a predefined number of 180 genera are selected. This feature selection is performed exclusively on the training dataset, which is an 80/20 split of the main dataset, to prevent information leakage into the test set. The selected genera are then consistently applied to the test dataset. The XGBoost model itself is optimized for multi-class logarithmic loss, which is ideal for the CD/UC/Healthy classification task.

2.4.3 Dysbiotic Taxa Prediction

The third critical component is the dysbiotic taxa identification.This tool identifies the top 2 dysbiotic taxa that are most altered compared to a healthy gut microbiome and are likely contributing to the inflammatory environment [25]. To achieve this, a healthy cohort baseline dataset is established by filtering all healthy samples from the same dataset used to train the XGBoost model and then taking the average of the relative abundances of each genus present in each healthy sample, thereby creating a baseline healthy taxonomic profile. This baseline is then used for comparison against the patient’s taxonomic profile.

Two statistical techniques are used for Dysbiosis Quantification: Z-score and Fold Change

Z-score measures how many standard deviations away a patient's genus abundance is from the mean abundance of that genus in the healthy baseline.The formulae is shown in equation (1) in

Figure 1 where X is the observed abundance, μ is the mean healthy abundance, and σ is the standard deviation of healthy abundances. Genera with a Z-score greater than 1.5 are considered to have a significant deviation and are ranked in descending order.

Fold Change measures the relative difference between the patient's genus abundance and the mean healthy abundance .The formula is shown in equation (2) in Figure 1 where X is the patient's abundance, μ is the mean healthy abundance, and a small pseudocount (1e-6) is added to avoid division by zero.Genera with a fold change > 1.5 (enriched) or < 0.67 (depleted) are considered dysbiotic and are also ranked.The top two dysbiotic taxa are selected primarily based on their Z-score ranking. Fold change serves as a secondary ranking criterion, acting as a tiebreaker when Z-scores are similar. Taxa with zero abundances in both healthy samples and the patient's profile are disregarded.

2.4.4 Probiotic Recommendation

The Probiotic Recommendation System utilizes a CrewAI setup [26], which is a multi-AI agent framework designed to automate personalized and precise probiotic recommendations

This setup features two main agents:

The Microbiome Research Scientist agent conducts in-depth literature reviews, leveraging current scientific articles and clinical guidelines related to probiotics based on the provided disease condition and dysbiotic taxa.

The Clinical Probiotic Specialist agent analyzes the findings from the research scientist agent and recommends the most biologically appropriate and clinically applicable probiotic for the patient, providing an explanation of its mechanism of action, particularly on the identified dysbiotic taxa

The interaction between these agents is coordinated by a Crew object.The system requires three specific inputs from the user: the Disease name (e.g., Crohn's Disease), Dysbiotic Taxa 1 genus (e.g., Klebsiella), and Dysbiotic Taxa 2 genus (e.g., Salmonella). Both agents are equipped with TavilySearchTool, an internet-based API integrated into the Crew pipeline, enabling them to access recent peer-reviewed literature, probiotic case studies, and clinical guidelines. This internet access is crucial for ensuring the recommendations are precise, accurate, and clinically relevant based on the most current research. The agents' reasoning and recommendation generation are powered by Gemma 3 [27], a large language model that is installed locally. The prompts for these agents are meticulously engineered, defining their specific Role and Goal, providing a Backstory to contextualize their domain, and offering Task Instructions with a clear schema for their expected output. This structured approach ensures the generated recommendations are clear, consistent, and easily interpretable.

2.5 Tool Interface integration

We use the Streamlit interface to integrate the two components, the dysbiotic taxa prediction pipeline and the probiotic recommendation pipeline, to form a user-friendly and intuitive tool that users can use to enter their gut metagenome and receive the dysbiosis analysis and the probiotic recommendation.

3. Results

3.1 XGBoost Model Results

3.1.1 RandomSearchCV results:

Hyperparameter	Value
subsample	**0.8**
reg_lambda	**1**
n_estimators	**200**
max_depth	**10**
learning_rate	**0.2**
gamma	**0.1**
colsample_bytree	**1.0**

Table 1: Best hyperparameter obtained through randomizedsearchcv

RandomSearchCV was used to tune the hyperparameters of the model. The best hyperparameters obtained for the model include a subsample of 0.8, meaning 80% of the training data is used per tree, and a reg_lambda of 1, which applies moderate regularization to balance bias and variance. The model uses 200 trees (n_estimators) to capture complex patterns, with a max_depth of 10 allowing deep trees, though regularization and subsampling mitigate overfitting risks. A learning rate of 0.2 ensures significant contributions from each tree, supported by regularization to prevent overfitting. A gamma of 0.1 ensures splits only occur with a minimum loss reduction, avoiding unnecessary complexity, and a colsample_bytree of 1 indicates all features are used, suitable when all are relevant for prediction.

3.1.2 XGBoost Model Results

The XGBoost model got an overall Accuracy of 86.56%, showing strong performance in classifying samples as CD, UC, or Healthy, as shown in Figure 2. The model excelled in identifying UC, with a Precision of 0.9082, a Recall of 0.8990, and F1 Score of 0.9036, suggesting high reliability for UC diagnosis.

For CD, the model showed a high recall of 93.98 %, ensuring most CD cases were correctly identified, though its precision of 84.32% indicates some false positives, as the F1 Score is 88.89 %. However, performance on the Healthy class was weaker, with a Recall of 71.96 % and an F1 Score of 078.57 %.The confusion matrix shown in Figure 3 shows strong UC performance, mostly satisfactory CD (minor misclassifications), but notable Healthy to CD misclassification, yielding Recall: 71.96%, F1: 78.57%, and Precision: 86.52% for Healthy.

Given in Figure 4, we can see the top 10 features contributing the most to the prediction. The immunosuppressant metadata feature contributes the most to prediction. Inflammatory diseases, like IBD, are autoimmune in nature . Immunosuppressants reduce the strength of the immune system, thereby reducing autoimmune symptoms as well. Patient response to immunosuppressants is also used by physicians as a way to diagnose IBD [28]. This features overcontribution can explain the misclassification of “Healthy” groups (Non-IBD) as CD.

3.2 Validation

For validation, the sample CSM9X233 (human gut metagenome) from the Inflammatory Bowel Disease Multi-omics Database (IBDMDB), part of the longitudinal multi-omics study "Human microbiome in Inflammatory Bowel Disease" was selected . The sample was sequenced using the Illumina HiSeq 2000 platform with whole-genome shotgun metagenomics (paired-end), yielding run SRR5950521 (~9.65 million reads, 626.7 MB compressed file size). The data were submitted by Harvard T.H. Chan School of Public Health and made publicly available on 2017-08-22.

The sample is taken from a Crohn’s disease(CD) patient aged 45 years old with antibiotic and immunosuppressant usage.The fastq files (forward and reverse reads)are uploaded to the tool.The metagenomic processing gives us a taxonomic profile of the patient. This prepared profile is then used for XGBoost prediction.

3.2.1 XGBoost model Output

We feed the input to the XGBoost model. The user is instructed to enter the metadata as we mentioned before. We enter the following details for metadata.The model prediction shown in Figure 5 shows the diagnosis as Crohn’s Disease , which is a misdiagnosis of the single Ulcerative Colitis patient sample.Since the validation sample size is very small the tool still shows potential as seen in performance metrics (Accuracy: 86.56%); further validation is required.

3.2.2 Dysbiosis analysis Output

Using Z score and fold change, the top 2 dysbiotic taxa or genera are found to be Faecalibacterium and Flavonifractor [29].Table 2 shows that Faecalibacterium and Flavonifractor pass the criteria and are the top two in rank for being dysbiotic, considering both z-score and fold change.

Genus	Z-score	Fold Change	Dysbiotic	Z-score Rank	Fold Change Rank
Faecalibacterium	3.107	4.05	True	1	4
Flavonifractor	1.63	5.13	True	2	1

Table 2: The table shows the Z- and fold change of the top 2 dysbiotic taxa. Faecalibacterium and Flavonifractor are two taxa or genera that cause.

3.2.3 Crew AI Probiotic Recommendation Output

The dysbiotic taxa and the predicted disease information is given to Crew AI to get the result.The recommendation by Crew AI setup is Faecalibacterium prausnitzii with the mechanistic reasoning on why it is recommended based on its action on the gut as shown Figure 6.This recommendation is for assistive purposes only. To give user an insight into their health, and should not be taken as medical advice.

4. Discussion

The methodology and results of this paper provide a robust framework for predicting inflammatory bowel disease (IBD) states and recommending probiotics using a multi-step pipeline that integrates metagenomic data processing, machine learning, dysbiosis analysis, and AI agent-driven recommendations. The pipeline successfully achieves its dual goals of disease prediction and probiotic recommendation, but there are several key aspects to discuss, including the strengths of the approach, limitations, clinical implications, and avenues for future improvement.

The computational pipeline excels in processing shotgun metagenomic data by integrating KneadData(for host DNA removal via GRCh38) and MetaPhlAn (genus-level profiling using mpa_v31_CHOCOPhlAn_201901) with machine learning. It resolves taxonomic inconsistencies across datasets by collapsing to genus level and imputing zero-abundance for missing taxa, yielding a unified training set of 1,858 samples with 195 genera and four metadata features. This ensures robust, consistent input for the XGBoost model.

The XGBoost model, tuned via RandomSearchCV, achieves 86.56% accuracy, with strong performance in UC (Precision: 0.91, Recall: 0.90, F1: 0.90) and CD (Recall: 0.94).It effectively handles high-dimensional taxonomic data (195 genera) and complex interactions, with tuned hyperparameters(max_depth=10,learning_rate=0.2,subsample=0.8,reg_lambda=1,gamma=0.1) balancing complexity and generalization to prevent overfitting.High recall for CD and UC minimizes false negatives, critical for timely clinical intervention

The dysbiosis analysis, using Z-scores and fold changes, identifies the top dysbiotic genera, providing a data-driven basis for probiotic recommendation, although this method is not standard

but it is conceptually sound enough. The Crew AI setup, powered by the Gemma 3 LLM and Tavily search, adds a layer of interpretability and clinical relevance by recommending probiotics with explanations of their mechanisms of action. This integration of machine learning with AI-driven recommendation systems is a significant step toward personalized probiotics, as it tailors interventions to the patient's specific disease state and dysbiotic taxa.

Despite its strengths, the pipeline has notable limitations. The pipeline's main limitation is the misdiagnosis of single validation Ulcerative Colitis sample as Crohn's Disease despite predicting the IBD status correctly, this could possibly be due to overlapping of microbial profile in UC and CD.The limited validation sample size suggests further validation is needed. Poor Healthy class performance (Recall: 0.72, F1: 0.79) due to class imbalance (485 Healthy vs. 839 CD, 536 UC). This imbalance likely contributes to the misclassification of Healthy samples as CD., This biases the model toward over-predicting CD/UC to improve overall accuracy, especially influenced by the high-weighted "Immunosuppressants" feature, causing false positives in Healthy patients on such drugs for other inflammatory conditions. Clinically, this risks unnecessary testing, treatment, anxiety, and costs.

The Crew AI setup depends on scientific literature accessed via Tavily. If recent studies or clinical guidelines on probiotics for specific taxa (e.g., Faecalibacterium, Flavonifractor) are limited, the recommendations may lack depth or accuracy. Nevertheless the tool still has potential to improve upon these limitations with more high quality training data and validation samples.

5. Conclusion

The pipeline's ability to predict IBD with high sensitivity (Recall for CD: 0.9398, UC: 0.8990) is promising for clinical applications, as it minimizes the risk of missing diseased patients who need timely intervention. However, the misclassification of Healthy samples as CD or UC (low Recall for Healthy: 0.7196) poses a challenge for clinical adoption, as overdiagnosis could lead to unnecessary procedures. Addressing class imbalance through techniques like SMOTE oversampling or adjusting class weights in XGBoost (e.g., increasing the weight for Healthy samples) could improve performance on the Healthy class, reducing false positives.

The identification of dysbiotic taxa like Faecalibacterium and Flavonifractor aligns with existing literature, as Faecalibacterium (specifically Faecalibacterium prausnitzii) is often depleted in IBD patients and associated with anti-inflammatory properties . The recommendation of Faecalibacterium prausnitzii by Crew AI is plausible, as this probiotic is a key anti-inflammatory commensal bacterium . However, clinical validation through trials is necessary to confirm the efficacy of these recommendations, as the tool's outputs are assistive and not a substitute for medical advice.

Future improvements could include incorporating species- or strain-level taxonomic profiling to enhance the precision of dysbiosis analysis and probiotic recommendations. Expanding the Healthy baseline dataset to include diverse populations would improve the generalizability of the dysbiosis analysis. Additionally, integrating more metadata features (e.g., dietary habits, disease severity) could improve the XGBoost model's performance and provide a more holistic view of the patient's condition. Finally, validating the Crew AI recommendations against clinical outcomes in IBD patients would strengthen the tool's utility in personalized medicine.

Acknowledgements

This work was carried out at GenAI Research Labs, an initiative of Sivasakthi Science Foundation. The authors gratefully acknowledge the guidance and supervision of Dr. Sabarinath Subramaniam, PhD, whose expertise, insightful feedback, and continuous support were instrumental throughout the project.

Ethics Statement

This study utilized anonymized, publicly available datasets. No ethical approval was required.

Declaration of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Funding Interest

This research was supported by the Sivasakthi Science Foundation. The authors acknowledge the collaboration and infrastructure support provided by the Amrita School of Biotechnology.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Notebooklm and Perplexity in order to gain insight from/summarize research papers and finding references respectively. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

1. Durack J, Lynch SV: The gut microbiome: Relationships with disease and opportunities for therapy. J Exp Med. 2019, 216:20–40. 10.1084/JEM.20180448

2. De Luca F, Shoenfeld Y: The microbiome in autoimmune diseases. Clin Exp Immunol. 2018, 195:74–85. 10.1111/CEI.13158

3. Belizário JE, Faintuch J: Microbiome and Gut Dysbiosis. Exp Suppl 2012. 2018, 109:459–76. 10.1007/978-3-319-74932-7_13

4. Alam MT, Amos GCA, Murphy ARJ, Murch S, Wellington EMH, Arasaradnam RP: Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels. Gut Pathog. 2020, 12:1. 10.1186/S13099-019-0341-6/FIGURES/3

5. Fasano A: Leaky gut and autoimmune diseases. Clin Rev Allergy Immunol. 2012, 42:71–8. 10.1007/S12016-011-8291-X/METRICS

6. Paray BA, Albeshr MF, Jan AT, Rather IA: Leaky Gut and Autoimmunity: An Intricate Balance in Individuals Health and the Diseased State. Int J Mol Sci 2020 Vol 21 Page 9770. 2020, 21:9770–9770. 10.3390/IJMS21249770

7. Sadeghpour Heravi F: Gut Microbiota and Autoimmune Diseases: Mechanisms, Treatment, Challenges, and Future Recommendations. Curr Clin Microbiol Rep. 2024, 11:18–33. 10.1007/S40588-023-00213-6/TABLES/2

8. Cross E, Saunders B, Farmer AD, Prior JA: Diagnostic delay in adult inflammatory bowel disease: A systematic review. Indian J Gastroenterol. 2023, 42:40–52. 10.1007/S12664-022-01303-X/TABLES/4

9. Vavricka SR, Spigaglia SM, Rogler G, et al.: Systematic Evaluation of Risk Factors for Diagnostic Delay in Inflammatory Bowel Disease. Inflamm Bowel Dis. 2012, 18:496–505. 10.1002/IBD.21719

10. Javaid A, Shahab O, Adorno W, Fernandes P, May E, Syed S: Machine Learning Predictive Outcomes Modeling in Inflammatory Bowel Diseases. Inflamm Bowel Dis. 2022, 28:819–29. 10.1093/IBD/IZAB187

11. D'Urso F, Broccolo F: Applications of Artificial Intelligence in Microbiome Analysis and Probiotic Interventions—An Overview and Perspective Based on the Current State of the Art. Appl Sci 2024 Vol 14 Page 8627. 2024, 14:8627–8627. 10.3390/APP14198627

12. Shan Y, Lee M, Chang EB: The Gut Microbiome and Inflammatory Bowel Diseases. Annu Rev Med. 2022, 73:455–68. 10.1146/ANNUREV-MED-042320-021020/CITE/REFWORKS

13. Guo N, Lv L lu: Mechanistic insights into the role of probiotics in modulating immune cells in ulcerative colitis. Immun Inflamm Dis. 2023, 11:e1045–e1045. 10.1002/IID3.1045

14. Derwa Y, Gracie DJ, Hamlin PJ, Ford AC: Systematic review with meta-analysis: the efficacy of probiotics in inflammatory bowel disease. Aliment Pharmacol Ther. 2017, 46:389–400. 10.1111/APT.14203

15. Estevinho MM, Yuan Y, Rodríguez-Lago I, et al.: Efficacy and safety of probiotics in IBD: An overview of systematic reviews and updated meta-analysis of randomized controlled trials. United Eur Gastroenterol J. 2024, 12:960–81. 10.1002/UEG2.12636

16. Westfall S, Carracci F, Estill M, et al.: Optimization of probiotic therapeutics using machine learning in an artificial human gastrointestinal tract. Sci Rep 2021 111. 2021, 11:1067. 10.1038/s41598-020-79947-y

17. McIver LJ, Abu-Ali G, Franzosa EA, et al.: bioBakery: a meta'omic analysis environment. Bioinformatics. 2018, 34:1235–7. 10.1093/BIOINFORMATICS/BTX754

18. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, et al.: Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol 2018 42. 2018, 4:293–305. 10.1038/s41564-018-0306-4

19. Han L, Liu X, Lan Y, Hua Y, Fan Z, Li Y: Metagenomic analysis demonstrates distinct changes in the gut microbiome of Kawasaki diseases children. Front Immunol. 2024, 15:1416185–1416185. 10.3389/FIMMU.2024.1416185/TEXT

20. Beghini F, McIver LJ, Blanco-Míguez A, et al.: Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. eLife. 2021, 10:. 10.7554/ELIFE.65088

21. Lin H, Peddada SD: Analysis of microbial compositions: a review of normalization and differential abundance analysis. Npj Biofilms Microbiomes 2020 61. 2020, 6:60. 10.1038/s41522-020-00160-w

22. Chang CC, Liu TC, Lu CJ, Chiu HC, Lin WN: Machine learning strategy for identifying altered gut microbiomes for diagnostic screening in myasthenia gravis. Front Microbiol. 2023, 14:1227300–1227300. 10.3389/FMICB.2023.1227300/BIBTEX

23. Ma B, Meng F, Yan G, Yan H, Chai B, Song F: Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Comput Biol Med. 2020, 121:103761–103761. 10.1016/J.COMPBIOMED.2020.103761

24. BergstraJames, BengioYoshua: Random search for hyper-parameter optimization. J Mach Learn Res. Published Online First: February 2012. 10.5555/2188385.2188395

25. Kim HS, Oh SJ, Kim BK, et al.: Dysbiotic signatures and diagnostic potential of gut microbial markers for inflammatory bowel disease in Korean population. Sci Rep 2024 141. 2024, 14:23701. 10.1038/s41598-024-74002-6

26. The Leading Multi-Agent Platform.

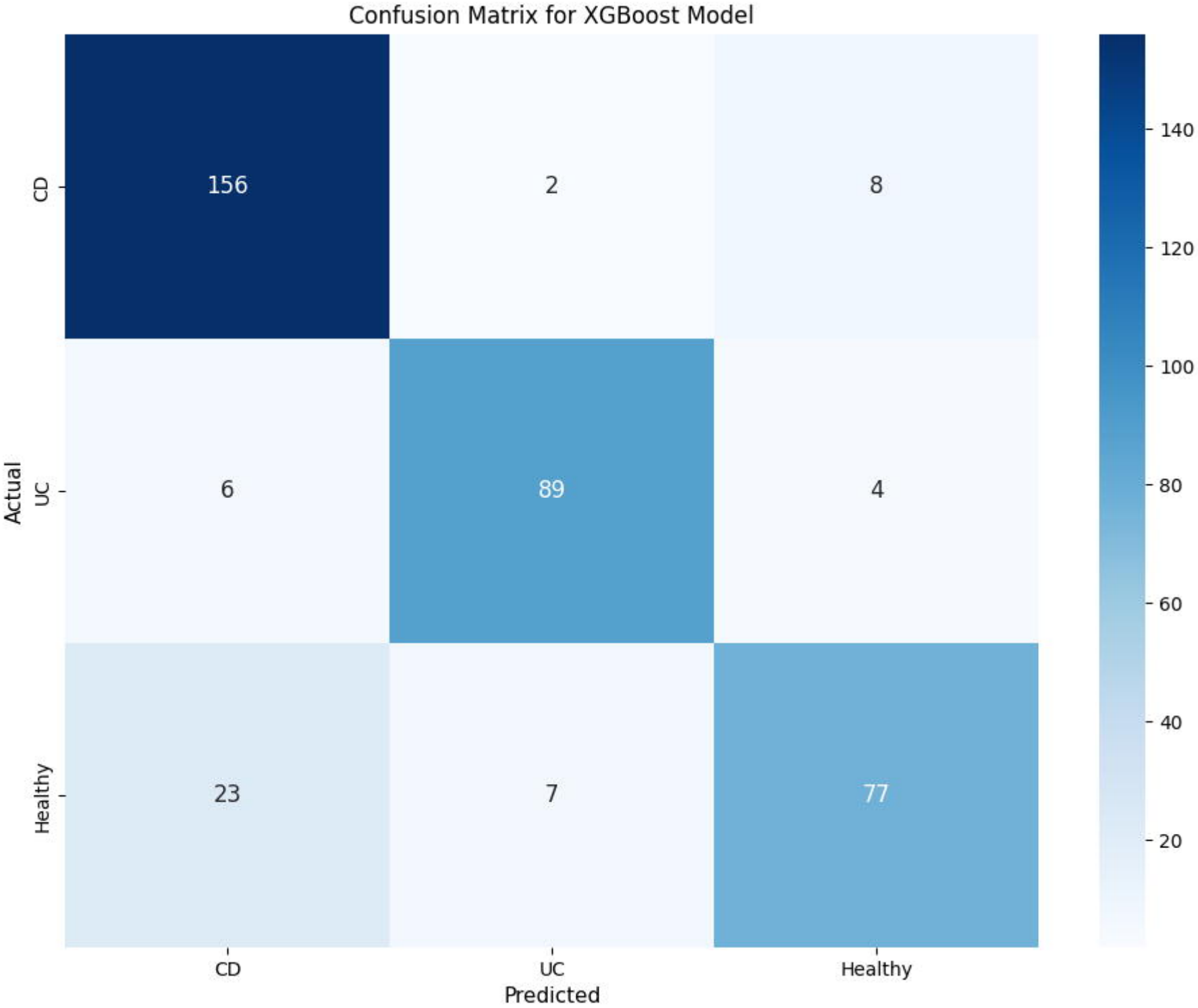
27. Team G, Kamath A, Ferret J, et al.: Gemma 3 Technical Report. Published Online First: March 2025.

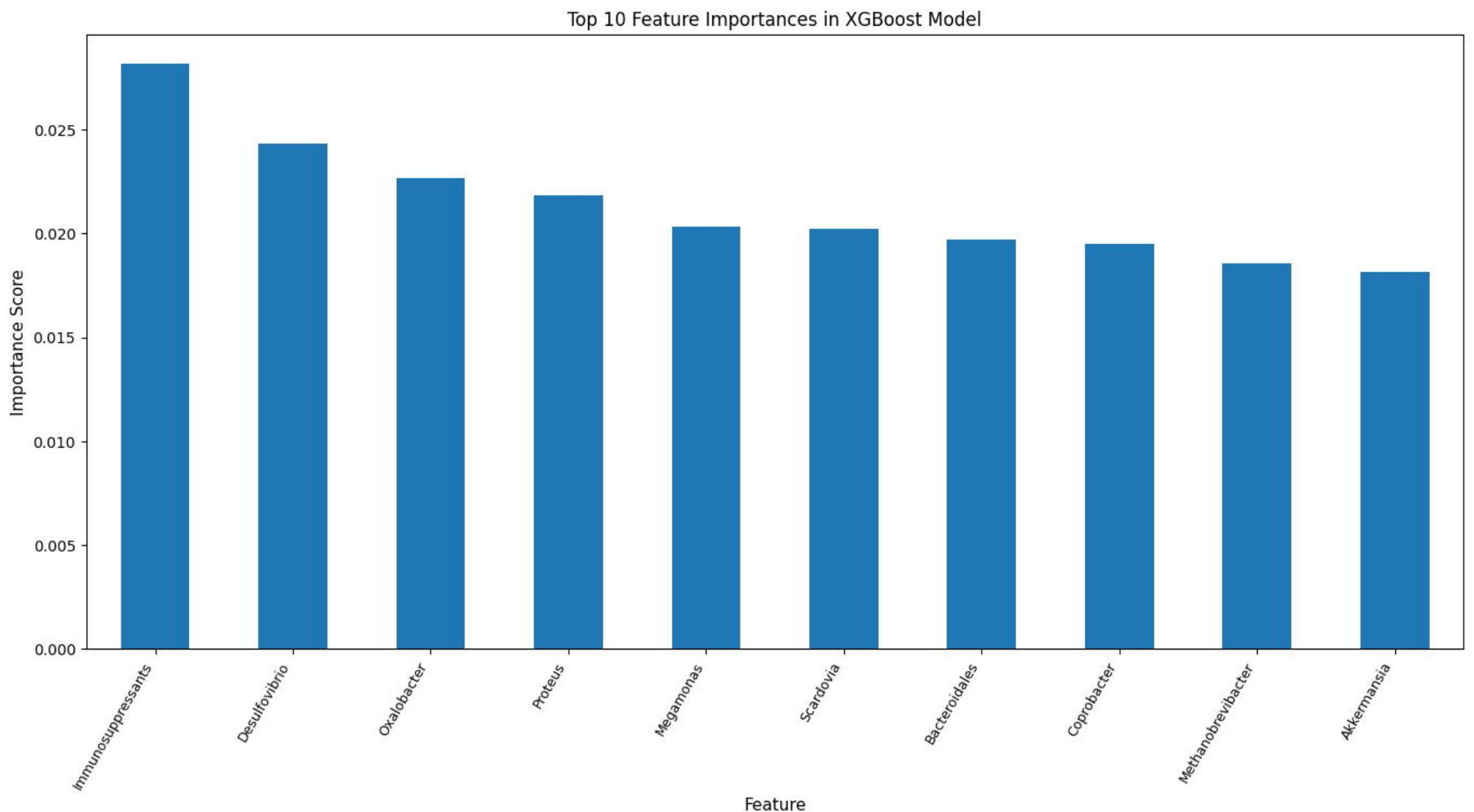
28. Nakase H, Uchino M, Shinzaki S, et al.: Evidence-based clinical practice guidelines for inflammatory bowel disease 2020. J Gastroenterol. 2021, 56:489–526. 10.1007/S00535-021-01784-1/FIGURES/9

29. Galazzo G, Tedjo DI, Wintjens DSJ, et al.: Faecal Microbiota Dynamics and their Relation to Disease Course in Crohn's Disease. J Crohns Colitis. 2019, 13:1273–82. 10.1093/ECCO-JCC/JJZ049

<u>Z score</u> - $$Z = \frac{X - \mu}{\sigma} \quad (1)$$

<u>Fold Change</u> - $$FC = \left(\frac{X}{\mu + 10^{-6}} \right) \quad (2)$$


Confusion Matrix for XGBoost Model
156
2
8
6
89
4
23
7
77
CD
UC
Healthy
Actual
CD
UC
Healthy
Predicted
140
120
100
80
60
40
20


Top 10 Feature Importances in XGBoost Model
Importance Score
0.025
0.020
0.015
0.010
0.005
0.000
Immunosuppressants
Desulfovibrio
Oxalobacter
Proteus
Megamonas
Scardovia
Bacteroidales
Coprobacter
Methanobrevibacter
Akkermansia
Feature

Pipeline Results

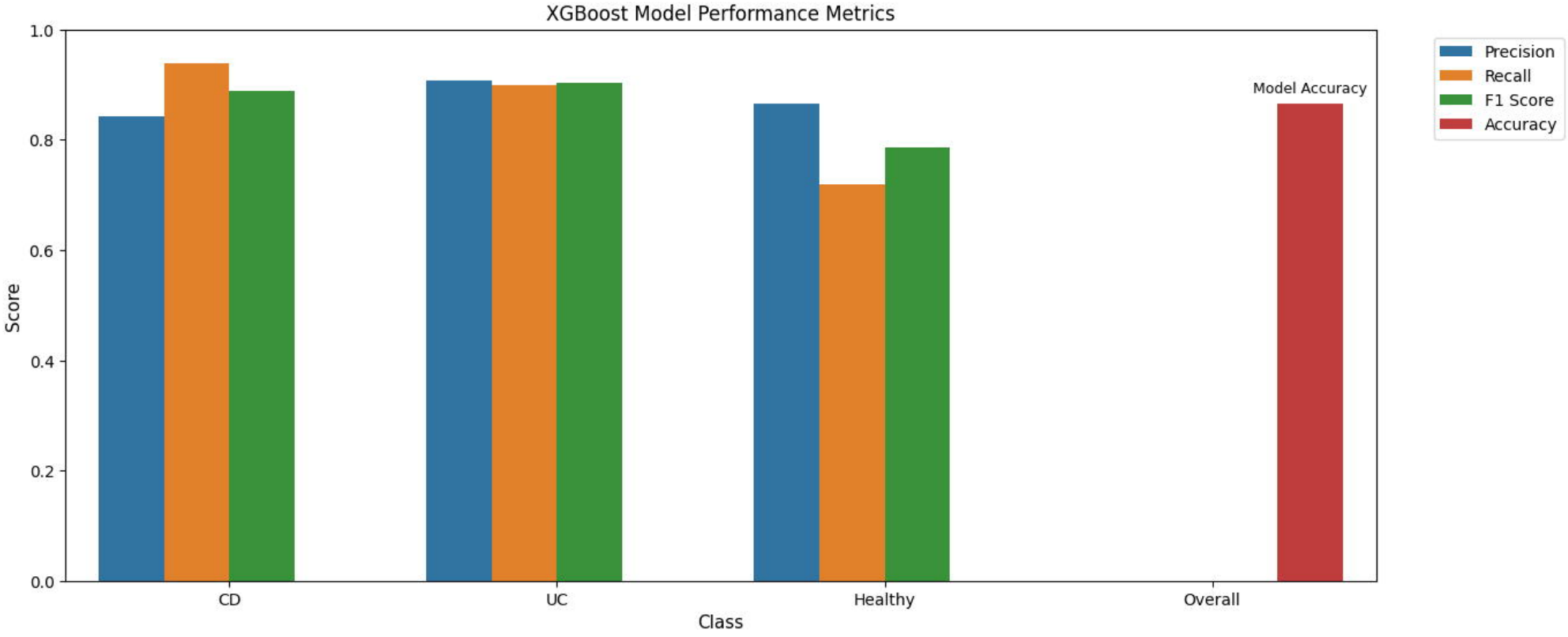
Kneaddata completed successfully!

MetaPhlAn completed successfully! Results saved to /home/mathew-thomas/Documents/project/metaphlan_genus_results.txt

CSV Converter completed successfully! Saved as: /home/mathew-thomas/Documents/project/metaphlan_results_2.csv

Diagnosis

Diagnosis: Crohn's Disease. Proceeding to dysbiosis analysis...


XGBoost Model Performance Metrics
Precision
Recall
F1 Score
Accuracy
Model Accuracy
Score
1.0
0.8
0.6
0.4
0.2
0.0
CD
UC
Healthy
Overall
Class

Probiotic Recommendation

Recommendation generated successfully!

Final Output:

Based on the synthesized information from the provided research articles, a targeted probiotic strategy for Crohn's disease should focus on enriching the gut microbiome with *Faecalibacterium prausnitzii* (F. prausnitzii). Several studies demonstrate that F. prausnitzii is a key anti-inflammatory commensal bacterium, deficient in active Crohn's patients, and its restoration can alleviate disease activity. Furthermore, the research highlights that combining F. prausnitzii with other beneficial species like *Bacteroides faecis* and *Roseburia intestinalis* can synergistically enhance the therapeutic effect, improving clinical symptoms, histological inflammation, and immune modulation. Therefore, a combined probiotic formulation targeting F. prausnitzii alongside other well-established beneficial bacteria represents the most promising approach for managing Crohn's disease, prioritizing the re-establishment of this critical microbial player within the gut ecosystem.

Previous Pipeline Results

Based on the synthesized information from the provided research articles, a targeted probiotic strategy for Crohn's disease should focus on enriching the gut microbiome with *Faecalibacterium prausnitzii* (F. prausnitzii). Several studies demonstrate that F. prausnitzii is a key anti-inflammatory commensal bacterium, deficient in active Crohn's patients, and its restoration can alleviate disease activity. Furthermore, the research highlights that combining F. prausnitzii with other beneficial species like *Bacteroides faecis* and *Roseburia intestinalis* can synergistically enhance the therapeutic effect, improving clinical symptoms, histological inflammation, and immune modulation. Therefore, a combined probiotic formulation targeting F. prausnitzii alongside other well-established beneficial bacteria represents the most promising approach for managing Crohn's disease, prioritizing the re-establishment of this critical microbial player within the gut ecosystem.