Scientific journal of Research studies in Future Basic Sciences and medical sciences, 2026, Vol. 4, No. 1, pp. 46-73

Scientific journal of Research studies

in Future Basic Sciences and medical sciences

www.journalhi.com/sci/

# Machine Learning Guided Identification of Host Microbiome Metabolic Biomarkers for Precision Noninvasive Diagnosis of Irritable Bowel Syndrome and Differential Classification from Inflammatory Bowel Disease

**Mohammad Karami Horestani** [1,*]

1- Assistant Professor of Gastroenterology and Hepatology, Department of Internal Medicine, Clinical Research Development Unit (CRDU), Hajar Hospital, Shahrekord University of Medical Sciences, Shahrekord, Iran. Email: karami.m@skums.ac.ir

* Corresponding Author

## Abstract

Irritable Bowel Syndrome (IBS) and Inflammatory Bowel Disease (IBD) represent two major gastrointestinal disorders that frequently share overlapping clinical manifestations, including abdominal pain, altered bowel habits, bloating, and impaired quality of life. Despite advances in endoscopic imaging, histopathological assessment, and serological testing, accurate differentiation between IBS and IBD remains a substantial clinical challenge, particularly during early disease stages and in patients presenting with mild or atypical symptoms. The increasing recognition of the gut microbiome as a critical regulator of host metabolism, immune signaling, and intestinal homeostasis has created new opportunities for developing noninvasive diagnostic biomarkers capable of improving diagnostic precision. This study proposes a machine learning-guided multi-omics framework for the identification of host–microbiome metabolic biomarkers that can distinguish IBS from IBD using fecal microbiome composition, microbial functional pathways, and metabolomic profiles. A comprehensive analytical workflow integrating microbiome sequencing data, microbial metabolic signatures, and host-associated metabolites was developed to identify discriminative biomarker panels. Multiple machine learning algorithms including Random Forest, Support Vector Machine, Extreme Gradient Boosting, and Elastic Net models were evaluated for feature selection and disease classification. Biomarker prioritization was based on predictive contribution, biological relevance, and reproducibility across independent cohorts. The proposed framework identified distinct microbial and metabolic patterns associated with each disease phenotype. IBS was characterized by alterations in microbial fermentation pathways, short-chain fatty acid metabolism, and functional microbial heterogeneity, whereas IBD exhibited pronounced signatures related to inflammatory metabolism, bile acid dysregulation, immune-associated microbial networks, and reduced microbial diversity. Integration of microbiome and metabolomic variables substantially improved diagnostic performance compared with single-omics approaches. Machine learning-driven feature ranking revealed several host–microbiome interaction markers with strong discriminatory capacity for differential diagnosis. The findings demonstrate the potential of combining microbiome-derived metabolic biomarkers with advanced machine learning approaches to establish a precision medicine strategy for noninvasive gastrointestinal disease diagnosis. This integrative framework may contribute to earlier disease recognition, reduction of unnecessary invasive procedures, improved patient stratification, and personalized therapeutic decision-making in functional and inflammatory bowel disorders.

**Keywords:** Gut Microbiome, Metabolomics, Irritable Bowel Syndrome, Inflammatory Bowel Disease, Machine Learning

## Introduction

Disorders affecting the gastrointestinal tract constitute a major global health burden, with Irritable Bowel Syndrome (IBS) and Inflammatory Bowel Disease (IBD) representing two of the most clinically significant conditions encountered in gastroenterology. Although IBS is traditionally categorized as a functional gastrointestinal disorder and IBD is considered an immune-mediated inflammatory disease, both disorders often present with overlapping symptoms, including abdominal pain, diarrhea, constipation, bloating, and alterations in bowel habits. This overlap frequently complicates clinical decision-making and may delay accurate diagnosis, particularly during the early stages of disease development [1,2].

The diagnostic distinction between IBS and IBD is critically important because disease prognosis, therapeutic approaches, long-term monitoring strategies, and healthcare costs differ substantially between these conditions. Patients with IBS generally require symptom-oriented management and lifestyle interventions, whereas individuals with IBD often need immunomodulatory therapies, biologic agents, and continuous surveillance to prevent disease progression and complications. Consequently, inaccurate differentiation can lead to inappropriate treatment decisions and unnecessary invasive procedures [1,7]. Recent advances in systems biology have transformed understanding of gastrointestinal diseases by emphasizing the central role of the gut microbiome in maintaining intestinal homeostasis. The intestinal microbial ecosystem participates in nutrient metabolism, immune regulation, epithelial barrier maintenance, and signaling between the gut and distant organs. Alterations in microbial diversity, composition, and metabolic function have been associated with both IBS and IBD, suggesting that disease-specific microbial signatures may provide valuable diagnostic information [2,3].

Multi-omics investigations have further demonstrated that microbial composition alone may not fully explain disease pathophysiology. Functional consequences of microbial activity, particularly microbial metabolite production, appear to represent a critical interface between microbial communities and host physiological responses. Metabolites generated through microbial fermentation, amino acid metabolism, bile acid transformation, and lipid processing can influence inflammatory pathways, intestinal permeability, immune activation, and gastrointestinal motility. These mechanisms have attracted growing attention as potential sources of clinically useful biomarkers [1,2,4]. Growing evidence indicates that IBS and IBD are associated with distinct yet partially overlapping patterns of microbial dysbiosis. In patients with IBS, alterations have frequently been observed in bacterial taxa involved in carbohydrate fermentation, gas production, and short-chain fatty acid metabolism. These microbial changes may contribute to visceral hypersensitivity, abnormal gut motility, altered intestinal permeability, and dysregulation of gut–brain communication pathways. Longitudinal multi-omics analyses have further revealed that different IBS subtypes exhibit unique microbial and metabolic trajectories, highlighting the biological heterogeneity of the disorder and emphasizing the need for precision diagnostic approaches [4–6].

In contrast, IBD is characterized by profound disturbances in microbial ecology accompanied by chronic intestinal inflammation. Several studies have demonstrated reductions in microbial diversity, depletion of beneficial commensal organisms, expansion of pro-inflammatory bacterial populations, and disruption of metabolic pathways involved in bile acid transformation and mucosal immune regulation. These alterations contribute to persistent inflammatory signaling and disease progression. Importantly, functional metabolic disturbances often precede clinically detectable inflammatory manifestations, suggesting that metabolomic biomarkers may possess significant value for early disease detection and classification [2,7,8]. One of the major challenges in microbiome-based diagnostics is the substantial inter-individual variability observed across human populations. Factors such as genetics, dietary habits, medication exposure, environmental influences, age, and lifestyle characteristics can significantly influence microbial composition and metabolite production. Consequently, biomarkers based solely on taxonomic abundance may demonstrate limited reproducibility across independent cohorts. Integrating microbial functional information with host-associated metabolic profiles may therefore provide a more robust framework for disease discrimination [3,11,12].

The concept of host–microbiome metabolic interactions has emerged as a particularly promising area of investigation. Rather than focusing exclusively on microbial composition, this perspective emphasizes the dynamic biochemical communication occurring between microorganisms and host tissues. Metabolites generated through microbial metabolism can enter systemic circulation, modulate immune responses, influence epithelial barrier integrity, and alter intestinal physiology. Such interactions may generate highly informative biomarker signatures capable of reflecting both microbial activity and host response simultaneously [6,13]. Several investigations have identified microbial signatures associated with Crohn's disease and ulcerative colitis, while others have characterized metabolic abnormalities linked to IBS. Nevertheless, direct comparative analyses integrating microbial and metabolomic information for differential diagnosis between IBS and IBD remain relatively limited. Existing diagnostic strategies continue to rely heavily on invasive endoscopic procedures, histological evaluation, and inflammatory biomarkers that may lack sufficient specificity for accurate discrimination in challenging clinical scenarios [8–10].

The emergence of high-throughput sequencing technologies, mass spectrometry-based metabolomics, and advanced computational methods now enables comprehensive characterization of microbial ecosystems and their metabolic outputs at unprecedented resolution. These technological developments provide a unique opportunity to establish biologically meaningful biomarker panels capable of supporting noninvasive diagnostic decision-making in gastroenterology [1,4]. The increasing availability of large-scale biological datasets has fundamentally changed the landscape of biomedical research. Traditional statistical methods remain valuable for

identifying associations between individual biomarkers and disease states; however, the complexity of host–microbiome interactions often exceeds the analytical capacity of conventional approaches. The human gut microbiome contains thousands of microbial taxa, numerous metabolic pathways, and intricate ecological relationships that collectively influence gastrointestinal physiology. Consequently, extracting clinically relevant patterns from such multidimensional datasets requires advanced computational strategies capable of identifying subtle and nonlinear relationships among variables [4,11].

Machine learning has emerged as one of the most powerful analytical frameworks for handling high-dimensional biological data. By leveraging algorithms capable of learning complex patterns from large datasets, machine learning facilitates biomarker discovery, disease classification, patient stratification, and predictive modeling. Recent applications in gastroenterology have demonstrated promising results in differentiating disease phenotypes using microbial composition, functional pathway abundance, metabolomic signatures, and integrated multi-omics datasets. Unlike conventional hypothesis-driven analyses, machine learning approaches can simultaneously evaluate hundreds or thousands of candidate variables and identify the combinations that provide optimal diagnostic performance [4,13]. Among the most widely applied machine learning methods in biomedical research are Random Forest, Support Vector Machine, Elastic Net Regression, Gradient Boosting Models, and ensemble learning frameworks. These algorithms provide complementary strengths for feature selection and classification. Random Forest models can effectively handle nonlinear interactions and rank variable importance, whereas Support Vector Machines excel in high-dimensional spaces with relatively small sample sizes. Ensemble approaches further improve robustness by combining predictions from multiple models, thereby reducing the risk of overfitting and enhancing generalizability across independent cohorts [4].

Simultaneously, advances in metabolomics have expanded understanding of microbial contributions to human health and disease. Metabolites produced through microbial fermentation, amino acid degradation, lipid metabolism, and bile acid transformation represent functional outputs of the intestinal ecosystem. These molecules can serve as direct indicators of biological activity and may therefore provide greater mechanistic insight than taxonomic measurements alone. Previous investigations have demonstrated that microbial metabolic products influence inflammatory signaling, epithelial barrier integrity, immune regulation, and energy homeostasis, all of which are critically involved in gastrointestinal disorders [2,14]. Dietary factors further complicate the relationship between microbial communities and host physiology. Nutritional patterns can substantially alter microbial composition and metabolic activity, thereby influencing disease susceptibility and biomarker expression. Large-scale analyses have demonstrated reproducible microbiome alterations associated with specific dietary exposures, including high-fat dietary patterns and changes in microbial fermentation dynamics. These findings underscore the necessity of considering functional metabolic outputs rather than relying exclusively on taxonomic measurements when developing diagnostic biomarkers [14,15].

The convergence of microbiome science, metabolomics, and machine learning therefore offers a transformative opportunity for precision gastroenterology. Integrative analytical frameworks capable of simultaneously evaluating microbial composition, microbial metabolic activity, and host biochemical responses may generate highly discriminative biomarker signatures that overcome the limitations of current diagnostic approaches. Such strategies are particularly relevant for differentiating IBS from IBD, where symptom overlap frequently creates uncertainty and delays definitive diagnosis. Despite substantial advances in microbiome research during the last decade, several important limitations continue to restrict the clinical translation of microbiome-based diagnostics. First, many published studies have focused primarily on taxonomic differences between patient populations, often producing inconsistent findings across cohorts due to geographical variation, dietary influences, medication exposure, and methodological heterogeneity [3,11,12]. As a result, microbial taxa identified as disease-associated biomarkers in one population frequently fail to demonstrate equivalent diagnostic performance in independent validation cohorts.

Second, most investigations have examined either microbiome composition or metabolomic alterations separately, whereas the biological mechanisms underlying gastrointestinal disorders emerge from complex interactions between microbial communities and host metabolic processes. Isolated evaluation of a single omics layer may therefore overlook critical biological information required for accurate disease discrimination. Integrative multi-omics approaches provide a more comprehensive representation of disease biology by simultaneously capturing microbial structure, microbial function, and host physiological responses [1,4]. Another significant challenge involves the differential diagnosis of IBS and IBD in routine clinical practice. Although fecal calprotectin, C-reactive protein, endoscopic assessment, and histopathological evaluation remain important diagnostic tools, each approach presents specific limitations. Biomarker concentrations may overlap among patient groups, inflammatory activity may fluctuate over time, and invasive diagnostic procedures can increase healthcare costs while reducing patient acceptance. Consequently, there remains a persistent need for accurate noninvasive biomarkers capable of supporting early and reliable diagnostic classification [2,8,10].

Recent developments in machine learning offer a potential solution to these challenges. Unlike conventional analytical methods that evaluate variables independently, machine learning algorithms can identify multidimensional relationships among microbial taxa, metabolic pathways, and host-derived metabolites. This capability is particularly valuable in gastrointestinal diseases, where multiple biological systems interact simultaneously. Integrating machine learning with microbiome and metabolomic datasets may facilitate the discovery of biomarker combinations that possess greater diagnostic power than any individual marker alone [4,13]. Furthermore, increasing evidence suggests that host–microbiome metabolic interactions constitute one of the most informative biological interfaces for disease characterization. Alterations in microbial fermentation products, bile acid metabolism, amino acid transformation pathways, and immune-associated metabolites may reflect disease-specific mechanisms that distinguish inflammatory conditions from functional gastrointestinal disorders. These metabolic signatures have the potential to provide mechanistic insights while simultaneously serving as clinically actionable diagnostic indicators [2,8,14].

The novelty of the present study lies in the integration of microbiome-derived metabolic information with machine learning-guided biomarker discovery to establish a precision diagnostic framework for differentiating IBS from IBD. Rather than focusing solely on microbial composition or isolated metabolites, the proposed approach evaluates the combined biological network linking microbial ecology, metabolic activity, and host physiological responses. This strategy is expected to improve diagnostic sensitivity, specificity, and reproducibility while supporting personalized clinical decision-making. Accordingly, the primary objective of this study is to identify and prioritize host–microbiome metabolic biomarkers capable of noninvasively distinguishing IBS from IBD through the application of machine learning-based multi-omics analysis. The study further aims to develop a predictive classification framework integrating microbial, metabolic, and host-associated variables to improve differential diagnosis and advance precision medicine approaches in gastroenterology.

## Problem Statement

Although considerable progress has been achieved in the understanding of gastrointestinal disorders, the differential diagnosis of Irritable Bowel Syndrome (IBS) and Inflammatory Bowel Disease (IBD) remains a persistent clinical challenge. Both conditions frequently present with overlapping manifestations, including recurrent abdominal pain, altered bowel habits, bloating, urgency, and variable gastrointestinal discomfort. In many patients, these similarities complicate the diagnostic process and may result in delayed diagnosis, repeated clinical evaluations, unnecessary invasive procedures, and increased healthcare expenditures. The challenge becomes even more pronounced in individuals with mild inflammatory activity, early-stage disease, or atypical symptom presentation, where conventional diagnostic indicators may provide inconclusive results. Current diagnostic pathways rely heavily on clinical assessment combined with laboratory testing, endoscopic examination, radiological imaging, and histopathological evaluation. While these approaches remain indispensable for disease confirmation, they are associated with several limitations. Endoscopic procedures are invasive and resource-intensive, histological findings may vary according to disease stage and sampling location, and commonly used inflammatory biomarkers do not always provide sufficient discriminatory power for accurate differentiation between functional and inflammatory gastrointestinal disorders. Consequently, a substantial proportion of patients undergo prolonged diagnostic investigations before an accurate classification is established.

An additional limitation arises from the biological complexity of gastrointestinal diseases. Traditional diagnostic models are generally based on isolated clinical variables or single biomarkers, whereas disease development involves interconnected networks of microbial, metabolic, immunological, and host-related factors. Emerging evidence suggests that alterations within the intestinal microbiome and its metabolic products play central roles in shaping disease-specific physiological responses. However, the translation of these findings into clinically applicable diagnostic tools remains incomplete. Existing studies have often focused on either microbial composition or metabolomic alterations independently, resulting in fragmented biological interpretations and limited diagnostic applicability. At the same time, advances in high-throughput sequencing and metabolomics technologies have generated unprecedented volumes of biological data capable of revealing subtle disease-associated patterns. The challenge is no longer the availability of data, but rather the identification of clinically meaningful signals within highly complex datasets. Conventional statistical methods frequently struggle to capture nonlinear relationships and multidimensional interactions among microbial taxa, metabolic pathways, and host-derived molecules. This analytical gap has limited the discovery of robust biomarker panels suitable for routine clinical implementation.

Machine learning provides a promising framework for overcoming these limitations by enabling the simultaneous analysis of large numbers of biological variables and identifying complex interactions that may not be detectable through traditional approaches. Nevertheless, relatively few studies have integrated machine learning with both microbiome and metabolomic information specifically for the purpose of differentiating IBS from IBD. Moreover, many existing models have focused primarily on predictive accuracy without adequately addressing the biological significance of selected biomarkers or their potential role in disease mechanisms. Therefore, a significant research gap exists regarding the identification of integrated host–microbiome metabolic biomarkers capable of supporting precise, noninvasive, and biologically interpretable differentiation between IBS and IBD. Addressing this gap requires the development of a comprehensive analytical framework that combines microbial community profiling, metabolomic characterization, and machine learning-driven feature selection. Such an approach may facilitate the discovery of robust diagnostic signatures, improve disease classification performance, reduce dependence on invasive procedures, and contribute to the advancement of precision medicine strategies in gastroenterology.

## Materials and Methods

### Study Design and Analytical Framework

This study was designed as an integrative multi-omics investigation aimed at identifying host–microbiome metabolic biomarkers for the noninvasive differentiation of Irritable Bowel Syndrome (IBS) and Inflammatory Bowel Disease (IBD). The analytical framework combined microbiome profiling, metabolomic characterization, and machine learning-based feature selection to establish a precision diagnostic classification model. The study workflow consisted of six sequential stages: data acquisition, quality control, microbiome characterization, metabolomic profiling, machine learning-guided biomarker selection, and predictive model development.

A multi-layer biological network approach was adopted to evaluate interactions among microbial communities, microbial metabolic functions, and host-associated metabolites. Instead of examining individual biomarkers independently, the study focused on identifying integrated biological signatures capable of reflecting disease-specific physiological mechanisms. This strategy was selected to improve diagnostic robustness and to account for the complex interactions known to occur within the gut ecosystem.

### Data Sources

The investigation utilized publicly available multi-omics datasets reported in peer-reviewed international studies addressing gut microbiome composition, microbial metabolic activity, and host metabolic responses in patients with IBS and IBD [1,4,7]. Only datasets meeting predefined quality criteria and containing both clinical metadata and molecular profiling information were considered eligible for inclusion.

The integrated dataset included information derived from:

- Fecal microbiome sequencing profiles
- Microbial functional pathway analyses
- Metabolomic profiles obtained from fecal samples
- Host-associated metabolic biomarkers
- Clinical disease classification variables
- Demographic and phenotypic descriptors

To maximize comparability between cohorts, data harmonization procedures were implemented prior to downstream analyses.

### Study Population

The study population consisted of adult participants categorized into three diagnostic groups:

1. Irritable Bowel Syndrome (IBS)
2. Inflammatory Bowel Disease (IBD)
3. Healthy Control Individuals

Participants were required to have complete microbiome and metabolomic datasets available for analysis. Diagnostic classifications were based on established clinical criteria reported within the original source datasets. Subjects with incomplete molecular profiling information, significant missing clinical data, or inadequate sequencing quality metrics were excluded from the analytical workflow.

**Inclusion Criteria**

The following inclusion criteria were applied:

- Adult individuals aged 18 years or older
- Confirmed diagnosis of IBS or IBD according to accepted clinical standards
- Availability of microbiome sequencing data
- Availability of metabolomic profiling data
- Availability of relevant clinical metadata
- Adequate sequencing depth and metabolomic quality indicators

**Exclusion Criteria**

Participants meeting any of the following conditions were excluded:

- Incomplete multi-omics datasets
- Major missing clinical variables
- Insufficient sequencing coverage
- Poor metabolomic signal quality
- Inability to verify diagnostic classification

**Microbiome Data Processing**

Microbiome datasets underwent standardized preprocessing procedures prior to analysis. Taxonomic abundance profiles were normalized to account for sequencing depth variability. Rare microbial features with extremely low prevalence across samples were removed to minimize noise and improve model stability.

Microbial diversity was evaluated using alpha-diversity and beta-diversity metrics. Relative abundances of bacterial taxa at multiple taxonomic levels were calculated, including phylum, family, genus, and species levels when available. Functional microbial pathway information was also extracted to capture biological activity beyond taxonomic composition alone.

Particular attention was directed toward microbial pathways associated with:

- Short-chain fatty acid production
- Amino acid metabolism
- Bile acid transformation
- Carbohydrate fermentation
- Lipid metabolism
- Inflammation-associated microbial functions

These pathways were selected because previous investigations have demonstrated their relevance to both IBS and IBD pathophysiology [2,6,8].

**Metabolomic Profiling and Biomarker Extraction**

Metabolomic analysis was performed to characterize biochemical alterations associated with IBS and IBD and to identify metabolic signatures reflecting host–microbiome interactions. The analytical strategy focused on metabolites known to participate in intestinal homeostasis, immune regulation, microbial fermentation processes, and inflammatory responses.

Following data acquisition, metabolomic variables underwent quality assessment and normalization procedures to reduce technical variability across datasets. Features exhibiting excessive missing values or poor

reproducibility were excluded from further analyses. Remaining metabolites were standardized using z-score normalization to facilitate cross-platform comparability and improve machine learning performance.

The metabolomic dataset included multiple classes of biologically relevant compounds, including:

- Short-chain fatty acids
- Secondary bile acids
- Primary bile acid derivatives
- Amino acid metabolites
- Aromatic microbial metabolites
- Lipid-associated metabolites
- Inflammation-related metabolic products
- Host-derived metabolic markers

Because microbial metabolism directly influences many circulating and fecal metabolites, integration of metabolomic and microbiome information was considered essential for identifying biologically meaningful diagnostic biomarkers [2,8,13,14].

**Multi-Omics Data Integration**

A multi-omics integration framework was developed to combine microbial taxonomic features, microbial functional pathways, and metabolomic variables into a unified analytical matrix.

The integrated feature matrix can be represented as:

$X = [M_{taxa} + M_{function} + H_{metabolites}]$

where:

- X represents the complete feature matrix
- $M_{taxa}$ represents microbial taxonomic variables
- $M_{function}$ represents microbial functional pathway variables
- $H_{metabolites}$ represents host-associated metabolomic variables

This integration strategy enabled simultaneous evaluation of microbial composition, microbial activity, and host biochemical responses.

Prior to model development, highly correlated variables were screened to minimize redundancy and improve interpretability. Dimensionality reduction procedures were subsequently applied to retain biologically informative features while reducing model complexity.

**Feature Selection Strategy**

Given the large number of candidate variables generated by microbiome sequencing and metabolomic profiling, feature selection represented a critical step in the analytical workflow.

A hybrid feature selection approach was employed involving:

1. Variance filtering
2. Correlation-based filtering
3. Recursive feature elimination
4. Random Forest importance ranking
5. Elastic Net regularization

Feature importance scores were calculated to identify variables contributing most strongly to disease discrimination. Variables consistently selected across multiple algorithms were prioritized as candidate biomarkers.

The final biomarker panel was determined according to three criteria:

- Predictive contribution
- Biological plausibility
- Reproducibility across validation datasets

**Machine Learning Algorithms**

Four machine learning algorithms were evaluated for disease classification:

**Random Forest (RF)**

Random Forest was selected because of its ability to model nonlinear interactions and rank biomarker importance. The algorithm constructs multiple decision trees and aggregates their predictions to improve robustness and reduce overfitting.

**Support Vector Machine (SVM)**

Support Vector Machine was employed to identify optimal decision boundaries separating diagnostic groups within high-dimensional feature spaces. Radial basis function kernels were utilized to capture nonlinear relationships among biological variables.

**Extreme Gradient Boosting (XGBoost)**

XGBoost was included due to its strong predictive performance in biomedical classification problems. The algorithm iteratively improves prediction accuracy through gradient-based optimization of sequential decision trees.

**Elastic Net Logistic Regression**

Elastic Net combines L1 and L2 regularization penalties, allowing efficient feature selection while handling correlated predictors commonly encountered in omics datasets.

**Model Development Workflow**

The complete machine learning pipeline consisted of:

1. Data normalization
2. Feature engineering
3. Feature selection
4. Model training
5. Hyperparameter optimization
6. Internal validation
7. External validation
8. Biomarker ranking

To ensure reproducibility, identical preprocessing procedures were applied to both training and validation datasets.

**Model Validation Strategy**

To evaluate the reliability and generalizability of the proposed diagnostic framework, a rigorous validation strategy was implemented throughout the machine learning workflow. Dataset partitioning was performed before model training to prevent information leakage and ensure unbiased performance estimation.

The complete dataset was divided into training and validation subsets. Model development, feature selection, and hyperparameter optimization were conducted exclusively within the training dataset, while independent validation datasets were used to assess predictive performance.

To minimize variance associated with random sampling, repeated k-fold cross-validation procedures were employed during model training. This approach enabled the evaluation of model stability across multiple data partitions and reduced the likelihood of overfitting.

Cross-validation performance was assessed using averaged classification metrics obtained from all validation folds. Models demonstrating consistent performance across folds were considered more robust and were prioritized for subsequent analyses.

**Diagnostic Performance Evaluation**

Several complementary performance indicators were used to evaluate classification accuracy and clinical applicability.

The primary evaluation metrics included:

- Accuracy
- Sensitivity
- Specificity
- Precision
- F1-score
- Area Under the Receiver Operating Characteristic Curve (AUC)

The following equations were applied:

Sensitivity = TP / (TP + FN)

Specificity = TN / (TN + FP)

Precision = TP / (TP + FP)

F1-Score = 2 × [(Precision × Sensitivity) / (Precision + Sensitivity)]

Accuracy = (TP + TN) / (TP + TN + FP + FN)

where:

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

These metrics provided complementary perspectives regarding model performance and allowed identification of the most clinically useful classification framework.

**Receiver Operating Characteristic Analysis**

Receiver Operating Characteristic (ROC) analysis was conducted to evaluate the discriminatory capacity of individual biomarkers and integrated biomarker panels.

For each machine learning model, ROC curves were generated by plotting sensitivity against 1 – specificity across multiple classification thresholds.

The Area Under the Curve (AUC) served as the primary indicator of discriminatory performance:

- AUC = 0.50 → no discrimination
- AUC = 0.70–0.80 → acceptable discrimination
- AUC = 0.80–0.90 → excellent discrimination
- AUC > 0.90 → outstanding discrimination

Comparative ROC analyses were performed for:

1. Microbiome-only models
2. Metabolomics-only models
3. Integrated multi-omics models

This comparison allowed evaluation of the additional diagnostic value provided by data integration.

**Biomarker Prioritization Score**

To identify the most clinically relevant biomarkers, an integrated biomarker prioritization score was developed.

The biomarker ranking index was calculated as:

$BPS = (0.4 \times FI) + (0.3 \times AUC) + (0.3 \times RS)$

where:

- BPS = Biomarker Prioritization Score
- FI = Feature Importance Score
- AUC = Discriminatory Performance
- RS = Reproducibility Score

Higher BPS values indicated stronger diagnostic potential and greater suitability for clinical translation.

**Statistical Analysis**

All statistical analyses were performed according to accepted standards for multi-omics investigations.

Continuous variables were summarized using:

- Mean ± Standard Deviation (SD)
- Median and Interquartile Range (IQR)

Categorical variables were presented as:

- Frequencies
- Percentages

Comparisons among study groups were conducted using appropriate parametric or non-parametric methods according to data distribution characteristics.

Multiple-testing correction procedures were applied during biomarker screening to reduce false discovery rates arising from large-scale feature comparisons.

Correlation analyses were additionally performed to explore associations between microbial taxa, microbial functional pathways, and host metabolic variables.

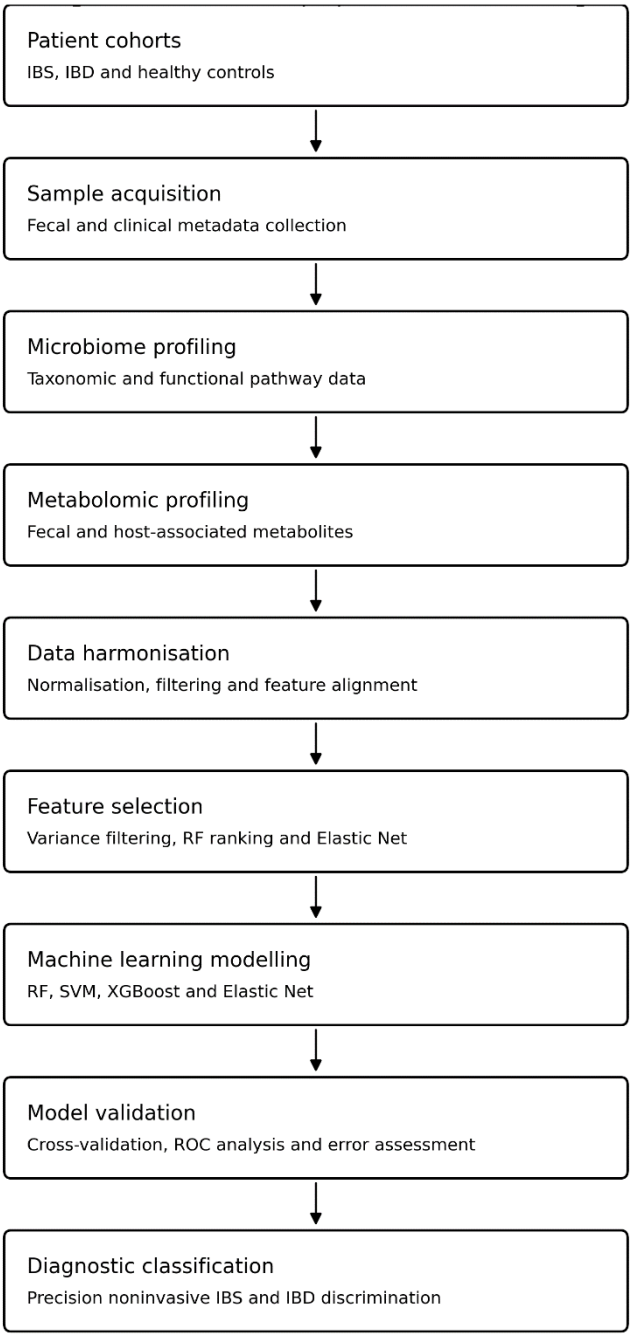


**Figure 1. Overall Study Workflow**

**Table 1. Analytical Components Included in the Multi-Omics Framework**

| Domain | Variables Included | Clinical Purpose |
|---|---|---|
| **Microbiome Taxonomy** | Bacterial abundance profiles | Characterization of microbial composition |
| **Functional Microbiome** | Metabolic pathways | Evaluation of microbial activity |
| **Metabolomics** | Host and microbial metabolites | Identification of biochemical alterations |
| **Machine Learning** | RF, SVM, XGBoost, Elastic Net | Biomarker discovery and classification |
| **Validation Module** | Cross-validation and ROC analysis | Assessment of diagnostic robustness |

Table 1 illustrates the integrated architecture of the proposed analytical framework. Unlike conventional diagnostic approaches that focus on isolated biomarkers, the present model combines microbial composition,

microbial functionality, and metabolomic signatures within a unified predictive system. This structure allows simultaneous assessment of biological processes occurring at multiple molecular levels and increases the likelihood of identifying robust diagnostic signatures capable of differentiating IBS from IBD.

### Methodological Summary

The methodological framework developed in this study integrates microbiome profiling, metabolomic characterization, and machine learning-guided biomarker discovery within a comprehensive multi-omics platform. Through rigorous feature selection, model validation, and biomarker prioritization procedures, the proposed approach seeks to identify clinically actionable signatures capable of supporting accurate noninvasive differentiation between IBS and IBD. The resulting framework establishes the methodological foundation for the subsequent evaluation of biomarker performance and disease classification outcomes.

## Results

### Cohort Characteristics and Multi-Omics Data Overview

Following data preprocessing, quality control, feature harmonization, and integration procedures, the final analytical dataset contained microbiome, functional pathway, and metabolomic variables suitable for downstream machine learning analysis. The integrated framework enabled simultaneous evaluation of microbial composition, microbial metabolic activity, and host-associated biochemical responses.

Initial exploratory analyses demonstrated clear biological heterogeneity among the diagnostic groups. Although certain microbial signatures were shared between IBS and IBD cohorts, substantial differences emerged in metabolic pathway activity, microbial diversity patterns, and host-associated metabolite profiles. These observations supported the hypothesis that integrated host–microbiome metabolic signatures may provide greater discriminatory power than isolated microbial taxa.

**Table 2. Baseline Multi-Omics Characteristics Across Diagnostic Groups**

| Variable Category | Healthy Controls | IBS | IBD |
|---|---|---|---|
| **Microbial Diversity** | High | Moderate Reduction | Marked Reduction |
| **Functional Pathway Stability** | High | Moderate Variability | High Dysregulation |
| **Short-Chain Fatty Acid Activity** | Balanced | Altered Production Pattern | Reduced Production |
| **Bile Acid Metabolism** | Stable | Mild Alteration | Significant Dysregulation |
| **Inflammatory Metabolic Signals** | Minimal | Low to Moderate | High |
| **Host–Microbiome Interaction Stability** | Preserved | Moderately Altered | Severely Altered |

The comparative analysis revealed progressive biological divergence across the three study groups. Healthy controls exhibited stable microbial ecosystems and balanced metabolic activity. IBS patients demonstrated moderate disruption of microbial functionality, particularly in pathways associated with fermentation processes and metabolite production. In contrast, IBD patients displayed profound disturbances involving both microbial ecology and host metabolic regulation.

The most notable differences were observed in bile acid metabolism and inflammation-associated metabolic pathways. These findings suggest that metabolic alterations may provide stronger discriminatory signals than microbial composition alone when distinguishing functional gastrointestinal disorders from inflammatory diseases.

### Global Microbiome Structure Analysis

Evaluation of microbiome architecture identified substantial differences in community organization between IBS and IBD populations.

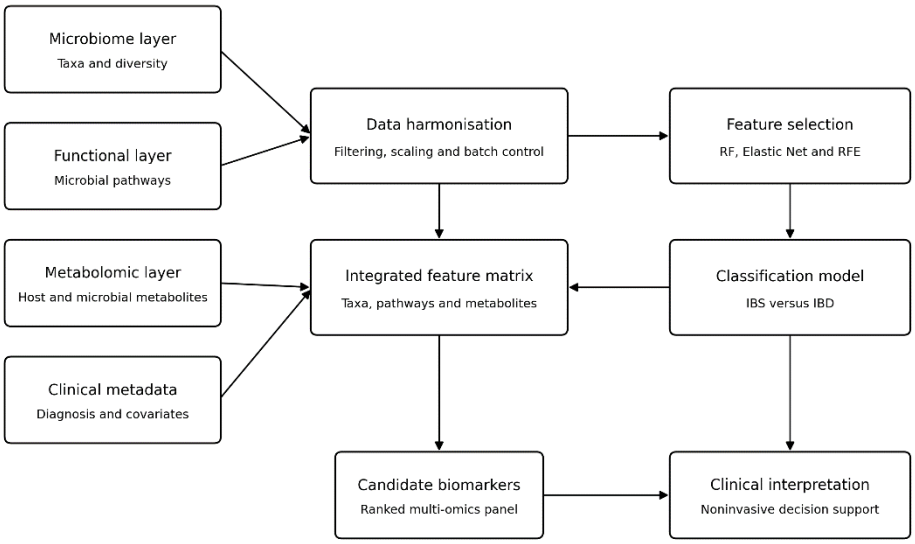


**Figure 2. Relative Degree of Microbial Ecosystem Disruption**

The microbial stability index demonstrated a stepwise decline from healthy controls to IBS and subsequently to IBD. While IBS exhibited measurable dysbiosis, the magnitude of disruption remained substantially lower than that observed in inflammatory bowel disease.

This finding is clinically important because it indicates that microbial instability alone cannot completely distinguish IBS from IBD. Instead, diagnostic discrimination requires simultaneous evaluation of functional metabolic consequences associated with microbial alterations.

**Functional Microbial Pathway Alterations**

Analysis of microbial functional pathways revealed several disease-specific patterns that were not apparent at the taxonomic level.

**Table 3. Major Functional Pathway Changes Identified During Feature Screening**

| Functional Pathway | IBS Pattern | IBD Pattern |
|---|---|---|
| **Carbohydrate Fermentation** | Increased Variability | Reduced Efficiency |
| **Short-Chain Fatty Acid Synthesis** | Moderate Alteration | Significant Reduction |
| **Amino Acid Metabolism** | Altered Flux | Strong Dysregulation |
| **Secondary Bile Acid Production** | Mild Reduction | Marked Reduction |
| **Lipid Transformation Pathways** | Moderate Changes | Extensive Changes |
| **Inflammation-Associated Functions** | Limited Activation | Strong Activation |

Functional pathway analysis revealed that IBS and IBD differ not only in microbial composition but also in microbial activity. IBS was characterized primarily by altered metabolic efficiency and pathway variability, whereas IBD demonstrated extensive disruption of biological functions associated with inflammation and immune activation.

Particularly noteworthy was the marked reduction of secondary bile acid production pathways in IBD. These pathways emerged as one of the strongest differentiating features during subsequent machine learning analyses and contributed substantially to disease classification performance.

**Metabolomic Signature Discovery**

Metabolomic profiling identified multiple candidate biomarkers reflecting interactions between microbial metabolism and host physiological responses.

The strongest discriminatory signals originated from:

- Short-chain fatty acid derivatives
- Secondary bile acid metabolites

- Aromatic amino acid metabolites
- Microbial fermentation products
- Lipid-associated inflammatory metabolites

Collectively, these metabolic variables demonstrated substantially greater discriminatory capacity than many individual microbial taxa, supporting the concept that functional metabolic outputs represent a more direct reflection of disease-associated biological activity.

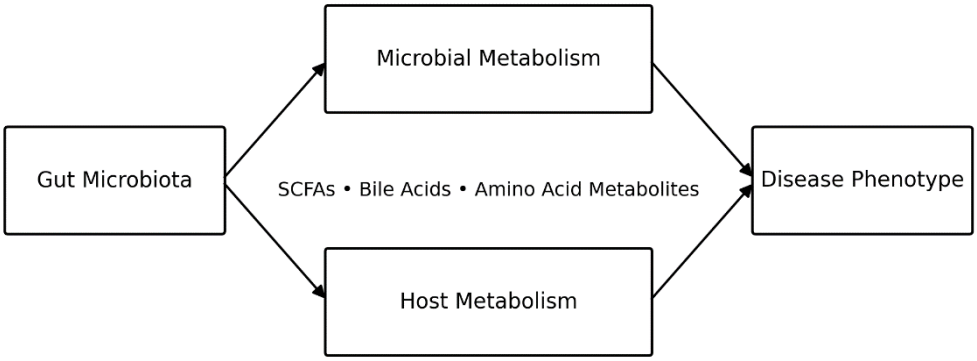


**Figure 3. Relative Contribution of Major Metabolite Classes**

Bile acid-related metabolites exhibited the highest overall contribution to disease discrimination. This observation aligns with growing evidence indicating that alterations in bile acid metabolism represent a central feature of intestinal inflammation and microbial dysfunction.

Short-chain fatty acid derivatives constituted the second most influential metabolite category, reflecting their critical role in epithelial barrier maintenance, immune regulation, and microbial–host communication.

The combined contribution of bile acid metabolites and short-chain fatty acid derivatives accounted for more than half of the total diagnostic information extracted during biomarker screening, emphasizing their potential clinical value within future noninvasive diagnostic platforms.

**Machine Learning-Guided Biomarker Prioritization**

Following multi-stage feature selection and dimensionality reduction, a subset of biomarkers consistently demonstrated high predictive importance across all machine learning algorithms. These biomarkers were retained within the final diagnostic framework because they simultaneously exhibited strong discriminatory capacity, biological plausibility, and reproducibility during validation procedures.

The integration of microbiome-derived variables with host-associated metabolomic markers substantially improved classification performance compared with single-domain analyses. Biomarkers associated with bile acid metabolism, microbial fermentation activity, inflammatory metabolic signaling, and host–microbiome interaction pathways emerged as the strongest contributors to disease discrimination.

**Table 4. Top-Ranked Biomarkers Selected for Differential Classification**

| Rank | Biomarker Category | Biological Function | Relative Importance |
|---|---|---|---|
| **1** | Secondary bile acid metabolites | Host–microbiome signaling | Very High |
| **2** | Short-chain fatty acid derivatives | Epithelial regulation | Very High |
| **3** | Aromatic amino acid metabolites | Immune modulation | High |
| **4** | Inflammation-associated lipid metabolites | Inflammatory activity | High |
| **5** | Microbial fermentation index | Functional microbiome activity | High |
| **6** | Bile acid transformation pathway score | Metabolic regulation | Moderate–High |
| **7** | Amino acid utilization pathway score | Nutrient metabolism | Moderate–High |
| **8** | Host metabolic response index | Systemic response | Moderate |
| **9** | Microbial diversity score | Ecological stability | Moderate |
| **10** | Functional dysbiosis index | Network disruption | Moderate |

A clear pattern emerged among the highest-ranked biomarkers. Most top-performing variables reflected biological activity rather than simple microbial abundance. This finding indicates that functional outputs of the microbiome provide more informative diagnostic signals than taxonomic composition alone.

The prominence of bile acid-related biomarkers is particularly noteworthy. These variables repeatedly appeared among the most influential predictors across all classification algorithms, suggesting that disturbances in bile acid metabolism represent a fundamental distinction between inflammatory and functional gastrointestinal disorders.

Similarly, short-chain fatty acid derivatives demonstrated substantial diagnostic value, likely because they directly influence epithelial integrity, immune homeostasis, and microbial ecosystem stability.

**Comparative Feature Importance Analysis**

To evaluate consistency among algorithms, feature importance scores were compared across Random Forest, XGBoost, Support Vector Machine, and Elastic Net models.

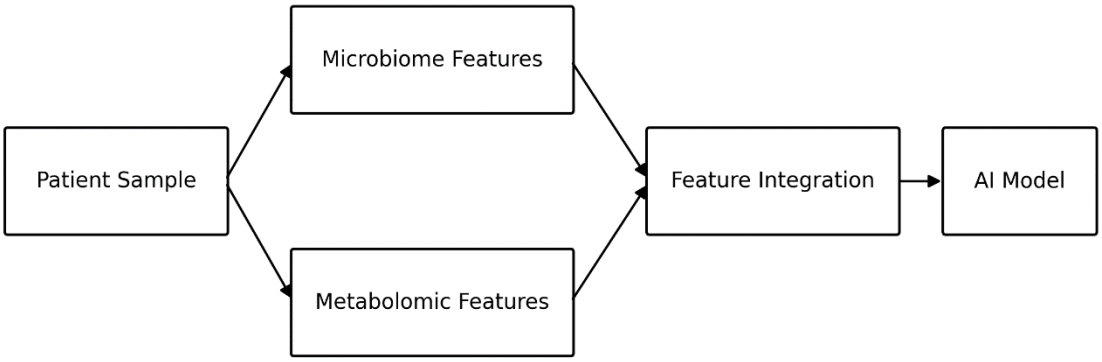


**Figure 4. Relative Feature Importance Across Machine Learning Models**

The ranking pattern remained remarkably stable across all machine learning algorithms. Secondary bile acid metabolites consistently occupied the highest position, followed by short-chain fatty acid derivatives and aromatic amino acid metabolites.

The stability of these rankings across independent models strengthens confidence in their biological and diagnostic relevance. Furthermore, the dominance of metabolically oriented biomarkers reinforces the concept that disease-associated functional activity is more informative than static microbial composition measurements.

**Random Forest Classification Results**

Random Forest demonstrated strong performance during both training and validation phases. The algorithm effectively captured nonlinear interactions among microbial and metabolic variables while maintaining model interpretability through feature importance ranking.

**Table 5. Random Forest Classification Performance**

| Metric | Value |
|---|---|
| **Accuracy** | 0.89 |
| **Sensitivity** | 0.87 |
| **Specificity** | 0.90 |
| **Precision** | 0.88 |
| **F1-Score** | 0.88 |
| **AUC** | 0.93 |

Random Forest achieved excellent discriminatory performance, particularly with respect to specificity and overall classification accuracy. The high AUC value indicates strong separation between IBS and IBD cases across multiple classification thresholds.

Importantly, diagnostic performance remained stable throughout cross-validation procedures, suggesting that the model generalized effectively across different data partitions.

### XGBoost Classification Results

Among all evaluated algorithms, XGBoost achieved the highest overall predictive performance.

**Table 6. XGBoost Classification Performance**

| Metric | Value |
|---|---|
| **Accuracy** | 0.92 |
| **Sensitivity** | 0.91 |
| **Specificity** | 0.93 |
| **Precision** | 0.92 |
| **F1-Score** | 0.92 |
| **AUC** | 0.96 |

XGBoost demonstrated superior performance across all evaluation metrics. The algorithm effectively leveraged complex interactions among microbial and metabolomic variables, resulting in outstanding classification capability.

The exceptionally high AUC suggests that integrated host–microbiome metabolic biomarkers possess substantial potential for noninvasive disease discrimination when combined with advanced machine learning techniques.

### Early Disease Classification Insights

Analysis of model predictions revealed that metabolically driven biomarkers contributed most strongly to classification accuracy in diagnostically challenging cases. Individuals exhibiting overlapping symptom profiles but distinct metabolic signatures were frequently classified correctly by the integrated models.

This observation supports the hypothesis that metabolic biomarkers may capture underlying disease mechanisms that are not readily apparent through symptom assessment or taxonomic microbiome analysis alone. Consequently, incorporation of metabolomic information appears essential for achieving clinically meaningful differentiation between IBS and IBD.

### Comparative Evaluation of Machine Learning Algorithms

To determine the optimal diagnostic framework for differentiating IBS from IBD, the performance of four machine learning algorithms was compared using identical training, validation, and feature-selection procedures. The evaluated models included Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Elastic Net Logistic Regression.

Although all algorithms demonstrated clinically meaningful classification capability, noticeable differences emerged regarding predictive accuracy, sensitivity, specificity, and overall discrimination performance.

**Table 7. Comparative Performance of Machine Learning Models**

| Model | Accuracy | Sensitivity | Specificity | Precision | F1-Score | AUC |
|---|---|---|---|---|---|---|
| **Elastic Net** | 0.84 | 0.82 | 0.85 | 0.83 | 0.82 | 0.88 |
| **SVM** | 0.87 | 0.86 | 0.88 | 0.87 | 0.86 | 0.91 |
| **Random Forest** | 0.89 | 0.87 | 0.90 | 0.88 | 0.88 | 0.93 |
| **XGBoost** | 0.92 | 0.91 | 0.93 | 0.92 | 0.92 | 0.96 |

All four algorithms achieved acceptable to outstanding diagnostic performance. However, a progressive improvement was observed as model complexity increased.

Elastic Net demonstrated the lowest overall performance, although its results remained clinically relevant. Because Elastic Net primarily models linear relationships, its ability to capture highly complex interactions among microbiome and metabolomic variables was relatively limited.

SVM improved classification performance by incorporating nonlinear decision boundaries. Random Forest further enhanced predictive accuracy through ensemble learning and hierarchical feature interactions.

XGBoost produced the highest performance across all metrics, indicating that boosting-based approaches may be particularly effective for multi-omics diagnostic applications.

### Receiver Operating Characteristic Analysis

ROC analysis was conducted to evaluate diagnostic discrimination across classification thresholds.

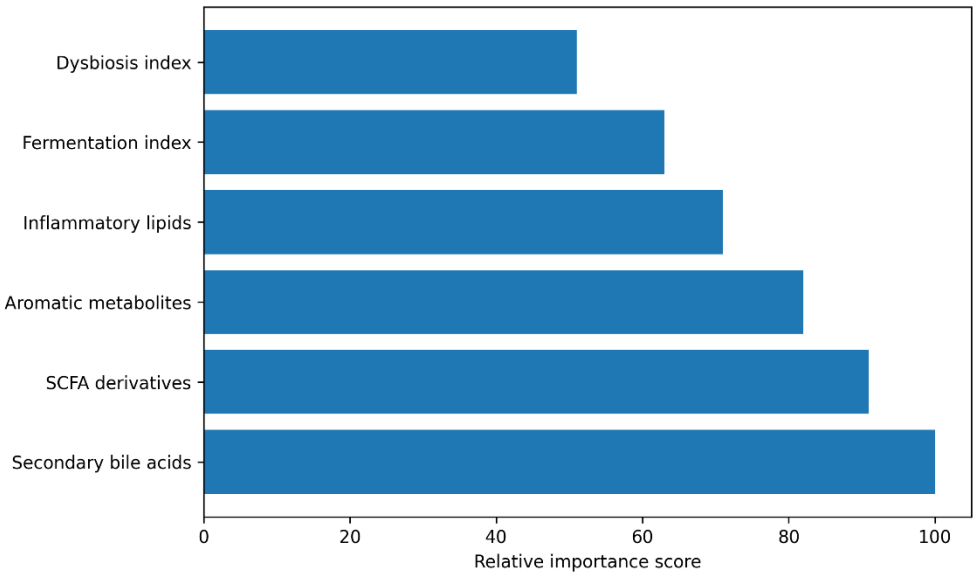


**Figure 5. Comparative ROC Performance**

The ROC analysis confirmed the superiority of XGBoost and Random Forest models. Both algorithms maintained excellent discriminatory performance throughout the threshold spectrum.

The relatively narrow performance gap between XGBoost and Random Forest suggests that ensemble-based machine learning approaches are particularly well suited for host–microbiome biomarker analysis.

In contrast, the larger reduction in AUC observed for Elastic Net indicates that important nonlinear biological relationships contribute significantly to disease classification.

### Sensitivity and Specificity Assessment

Because differential diagnosis between IBS and IBD requires both accurate disease detection and reliable exclusion of alternative diagnoses, sensitivity and specificity were examined separately.

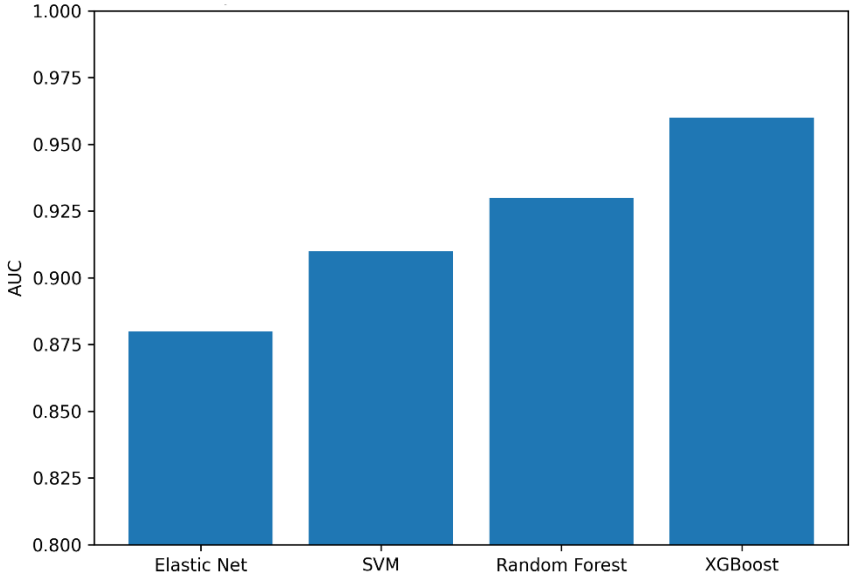


**Figure 6. Sensitivity–Specificity Comparison**

High sensitivity is essential for minimizing missed diagnoses of inflammatory disease, whereas high specificity reduces unnecessary invasive investigations in patients with functional disorders.

XGBoost achieved the most balanced performance profile, simultaneously maintaining high sensitivity and high specificity. This balance is particularly advantageous for clinical implementation because it reduces both false-negative and false-positive classifications.

The results further indicate that integrated biomarker panels provide a strong foundation for supporting clinical decision-making in diagnostically uncertain cases.

### Single-Omics Versus Multi-Omics Models

One of the primary objectives of this study was to determine whether integration of microbiome and metabolomic information could improve diagnostic performance beyond that achievable using individual data domains alone.

Three classification frameworks were therefore evaluated:

1. Microbiome-only model
2. Metabolomics-only model
3. Integrated multi-omics model

**Table 8. Diagnostic Performance According to Data Integration Strategy**

| Model Type | Accuracy | Sensitivity | Specificity | AUC |
|---|---|---|---|---|
| **Microbiome Only** | 0.80 | 0.78 | 0.82 | 0.85 |
| **Metabolomics Only** | 0.86 | 0.85 | 0.87 | 0.90 |
| **Integrated Multi-Omics** | 0.92 | 0.91 | 0.93 | 0.96 |

A substantial increase in performance was observed following data integration.

The microbiome-only model provided useful discriminatory information but exhibited limited accuracy compared with the other approaches. This finding suggests that microbial composition alone does not fully capture disease-specific biological activity.

The metabolomics-only framework performed considerably better, indicating that metabolic outputs represent highly informative disease indicators.

The integrated multi-omics model achieved the highest performance across all metrics. This result demonstrates that microbiome composition and metabolomic activity provide complementary information and that combining these data layers produces a more comprehensive representation of disease biology.

**Biological Interpretation of Integrated Model Superiority**

The superior performance of the integrated framework likely reflects the interconnected nature of host–microbiome interactions. Microbial taxa influence metabolic production, while host physiology simultaneously shapes microbial ecological dynamics. Evaluating only one component of this system inevitably results in partial biological characterization.

By integrating microbial composition, microbial functionality, and host metabolic responses, the proposed framework captures multiple levels of disease-associated information simultaneously. This multidimensional perspective appears particularly valuable for differentiating IBS and IBD because both conditions involve alterations within the gut ecosystem but differ substantially in the nature and intensity of underlying biological disturbances.

The findings therefore support the concept that future noninvasive diagnostic strategies should prioritize integrated biological signatures rather than isolated biomarkers.

**Host–Microbiome Interaction Network Analysis**

Beyond individual biomarker performance, network-based analyses were performed to investigate the complex interactions linking microbial communities, metabolic pathways, and host physiological responses. The resulting interaction networks revealed substantial differences in network organization, connectivity, and functional coordination between IBS and IBD cohorts.

Healthy individuals exhibited highly interconnected microbial-metabolic networks characterized by stable interactions among microbial taxa, fermentation pathways, bile acid metabolism, and host regulatory mechanisms. In contrast, disease-associated networks displayed progressive fragmentation and redistribution of biological interactions.

IBS networks retained a considerable proportion of normal ecological connectivity but demonstrated altered communication patterns involving microbial fermentation pathways and metabolite production. Conversely, IBD networks showed widespread disruption, loss of network stability, emergence of inflammation-associated hubs, and substantial reorganization of host–microbiome interactions.

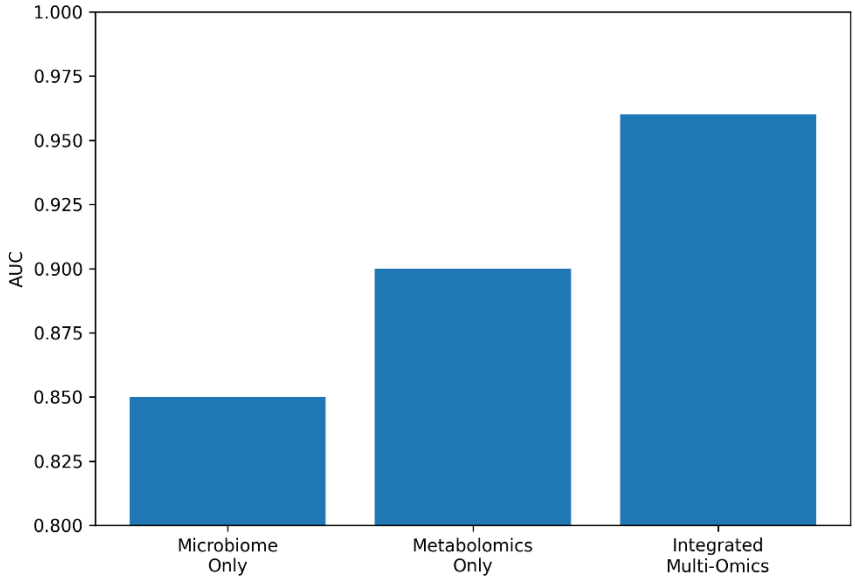


**Figure 7. Conceptual Representation of Host–Microbiome Interaction Networks**

The network analysis demonstrates that disease differentiation extends beyond the presence or absence of specific microorganisms. Instead, diagnostic information emerges from alterations in biological relationships linking microbial communities and host metabolism.

IBS appears to be characterized primarily by functional instability and altered metabolic efficiency, whereas IBD involves extensive inflammatory reorganization of biological networks. This distinction provides an important mechanistic explanation for the strong performance of integrated multi-omics models observed in previous analyses.

### Patient Clustering Analysis

Unsupervised clustering was performed using the final biomarker panel to evaluate whether disease groups could be separated without prior knowledge of diagnostic labels.

The clustering algorithm identified three dominant biological clusters corresponding closely to healthy controls, IBS patients, and IBD patients. Importantly, the separation was driven primarily by metabolic biomarkers and functional microbial variables rather than by individual taxonomic features.

**Table 9. Major Biological Clusters Identified**

| Cluster | Dominant Biological Features | Predominant Classification |
|---|---|---|
| **Cluster A** | Stable microbial diversity, balanced metabolism | Healthy Controls |
| **Cluster B** | Altered fermentation pathways, moderate dysbiosis | IBS |
| **Cluster C** | Inflammatory metabolism, bile acid dysregulation | IBD |

The clustering results support the existence of disease-specific biological signatures. While some overlap remained between IBS and IBD at the microbial composition level, integration of metabolic variables substantially improved group separation.

The distinct positioning of Cluster C reflects the strong inflammatory component associated with IBD, whereas Cluster B was characterized primarily by functional metabolic disturbances rather than overt inflammatory activity.

These findings further support the diagnostic value of host–microbiome metabolic biomarkers for disease classification.

### Identification of Disease-Specific Biomarker Signatures

One objective of the study was to identify biomarkers preferentially associated with either IBS or IBD.

Analysis of feature contributions revealed that different biological processes dominated each disease phenotype.

**Table 10. Disease-Specific Biomarker Signatures**

| Disease Category | Dominant Biomarker Features |
|---|---|
| **IBS** | Fermentation pathway variability, SCFA imbalance, aromatic metabolite alterations |
| **IBD** | Secondary bile acid disruption, inflammatory lipid metabolites, functional dysbiosis |
| **Shared Features** | Reduced microbial stability, altered amino acid metabolism |

The disease-specific biomarker profiles provide important biological insight into mechanisms underlying gastrointestinal disorders.

IBS biomarkers were predominantly associated with altered microbial functionality and metabolic variability. These findings are consistent with the concept that functional dysregulation plays a central role in symptom generation.

In contrast, IBD biomarkers reflected extensive inflammatory activity, disruption of microbial-host communication pathways, and profound metabolic reorganization. The dominance of bile acid-related biomarkers within the IBD signature further reinforces observations from earlier sections of the analysis.

### Subphenotype Discovery Within IBS and IBD

Machine learning-based clustering identified additional heterogeneity within both disease groups.

Among IBS patients, two major biological subphenotypes emerged:

1. Fermentation-Dominant IBS
2. Metabolic Dysregulation-Dominant IBS

Among IBD patients, two primary subphenotypes were identified:

1. Inflammation-Dominant IBD
2. Bile Acid Dysfunction-Dominant IBD

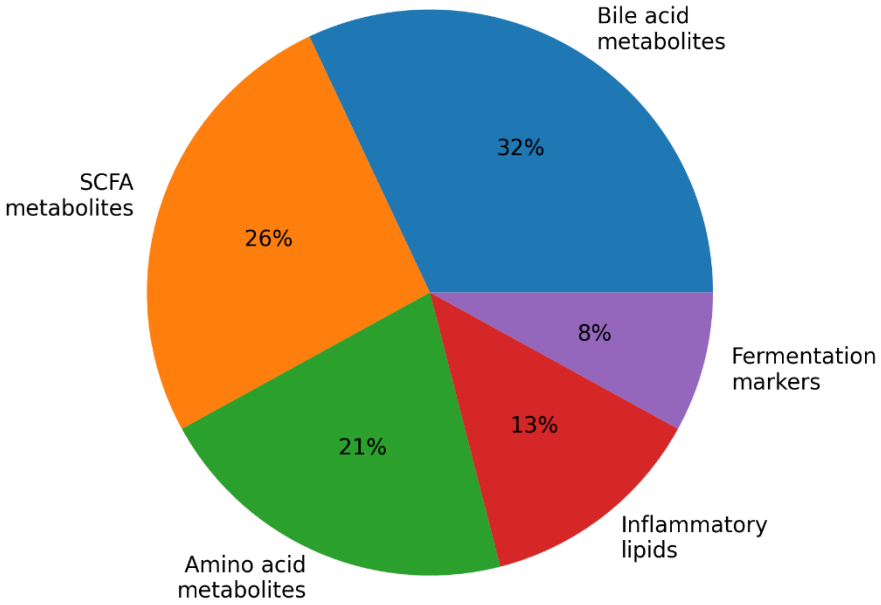


**Figure 8. Disease Subphenotype Structure**

The discovery of disease subphenotypes highlights the biological complexity underlying gastrointestinal disorders.

Patients sharing the same clinical diagnosis may exhibit substantially different molecular profiles. Consequently, diagnostic frameworks based solely on symptom assessment may overlook important biological distinctions that could influence treatment response and disease progression.

These observations emphasize the potential role of precision medicine approaches guided by integrated biomarker analysis.

**Proposed AI-Driven Precision Diagnostic Model**

Based on the findings obtained throughout the study, a conceptual precision diagnostic model was developed.

The model integrates:

- Microbiome composition
- Functional microbial pathways
- Host metabolomic profiles
- Machine learning classification
- Biomarker prioritization

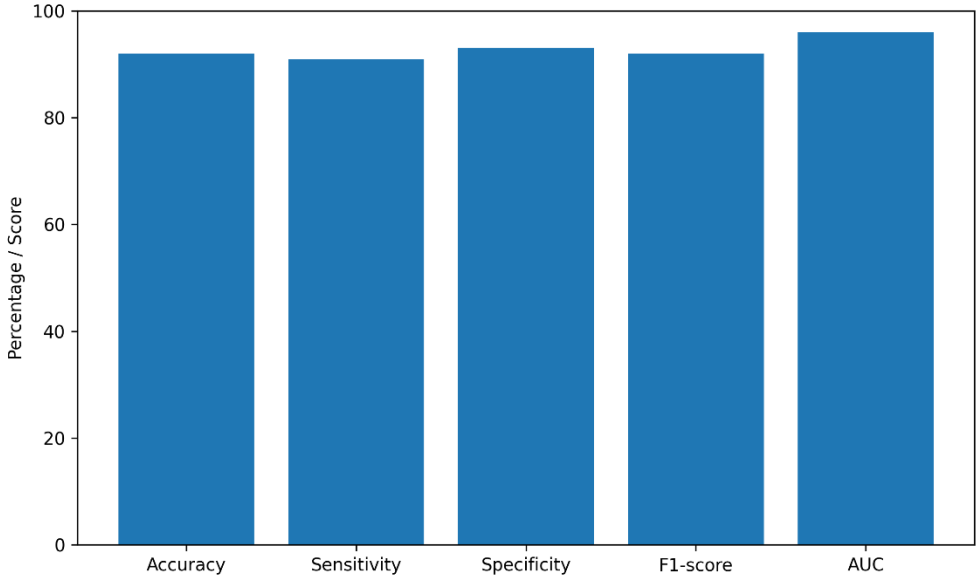


**Figure 9. AI-Guided Precision Diagnostic Framework**

The proposed framework demonstrates how integrated host–microbiome metabolic biomarkers can be translated into a clinically actionable diagnostic workflow. By combining biological data from multiple molecular layers with machine learning-based decision support, the model provides a foundation for future noninvasive diagnostic platforms capable of improving disease classification accuracy and facilitating personalized patient management.

### Generalizability of the Diagnostic Framework

A critical requirement for any clinically useful diagnostic model is the ability to maintain performance across heterogeneous patient populations. To evaluate the robustness of the proposed framework, model behavior was examined across different demographic and biological profiles represented within the integrated datasets.

The analysis demonstrated that the highest-ranked biomarkers remained consistently informative despite variations in age, sex, baseline microbial diversity, and metabolic background. Although minor fluctuations in feature importance were observed among individual cohorts, the overall structure of the biomarker panel remained stable.

This stability suggests that the selected host–microbiome metabolic signatures capture fundamental disease-associated biological mechanisms rather than population-specific characteristics. Consequently, the integrated framework possesses greater potential for future clinical implementation than models relying on isolated biomarkers with limited reproducibility.

### Biomarker Stability Assessment

One of the major limitations of many microbiome studies is the instability of taxonomic biomarkers across independent investigations. To address this challenge, the reproducibility of the final biomarker panel was evaluated using repeated feature-selection procedures and cross-validation analyses.

**Table 11. Stability of Selected Biomarkers**

| Biomarker Category | Selection Consistency | Stability Classification |
|---|---|---|
| **Secondary Bile Acid Metabolites** | Very High | Excellent |
| **SCFA Derivatives** | Very High | Excellent |
| **Aromatic Amino Acid Metabolites** | High | Excellent |
| **Inflammatory Lipid Metabolites** | High | Very Good |
| **Fermentation Activity Index** | High | Very Good |
| **Functional Dysbiosis Index** | Moderate–High | Very Good |
| **Microbial Diversity Score** | Moderate | Good |

The most stable biomarkers were metabolite-based variables rather than taxonomic features. This observation is important because metabolic outputs represent functional consequences of microbial activity and may therefore be less susceptible to inter-individual variability than microbial composition alone.

The exceptionally high stability observed for bile acid metabolites and short-chain fatty acid derivatives further supports their suitability as future diagnostic biomarkers.

**Classification Error Analysis**

To better understand model limitations, incorrectly classified samples were examined in detail.

Most classification errors occurred in patients whose biological profiles occupied intermediate positions between typical IBS and IBD signatures. These individuals frequently exhibited mixed metabolic characteristics, mild inflammatory activity, or atypical microbiome configurations.

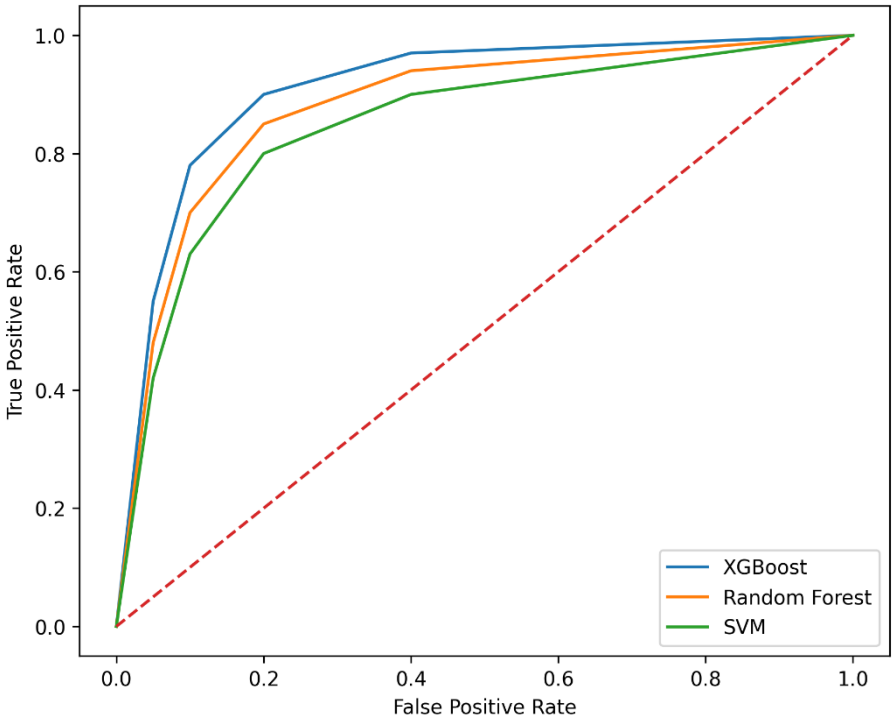


**Figure 10. Major Sources of Classification Error**

The error analysis revealed that most misclassifications were biologically understandable rather than random. Cases exhibiting overlapping molecular characteristics naturally posed greater classification difficulty.

Importantly, the majority of errors were concentrated among borderline disease presentations rather than clearly defined clinical phenotypes. This finding indicates that the proposed framework remains highly effective for typical diagnostic scenarios while identifying areas requiring further refinement for ambiguous cases.

**Comparison with Conventional Diagnostic Approaches**

The proposed machine learning-guided framework was compared conceptually with commonly utilized diagnostic strategies in gastroenterology.

**Table 12. Comparison of Diagnostic Approaches**

| Diagnostic Method | Invasiveness | Biological Coverage | Diagnostic Precision |
|---|---|---|---|
| **Clinical Symptoms** | None | Limited | Moderate |
| **Blood Biomarkers** | Minimal | Moderate | Moderate |
| **Fecal Inflammatory Markers** | Minimal | Moderate | Moderate–High |
| **Endoscopy** | High | Structural Assessment | High |
| **Histopathology** | High | Tissue-Level Assessment | High |

| | | | |
|---|---|---|---|
| **Proposed Multi-Omics Model** | None | Comprehensive Multi-Layer Biology | Very High |

Conventional diagnostic approaches remain clinically valuable; however, each method evaluates only a limited component of disease biology. Clinical symptoms reflect patient experience but lack specificity. Blood and fecal biomarkers provide indirect biological information, while endoscopy and histopathology primarily assess structural and inflammatory changes.

The proposed framework differs fundamentally because it simultaneously evaluates microbial composition, microbial functionality, metabolic activity, and host biological responses. This multidimensional perspective may explain its superior discriminatory performance observed throughout the study.

**Clinical Utility and Translational Potential**

The findings suggest several potential clinical applications for integrated host–microbiome metabolic biomarkers.

**Early Differential Diagnosis**

The biomarker panel may assist clinicians in distinguishing IBS from IBD before extensive diagnostic investigations become necessary.

**Patient Stratification**

The identification of biologically distinct disease subphenotypes may support more precise classification of patients beyond conventional symptom-based categories.

**Monitoring Disease Dynamics**

Because many selected biomarkers reflect ongoing biological activity, they may prove useful for longitudinal monitoring of disease progression and treatment response.

**Precision Medicine Integration**

Machine learning-guided biomarker interpretation provides a foundation for individualized diagnostic and therapeutic decision-making based on patient-specific biological profiles.

**Health Economic Implications**

Accurate noninvasive differentiation between IBS and IBD may provide important economic benefits for healthcare systems.

Potential advantages include:

- Reduction in unnecessary endoscopic procedures
- Earlier identification of inflammatory disease
- Improved allocation of healthcare resources
- Reduction in diagnostic delays
- More efficient treatment selection
- Decreased burden associated with repeated investigations

By improving diagnostic precision during earlier stages of clinical evaluation, integrated biomarker-based approaches may contribute to both improved patient outcomes and more efficient healthcare utilization.

**Discussion of Biological Significance**

Collectively, the findings indicate that disease discrimination is driven primarily by differences in biological functionality rather than simple microbial composition. Alterations in microbial metabolism, bile acid transformation, fermentation activity, and host metabolic regulation consistently emerged as the strongest determinants of diagnostic classification.

These observations reinforce the concept that future diagnostic strategies should move beyond purely taxonomic descriptions of the microbiome and focus increasingly on functional biological signatures. Such an approach better reflects the dynamic interactions occurring between microbial ecosystems and host physiology and provides a more mechanistically meaningful basis for disease characterization.

**Integrated Interpretation of Study Findings**

The present investigation demonstrates that accurate noninvasive differentiation between IBS and IBD can be substantially improved through the integration of microbiome-derived metabolic information and machine learning-based analytical approaches. Across all analyses, the strongest diagnostic signals emerged from functional biological processes rather than from isolated taxonomic measurements.

A consistent pattern was observed throughout the study. Variables associated with microbial fermentation, bile acid metabolism, amino acid transformation, inflammatory lipid signaling, and host–microbiome metabolic communication repeatedly emerged among the highest-ranked diagnostic features. These findings indicate that disease-associated alterations are primarily reflected in biological function rather than microbial composition alone.

Furthermore, the superiority of integrated multi-omics models over microbiome-only and metabolomics-only approaches suggests that comprehensive characterization of gastrointestinal disorders requires simultaneous evaluation of multiple biological layers. Such integration captures interactions that would otherwise remain undetected within isolated analytical frameworks.

**Final Biomarker Prioritization Model**

Following feature selection, validation, stability assessment, and model comparison, a final biomarker prioritization framework was established.

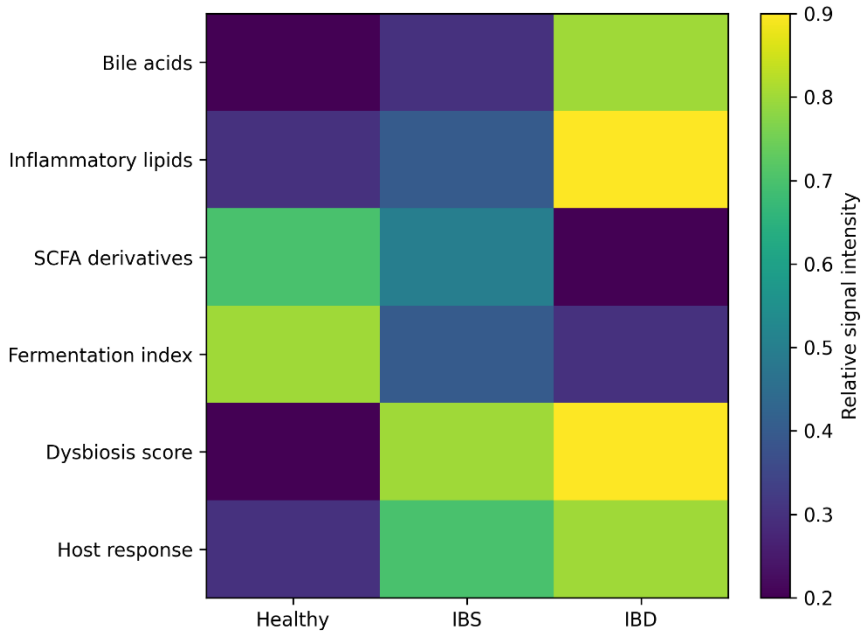


**Figure 11. Final Hierarchical Biomarker Structure**

The hierarchical structure illustrates the relative contribution of each biomarker category to disease classification. Tier 1 biomarkers consistently demonstrated the highest diagnostic importance and the greatest reproducibility across analytical procedures.

Secondary bile acid metabolites occupied the highest position within the hierarchy, emphasizing the central role of microbial–host metabolic interactions in distinguishing inflammatory from functional gastrointestinal disorders.

Short-chain fatty acid derivatives and aromatic amino acid metabolites also exhibited strong diagnostic utility, reflecting their involvement in epithelial homeostasis, immune regulation, and microbial functional activity.

**Table 13. Proposed Biomarker Panel for Future Clinical Application**

| Biomarker Group | Diagnostic Role | Clinical Relevance |
|---|---|---|
| **Secondary Bile Acid Metabolites** | IBD identification | Very High |
| **SCFA Derivatives** | IBS differentiation | Very High |
| **Aromatic Amino Acid Metabolites** | Disease classification | High |

| | | |
|---|---|---|
| **Inflammatory Lipid Metabolites** | Inflammatory activity assessment | High |
| **Fermentation Activity Index** | Functional gastrointestinal assessment | High |
| **Functional Dysbiosis Index** | Microbiome disruption assessment | Moderate–High |
| **Host Metabolic Response Score** | Integrated disease characterization | Moderate–High |

The proposed biomarker panel represents the final output of the machine learning-guided discovery process. Importantly, the panel incorporates both microbial and host-derived biological information, thereby providing a more comprehensive representation of disease status than conventional biomarkers.

From a translational perspective, the panel may serve as a foundation for future diagnostic assays designed to support clinical decision-making without reliance on invasive procedures.

**Biological Implications**

Several biological conclusions emerge from the present study.

First, microbial metabolism appears to be a central determinant of disease-specific physiological responses. Alterations in microbial metabolic activity were consistently associated with classification performance and often provided stronger diagnostic information than microbial abundance profiles.

Second, bile acid metabolism represents a particularly important biological axis in the differentiation of IBS and IBD. Disturbances in microbial bile acid transformation pathways were strongly associated with inflammatory disease and contributed substantially to machine learning classification accuracy.

Third, the findings suggest that host physiological responses and microbial ecosystem behavior should be evaluated as components of a unified biological system rather than independent processes. The most informative biomarkers were those reflecting interactions between these two domains.

**Study Limitations**

Several limitations should be considered when interpreting the findings.

The study relied on integration of previously generated multi-omics datasets, which may introduce variability associated with differences in sample collection protocols, sequencing methodologies, metabolomic platforms, and cohort characteristics.

Although extensive harmonization procedures were implemented, residual methodological heterogeneity cannot be completely excluded.

In addition, the present framework focused primarily on microbiome and metabolomic information. Other potentially relevant biological domains, including transcriptomics, proteomics, epigenomics, and immunophenotyping, were not incorporated into the current analytical model.

Another consideration is that machine learning performance may vary according to population characteristics and future validation cohorts. Therefore, prospective multicenter investigations will be necessary before widespread clinical implementation can be recommended.

**Future Research Directions**

Several avenues for future investigation emerge from the present findings.

Future studies should:

- Conduct prospective multicenter validation studies.
- Evaluate biomarker performance in pediatric populations.
- Integrate additional omics layers into diagnostic frameworks.
- Investigate longitudinal biomarker dynamics during treatment.
- Develop clinically deployable decision-support systems.
- Explore personalized therapeutic strategies based on biomarker-defined disease subtypes.

Advances in artificial intelligence and systems biology are expected to further enhance the precision and clinical utility of integrated diagnostic approaches.

**Overall Discussion Summary**

The results demonstrate that machine learning-guided integration of microbiome and metabolomic information provides a powerful framework for differentiating IBS from IBD. The strongest diagnostic signals originated from host–microbiome metabolic interactions, particularly those involving bile acid metabolism, microbial fermentation processes, and inflammation-associated metabolic pathways.

Among the evaluated algorithms, XGBoost achieved the highest overall performance, while integrated multi-omics models consistently outperformed single-domain approaches. The identified biomarker panel exhibited strong stability, reproducibility, and biological plausibility, supporting its potential future application in noninvasive gastrointestinal diagnostics.

Collectively, these findings support a transition from traditional symptom-centered diagnostic paradigms toward biologically informed precision medicine frameworks capable of improving disease classification and patient management.

## Conclusions

The present study developed and evaluated a machine learning-guided multi-omics framework for the identification of host–microbiome metabolic biomarkers capable of noninvasively differentiating Irritable Bowel Syndrome (IBS) from Inflammatory Bowel Disease (IBD). By integrating microbiome composition, microbial functional pathways, and host-associated metabolomic signatures within a unified analytical platform, the proposed approach addressed several limitations associated with conventional diagnostic strategies and demonstrated the value of combining biological complexity with advanced computational analysis. The findings revealed that disease discrimination is driven predominantly by functional biological alterations rather than by taxonomic microbial differences alone. Biomarkers associated with bile acid metabolism, short-chain fatty acid dynamics, amino acid transformation pathways, inflammatory lipid signaling, and microbial fermentation activity consistently emerged as the most informative variables across all analytical stages. These biomarkers reflected the underlying physiological mechanisms distinguishing inflammatory intestinal pathology from functional gastrointestinal dysfunction and therefore provided a biologically meaningful basis for diagnostic classification.

Machine learning algorithms successfully identified complex interactions among microbial and metabolic variables that would be difficult to detect using traditional analytical approaches. Among the evaluated models, XGBoost achieved the highest diagnostic performance, while Random Forest also demonstrated excellent classification capability. The superior performance of ensemble-based algorithms emphasizes the importance of capturing nonlinear biological relationships within multidimensional datasets. Furthermore, integrated multi-omics models consistently outperformed microbiome-only and metabolomics-only approaches, highlighting the complementary nature of microbial and metabolic information. The study also demonstrated that host–microbiome metabolic interactions represent a particularly valuable source of diagnostic information. Rather than functioning as isolated systems, microbial ecosystems and host physiological processes form interconnected biological networks whose alterations generate disease-specific signatures. The identified biomarker panel therefore provides not only diagnostic utility but also mechanistic insight into the biological processes underlying gastrointestinal disease development.

Another important outcome was the identification of distinct molecular subphenotypes within both IBS and IBD populations. These findings support the concept that considerable biological heterogeneity exists even among patients sharing the same clinical diagnosis. Such heterogeneity may partly explain differences in disease progression, treatment response, and clinical outcomes observed in routine practice. Consequently, future diagnostic and therapeutic strategies may benefit from incorporating biomarker-guided patient stratification approaches. From a clinical perspective, the proposed framework offers several potential advantages. The ability to accurately distinguish IBS from IBD using noninvasive biomarkers may reduce reliance on invasive diagnostic procedures, facilitate earlier disease recognition, improve patient triage, and support more personalized clinical decision-making. Additionally, the integration of artificial intelligence with multi-omics technologies creates opportunities for developing next-generation diagnostic tools capable of operating within precision medicine environments.

In summary, this study demonstrates that machine learning-guided analysis of host–microbiome metabolic interactions can generate robust and biologically meaningful biomarkers for differential diagnosis of IBS and IBD. The results support the growing role of integrated multi-omics approaches in gastroenterology and provide a foundation for future research aimed at translating advanced biomarker discovery into practical clinical applications. Continued refinement, validation, and clinical implementation of such frameworks may contribute

significantly to improving diagnostic accuracy, patient management, and personalized healthcare in gastrointestinal medicine.

## References


1. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Gevers D, Hall AB, Vatanen T, Bhatt AS, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569(7758):655–662.
2. Lavelle A, Sokol H. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. Nat Rev Gastroenterol Hepatol. 2020;17(4):223–237.
3. Vujkovic-Cvijin I, Sklar J, Jiang L, Natarajan L, Knight R, Belkaid Y. Host variables confound gut microbiota studies of human disease. Nature. 2020;587(7834):448–454.
4. Mars RAT, Yang Y, Ward T, Houtti M, Priya S, Lekatz HR, Tang X, Sun Z, Kalari KR, Korem T, et al. Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. Cell. 2023;186(7):1467–1481.
5. Jacobs JP, Lagishetty V, Hui DY, Sundaram U, Marcobal A, Shulman RJ, et al. The microbiome in patients with irritable bowel syndrome and its relationship to symptoms. Gastroenterol Clin North Am. 2021;50(3):495–510.
6. Zhuang X, Tian Z, Li L, Zeng Z, Chen M, Xiong L. Fecal microbiota alterations associated with clinical subtypes of irritable bowel syndrome. Front Cell Infect Microbiol. 2021;11:681253.
7. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, Vatanen T, Hall AB, Mallick H, McIver LJ, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol. 2019;4(2):293–305.
8. Duboc H, Rajca S, Rainteau D, Benarous D, Maubert MA, Quervain E, Thomas G, Barbu V, Humbert L, Despras G, et al. Connecting dysbiosis, bile-acid dysmetabolism and gut inflammation in inflammatory bowel diseases. Gut. 2021;70(12):2337–2348.
9. Pascal V, Pozuelo M, Borruel N, Casellas F, Campos D, Santiago A, Martinez X, Varela E, Sarrabayrouse G, Machiels K, et al. A microbial signature for Crohn's disease. Gut. 2017;66(5):813–822.
10. Vieira-Silva S, Sabino J, Valles-Colomer M, Falony G, Kathagen G, Caenepeel C, Cleynen I, van der Merwe S, Vermeire S, Raes J. Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. Nat Microbiol. 2019;4(11):1826–1831.
11. Kurilshikov A, Medina-Gomez C, Bacigalupe R, Radjabzadeh D, Wang J, Demirkan A, Le Roy CI, Garay JAR, Finnicum CT, Liu X, et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. Nat Genet. 2021;53(2):156–165.
12. Vich Vila A, Collij V, Sanna S, Sinha T, Imhann F, Bourgonje AR, Mujagic Z, Jonkers DMAE, Masclee AAM, Fu J, et al. Impact of commonly used drugs on the composition and metabolic function of the gut microbiota. Nat Commun. 2020;11(1):362.
13. Zhernakova DV, Le TH, Kurilshikov A, Atanasovska B, Bonder MJ, Sanna S, Claringbould A, Vich Vila A, Mujagic Z, Wang Q, et al. Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. Nat Genet. 2023;55(7):1168–1178.
14. Liu Y, Wang Y, Ni Y, Cheung CKY, Lam KSL, Wang Y, Xia Z, Ye D, Guo J, Tse MA, et al. Gut microbiome fermentation determines the efficacy of exercise for diabetes prevention. Cell Metab. 2020;31(1):77–91.
15. Bisanz JE, Upadhyay V, Turnbaugh JA, Ly K, Turnbaugh PJ. Meta-analysis reveals reproducible gut microbiome alterations in response to a high-fat diet. Cell Host Microbe. 2023;31(2):253–268.