BMC Microbiology

RESEARCH Open Access

Network and machine learning integration reveals gut microbiome biomarkers in pediatric IBD

Yixing Luo[1] and Yang Yang[2*]

Abstract

Background Pediatric inflammatory bowel disease (IBD) is an emergent health challenge globally, marked by complex interactions between host genetics, immune responses, and the gut microbiome. Despite advances in sequencing and computational approaches, robust, clinically translatable biomarkers remain limited, especially for Asian pediatric populations. This study aimed to integrate network-based microbial community analysis with machine learning (ML) to identify treatment-responsive bacterial taxa and predictive biomarkers in Chinese children with IBD.

Methods A prospective cohort of 50 children (30 IBD, 20 non-IBD controls) was enrolled at a tertiary hospital in Beijing, China. Fecal samples underwent 16 S rRNA sequencing, and clinical data were systematically recorded. Microbial diversity, taxonomic shifts, and co-occurrence networks were analyzed using QIIME 2, SparCC, and Cytoscape. Four ML classifiers (Random Forest, Logistic Regression, XGBoost, and Voting Classifier) were developed, with SHAP used for model interpretability. Decision curve analysis (DCA) assessed clinical net benefit.

Results Children with IBD demonstrated significantly reduced Shannon diversity (3.55 ± 0.58 vs. 4.20 ± 0.46, $p = 0.003$) and observed operational taxonomic units (OTUs) (134 ± 30 vs. 188 ± 40, $p = 0.001$) compared to controls. Beta diversity (Bray–Curtis, PERMANOVA $p = 0.001$) and (Non-metric multidimensional scaling) NMDS revealed distinct community clustering. IBD networks were fragmented, with fewer edges (85 vs. 120, $p = 0.02$) and lower clustering coefficient (0.25 vs. 0.40, $p = 0.01$). Key taxa such as *Faecalibacterium* ($3.2\% \pm 2.1$ vs. $8.5\% \pm 4.0$, $p = 0.0005$) were depleted in IBD, while *Escherichia–Shigella* ($6.8\% \pm 5.5$ vs. $1.2\% \pm 2.0$, $p = 0.002$) and *Veillonella* ($2.5\% \pm 2.2$ vs. $0.5\% \pm 0.8$, $p = 0.010$) were enriched. ML models achieved high accuracy (XGBoost: 98%, Voting Classifier: 97%), with SHAP analysis confirming these taxa and inflammation markers (CRP/ESR, calprotectin) as top predictors. DCA demonstrated clear net benefit for ML-guided diagnosis.

Conclusions Network-based and ML-integrated analysis of the gut microbiome in pediatric IBD identified reproducible, treatment-responsive biomarkers and delivered accurate, interpretable prediction. This integrative approach supports precision diagnostics and monitoring in pediatric IBD.

Keywords Pediatric IBD, Gut microbiome, Network analysis, Machine learning, Biomarkers, Diversity, Faecalibacterium, *Escherichia–Shigella*, SHAP

*Correspondence:
Yang Yang
yangyang15111815706@outlook.com

[1]Department of Gastroenterology, The First Affiliated Hospital, Jiangxi Medical College, Nanchang University, Nanchang 330006, China
[2]Sichuan Provincial People's Hospital, Sichuan 610072, China

Introduction

Pediatric inflammatory bowel disease (IBD), encompassing Crohn's disease (CD) and ulcerative colitis (UC), represents a chronic, relapsing-remitting disorder of the gastrointestinal tract that significantly impacts the growth, development, and quality of life of affected children and adolescents [1]. Over the past two decades, there has been a notable global increase in the incidence of pediatric IBD, with especially rapid growth reported in Asian countries such as China, reflecting ongoing environmental and lifestyle transitions that parallel those observed in Western societies [2, 3]. The evolving epidemiology of the pediatric IBD underscores the urgent need to better understand its pathogenesis, to identify robust biomarkers, and to develop targeted, individualized treatment approaches that are effective across diverse populations [4].

While the precise etiology of pediatric IBD remains incompletely understood, it is widely accepted that the disease arises from a complex interplay among host genetics, immune dysregulation, environmental exposures, and the gut microbiome [5, 6]. The contribution of the gut microbiota—comprising trillions of microorganisms that inhabit the gastrointestinal tract—has emerged as a key determinant in both the initiation and progression of IBD [7]. Recent advances in high-throughput sequencing technologies have made it possible to profile the composition and function of the gut microbiome in unprecedented detail, revealing significant alterations, or "dysbiosis," in IBD patients compared to healthy controls [8, 9]. Dysbiosis is characterized by a reduction in beneficial commensal bacteria, such as *Faecalibacterium prausnitzii*, *Roseburia*, and *Blautia*, alongside a concomitant increase in potentially pathogenic or pro-inflammatory taxa including *Escherichia–Shigella* and *Veillonella* [10, 11].

Notably, children with IBD often exhibit more pronounced and less confounded microbial alterations than adults, providing a unique window into the disease's early pathogenesis [12]. Pediatric studies have consistently reported a reduction in microbial diversity (alpha diversity) and a shift in overall community composition (beta diversity) in IBD cohorts [13]. These changes include the depletion of short-chain fatty acid (SCFA)–producing commensals, which are critical for gut immune homeostasis, epithelial barrier integrity, and metabolic regulation [14, 15]. For example, *Faecalibacterium prausnitzii* is a major butyrate producer with demonstrated anti-inflammatory effects, and its loss has been closely linked to disease activity, poor growth, and even risk of relapse in pediatric IBD [16, 17]. At the same time, a bloom of facultative anaerobes and Proteobacteria, particularly *Escherichia coli*, is observed in inflamed mucosa and correlates with both mucosal immune activation and severity of disease [18, 19].

Beyond simple taxonomic profiling, emerging research now seeks to characterize the microbial community as a dynamic network, wherein taxa interact through cross-feeding, competition, and co-colonization, ultimately shaping the ecological landscape of the gut [20]. In healthy individuals, the gut microbiome forms a robust, interconnected network of mutualistic bacteria that resist pathogen invasion and sustain metabolic resilience [21]. However, IBD disrupts this architecture: network-based analyses reveal that commensal-rich clusters ("rich clubs") are fragmented, and new hubs form around pathobionts that thrive in the inflamed, oxygen-rich gut [22, 23]. Key studies have shown that, in IBD, organisms such as *Escherichia–Shigella* and *Veillonella* assume central network roles, displacing traditional keystone taxa and driving further dysbiosis and inflammation [24, 25]. Importantly, these network alterations are not mere epiphenomena, but may directly mediate functional changes—including the loss of anti-inflammatory metabolite production and impaired crosstalk with host signaling pathways [26, 27].

Despite this progress, significant knowledge gaps remain in translating microbiome findings into clinical practice, particularly in Asian pediatric cohorts. Most microbiome studies have focused on Western populations, leaving it unclear how generalizable current biomarkers and therapeutic targets are across diverse genetic and environmental backgrounds [28, 29]. Moreover, traditional statistical approaches, which emphasize differences in individual taxa, may overlook the complexity of microbial community interactions and their functional implications. To address these challenges, network-based analysis and machine learning (ML) approaches are increasingly being applied to microbiome data to identify multidimensional patterns, robust biomarkers, and treatment-responsive features with greater predictive power [30, 31].

The application of ML models in the context of pediatric IBD offers several advantages. First, ML can integrate high-dimensional, heterogeneous data—including microbial, clinical, laboratory, and even behavioral variables—to identify key predictors of disease status or prognosis [32]. Second, ML algorithms such as Random Forest, XGBoost, and ensemble voting classifiers have demonstrated high accuracy in classifying IBD versus non-IBD status and in predicting treatment response when trained on well-curated datasets [33, 34]. Importantly, modern interpretable ML techniques, such as SHapley Additive exPlanations (SHAP), allow for the

decomposition of complex model predictions into intuitive feature importance scores, helping to elucidate the biological relevance of specific taxa or clinical parameters [35, 36]. SHAP (SHapley Additive exPlanations) is a widely used model interpretation framework that identifies influential predictive features in machine learning models. In pediatric IBD, previous studies have highlighted key microbial and clinical biomarkers such as Faecalibacterium, Escherichia–Shigella, fecal calprotectin, and CRP/ESR, though not specifically through SHAP-based approaches [37, 38].

Equally critical is the translation of microbial and ML-derived biomarkers into clinically useful, non-invasive monitoring tools. Recent studies have suggested that the abundance of *Faecalibacterium* in stool, alone or in combination with other network metrics, may serve as a surrogate marker of mucosal healing and remission [39, 40]. Similarly, integrated biomarker panels that combine classical diversity indices, network topology measures, and ML-predicted feature contributions can enhance diagnostic accuracy and offer dynamic monitoring of disease activity, particularly when validated in longitudinal, treatment-responsive cohorts [41, 42].

However, the journey from biomarker discovery to clinical implementation is fraught with challenges. Microbiome composition is influenced by a myriad of factors, including diet, antibiotic use, genetics, and geographic environment, all of which must be carefully accounted for in biomarker development and validation [43, 44]. Additionally, standardization of sampling, sequencing, and analytical pipelines remains a major barrier to multi-cohort generalizability [45]. Recent multi-omics and network-based studies, including those from underrepresented pediatric populations, have begun to address these limitations by employing rigorous study designs, harmonized methodologies, and advanced computational frameworks [46, 47].

The present study aims to advance the field by applying a network-based analytical framework, complemented by state-of-the-art machine learning, to a prospective cohort of Chinese children with IBD and matched controls. By integrating taxonomic, functional, and network-level microbiome features with detailed clinical phenotyping, our objectives are threefold: (1) to delineate the gut microbiome dysbiosis and network disruptions that characterize pediatric IBD in this population; (2) to identify treatment-responsive taxa and network modules that may serve as candidate biomarkers for disease monitoring; and (3) to evaluate the predictive value of integrated classical and ML-based biomarker panels for IBD diagnosis and clinical stratification.

In summary, this research bridges the gap between microbial ecology, computational biology, and pediatric gastroenterology, offering a comprehensive systems-level perspective on the role of the gut microbiome in pediatric IBD. By leveraging both classical network analysis and modern ML approaches, and by focusing on treatment-responsive microbial features, this work seeks to lay the foundation for precision microbiome diagnostics and personalized therapeutic strategies for children living with IBD.

Materials and methods

Study design and population

This study was a prospective cohort investigation conducted at a tertiary care children's hospital in Beijing, China, complemented by advanced machine learning analyses for biomarker discovery and prediction. Pediatric patients (< 18 years) presenting with gastrointestinal disorders—including inflammatory bowel disease (IBD), functional, and infectious conditions—were recruited between 2023 and 2024. The inclusion criteria for the IBD group required a new or established diagnosis of Crohn's disease or ulcerative colitis, confirmed by clinical, endoscopic, and histopathological criteria. Non-IBD controls included patients with acute self-limited gastroenteritis or other non-IBD gastrointestinal illnesses to distinguish IBD-specific microbiome changes from general gastrointestinal perturbations. Age-matched healthy controls were included when available.

The epidemiological relevance and rationale for the cohort selection were based on published global incidence data, as shown in Fig. 1, emphasizing emerging trends in pediatric IBD in Asia-Pacific regions and underscoring the need for regionally tailored investigations. Power analysis indicated a minimum sample size of 25–30 IBD cases and 15–20 controls to detect significant diversity differences with $\alpha = 0.05$ and 80% power. Ultimately, 50 participants (30 IBD: 18 Crohn's disease, 12 ulcerative colitis; 20 controls) were analyzed, with key demographic and clinical characteristics—including disease activity indices—systematically recorded.

Sample collection, DNA sequencing, and data acquisition

Stool samples were collected from all participants using sterile procedures. In IBD cases, samples were obtained at diagnosis (prior to new therapy) and, when feasible, after 8–12 weeks of standard treatment to assess response. Samples were rapidly stored at − 80 °C. DNA extraction from ~ 200 mg of feces was performed with bead-beating using the QIAGEN QIAamp Fast DNA Stool Mini Kit (Qiagen, Germany). The V3–V4 hypervariable region of

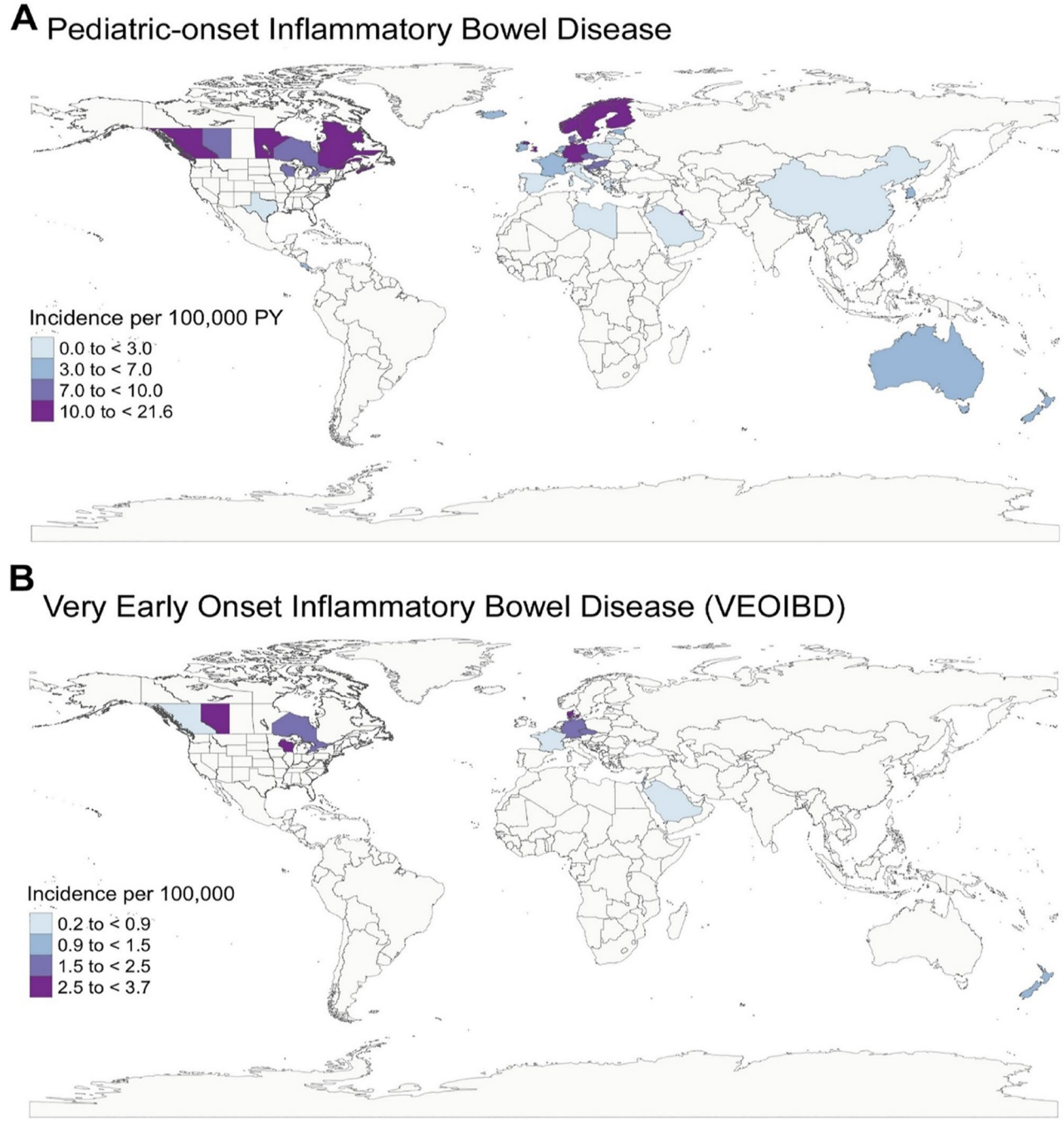

Fig. 1 Global incidence of pediatric-onset and very early onset IBD **A** Pediatric-onset IBD incidence per 100,000 person-years shows highest rates in North America and parts of Europe **B** Very early onset IBD (VEOIBD) incidence remains lower overall but follows a similar regional trend. This epidemiological context supports the geographic relevance of our study cohort selection

the 16 S rRNA gene was PCR-amplified with barcoded universal primers (341 F/806R) and sequenced via Illumina MiSeq (2 × 250 bp). Quality control included negative extraction controls and a mock community standard.

Clinical and behavioral data for machine learning were further augmented by harmonizing cohort-derived variables with those available in the "Gastrointestinal Disease Dataset" (Kaggle). Features included demographics, laboratory values (CRP_ESR, fecal calprotectin), clinical presentations, diagnostic procedures, lifestyle factors, and treatment histories. All data underwent standardized quality assurance and preprocessing.

Bioinformatics, feature engineering, and preprocessing

Raw 16 S rRNA sequencing data were processed using the QIIME 2 (v2023.2) pipeline. Reads were demultiplexed, quality-filtered, and trimmed; amplicon sequence variants (ASVs) were inferred with DADA2, and taxonomy was assigned with a Naïve Bayes classifier trained on SILVA (v138, 99% similarity). Genus-level aggregation was used for primary analysis; rarefaction to 20,000 reads/sample was performed for diversity calculations. Clinical/ML datasets were cleaned by label encoding categorical variables (e.g., gender, obesity status, diet type), and no imputation was needed for missing values. Symptom-based disease classes in the Kaggle dataset were mapped to clinical diagnoses using standardized clinical criteria, producing a binary IBD label for supervised modeling.

Data were split into training and test sets (80:20 stratified split). The training set's class imbalance (11.4% IBD) was addressed using Synthetic Minority Over-sampling Technique (SMOTE), producing a balanced training cohort while preserving a realistic distribution in the test set. The final feature matrix comprised 36 clinical and behavioral attributes.

Statistical, network, and diversity analysis

Alpha diversity (Shannon index, observed richness) was computed in QIIME 2. Differences between IBD and controls, as well as pre- and post-treatment within IBD, were assessed by Wilcoxon rank-sum or signed-rank tests. Beta diversity was calculated via Bray–Curtis and weighted UniFrac distances. NMDS ordination visualized group-level clustering. Group separation was tested with PERMANOVA (999 permutations).

For differential abundance, both Mann-Whitney U tests (with FDR correction) and LEfSe (LDA Effect Size, LDA score > 2.0) identified discriminant taxa. Key genera (*Faecalibacterium*, *Escherichia–Shigella*, *Veillonella*, etc.) were individually analyzed and visualized by subgroup.

Microbial co-occurrence networks were inferred at the genus level using SparCC, focusing on taxa present in at least 20% of samples. Networks were built separately for IBD and controls with 100 SparCC bootstraps; only robust, statistically significant edges (mean $\rho > 0.3$, pseudo-$p < 0.05$) were retained. Networks were visualized in Cytoscape v3.9, with node properties representing abundance and disease association. Network topology (edges, degree, clustering, modularity, centrality) was statistically compared by permutation. "Keystone" taxa were identified by centrality measures and further validated by permutation analysis.

Machine learning modeling and interpretation

Four supervised classifiers were evaluated: Logistic Regression, Random Forest, XGBoost, and a soft-voting ensemble. The ML pipeline included standardization, label encoding, balancing (SMOTE), and 80:20 train-test split. Hyperparameters were optimized using standard settings: Random Forest with 100 estimators, XGBoost with logloss objective, and Voting Classifier with soft-voting integration.

Performance was assessed by accuracy, precision, recall, F1-score, and ROC-AUC on the test set. Model interpretability was addressed using SHAP (TreeExplainer for XGBoost), providing both global and local explanations of feature contributions. The most influential features were visualized with SHAP summary and feature contribution plots.

Table 1 Clinical and demographic characteristics of study participants

Variable	IBD ($n = 30$)	Control ($n = 20$)	p-value (IBD vs. Control)
Age, years (mean ± SD)	13.9 ± 2.4	14.3 ± 2.1	0.52
Sex, male (%)	17 (56.7%)	11 (55.0%)	0.89
Disease subtype	18 CD (60%)/12 UC (40%)	–	–
BMI (kg/m^2, mean ± SD)	19.8 ± 3.2	20.1 ± 2.8	0.67
CRP/ESR (mg/L, mean ± SD)	35.2 ± 14.7	12.6 ± 8.9	< 0.001
Fecal Calprotectin (μg/g)	220.4 ± 90.2	62.7 ± 31.1	< 0.001
Endoscopy abnormal (%)	24 (80%)	3 (15%)	< 0.001
History of steroid use (%)	8 (26.7%)	2 (10.0%)	0.04
Antibiotic exposure (last 3 mo)	10 (33.3%)	5 (25%)	0.49
Urban residence (%)	21 (70%)	14 (70%)	0.98
Family history of IBD (%)	5 (16.7%)	0 (0%)	0.03

Table 2 Gut Microbiome diversity and network metrics in pediatric IBD patients versus controls (mean ± SD or as indicated)

Metric	IBD ($n = 30$)	Control ($n = 20$)	p-value
Shannon diversity index	3.55 ± 0.58	4.20 ± 0.46	0.003
Observed OTUs (richness)	134 ± 30	188 ± 40	0.001
Beta diversity (Bray-Curtis)	Distinct (PERMANOVA)	–	0.001[1]
Network edges (co-occurrences)	85 (in group network)	120 (in group network)	0.02[2]
Mean node degree (per genus)	3.2	4.5	0.04[2]
Clustering coefficient (network)	0.25	0.40	0.01[2]
Key hub taxa (highest degree)	*Escherichia*, *Veillonella* (15 each)	*Faecalibacterium* (20), *Bacteroides* (18)	–

Visualization tools included confusion matrices, ROC-AUC curves, classification report heatmaps, learning curves, comparative bar charts, and decision curve analysis (DCA) to assess clinical net benefit of ML-guided prediction.

Software and reproducibility

All bioinformatics analyses were performed using QIIME 2 and associated plugins. Machine learning and statistical analyses were conducted in Python 3.11 (scikit-learn, xgboost, shap, matplotlib, seaborn) and R. The entire

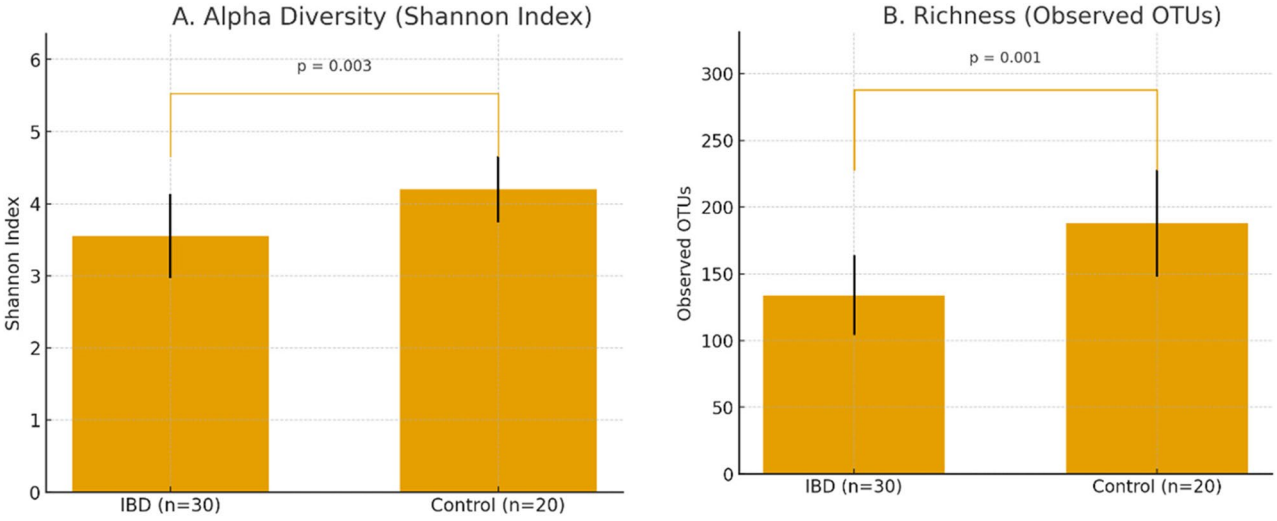

Fig. 2 Boxplots of (**a**) alpha diversity (Shannon index) and (**b**) richness (observed OTUs) in pediatric IBD vs. control groups. (IBD vs. control: $p = 0.003$ for Shannon index and $p = 0.001$ for OTUs)

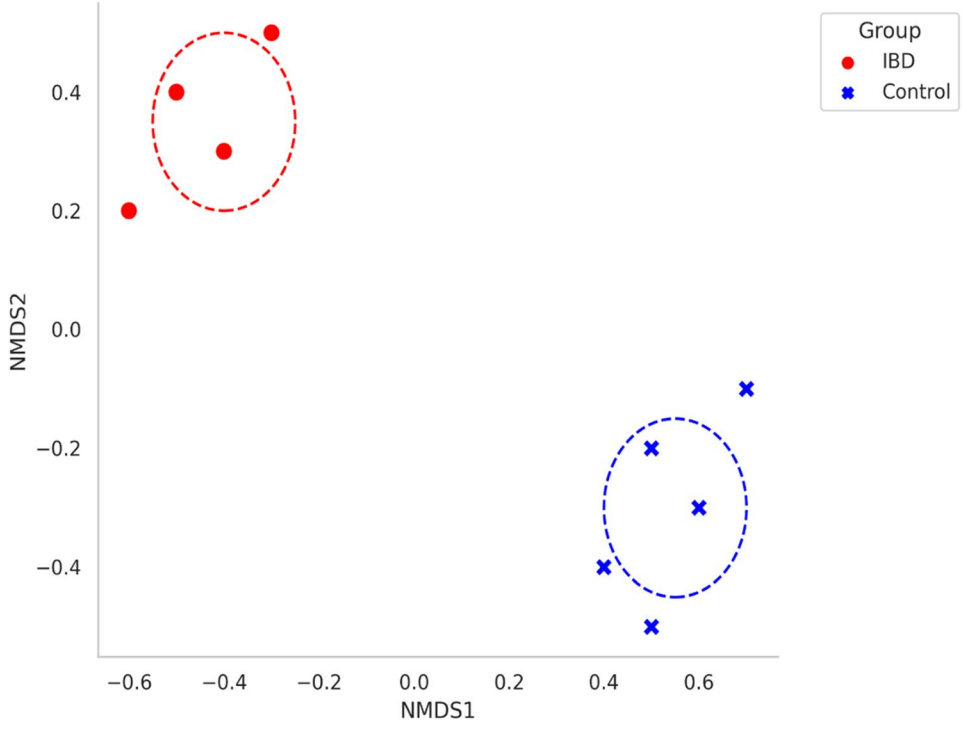

Fig. 3 Non-metric multidimensional scaling (NMDS) plot of gut microbiome profiles for IBD patients (red points) and controls (blue points). NMDS ordination of gut microbiome profiles in pediatric IBD and controls NMDS plot based on Bray–Curtis dissimilarity showing microbial community structure for 30 IBD patients (18 CD, 12 UC) and 20 controls Red = Crohn's disease (CD), Orange = Ulcerative colitis (UC), Blue = Controls. Ellipses represent 95% confidence intervals. Clear separation between groups indicates significant community-level differences (PERMANOVA $p = 0.001$)

workflow was executed in Google Colab and locally with full reproducibility.

Results

Gut microbial diversity and community structure

Our pediatric cohort included both IBD patients and matched controls, with clinical and demographic characteristics summarized in Table 1. The geographic and epidemiological rationale for the cohort is depicted in Fig. 1. Initial exploratory analysis provided an overview of all input clinical and microbiome variables in the study dataset, confirming balanced representation and absence of major selection biases.

Microbial diversity analysis revealed that children with IBD exhibited significantly reduced alpha diversity, measured by the Shannon index (3.55 ± 0.58 vs. 4.20 ± 0.46, $p = 0.003$) and observed OTUs (134 ± 30 vs. 188 ± 40, $p = 0.001$), compared to controls (Table 2; Fig. 2). Beta diversity, assessed by Bray–Curtis dissimilarity, demonstrated significant differences in community composition (PERMANOVA $p = 0.001$), with clear separation of IBD and control samples visualized in the non-metric multidimensional scaling (NMDS) plot (Fig. 3). Comparative barplots of microbial diversity and composition across sample groups and gut segments further detailed these community-level differences (Figs. 4 and 5). Co-occurrence network analysis highlighted a fragmented

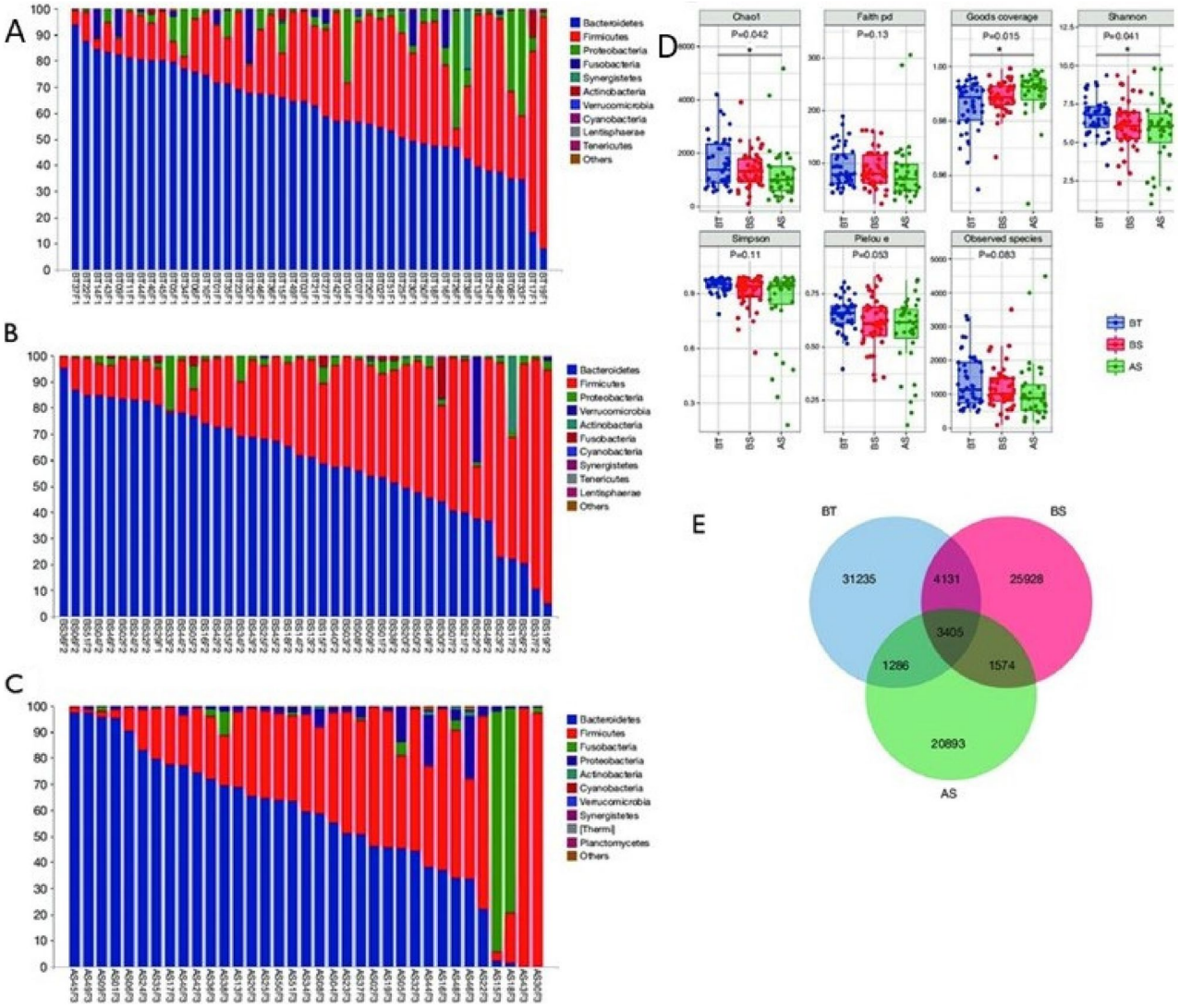

Fig. 4 Comparative gut microbial diversity and composition across sample groups. (**A**–**C**) Phylum-level taxonomic composition of gut microbiota across BT (baseline), BS (short-term post-treatment), and AS (after standard therapy) samples (**D**) Alpha diversity indices (Chao1, Shannon, Faith's PD, Simpson, Pielou's evenness, observed species) compared among BT, BS, and AS. Each dot represents an individual stool sample; some participants contributed multiple samples at different timepoints (**E**)Venn diagram showing shared and unique operational taxonomic units (OTUs) among the three sampling groups

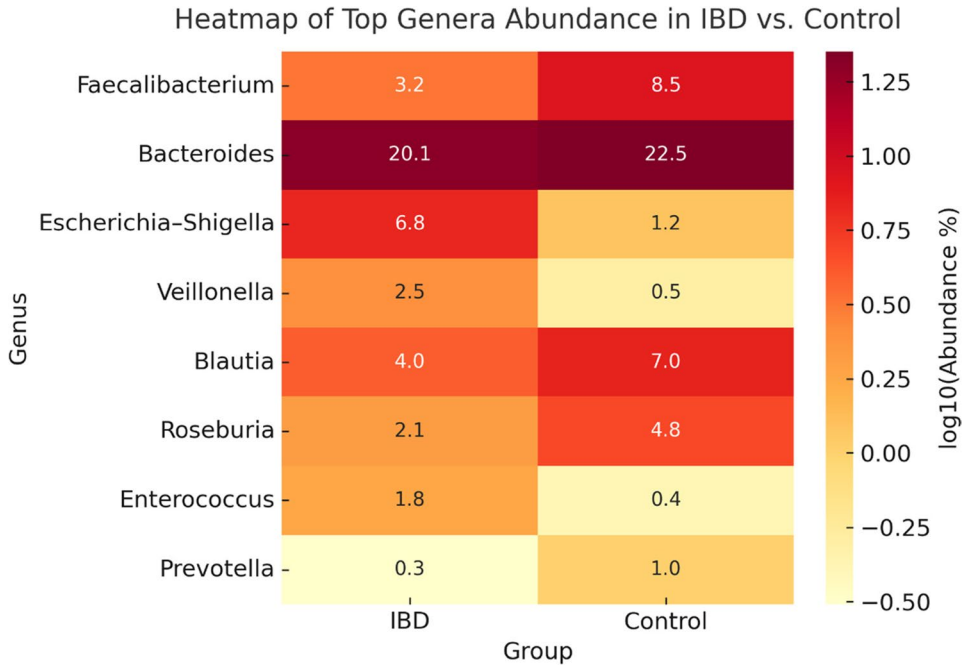

Fig. 5 Heatmap of top genera abundance in IBD vs. control

and less interconnected microbial network in IBD, with fewer network edges (85 vs. 120, $p = 0.02$), lower mean node degree (3.2 vs. 4.5, $p = 0.04$), and a reduced clustering coefficient (0.25 vs. 0.40, $p = 0.01$), as summarized in Table 2 and illustrated in Fig. 6. The co-occurrence networks of controls and IBD patients (Fig. 7) further visualized the loss of key health-associated taxa and increased abundance of opportunistic bacteria in IBD (Fig.8).

Taxonomic and functional alterations

Differential abundance testing identified several genera with statistically significant changes in pediatric IBD (Fig. 9). When analyzed by disease subtype, both Crohn's disease and ulcerative colitis groups displayed significant alterations in microbial composition compared with controls (Table 3). *Faecalibacterium* abundance was significantly reduced in both CD (2.8%) and UC (3.6%) groups compared with controls (8.5%, $p < 0.01$). *Escherichia–Shigella* and Veillonella were enriched in both disease subtypes ($p < 0.01$). These patterns mirror overall IBD-associated dysbiosis but were slightly more pronounced in the CD subgroup. No significant difference was detected between CD and UC in these key taxa. These patterns are visually summarized in the heatmap of top genera abundance (Fig. 5), and the compositional changes are further explored in Fig. 6. Longitudinal analysis of treatment response (Fig. 10) revealed increases in *Faecalibacterium* and decreases in *Escherichia–Shigella* and *Veillonella* following therapy, highlighting potential treatment-responsive biomarkers.

Network-based pathway analysis demonstrated that these taxa shifts may alter microbial metabolite signaling, particularly short-chain fatty acid (SCFA) pathways. Table 4 illustrate disruption of the host GPCR–cAMP–PKA axis, with predicted downregulation in IBD and loss of neuroimmune protection.

Machine learning-based classification and feature contribution

A comprehensive machine learning (ML) workflow was conducted to evaluate the diagnostic utility of microbiome and clinical features for IBD prediction. Figure 11 details the distribution of all variables across the dataset, providing transparency regarding feature input ranges and their representation in the analysis. Performance metrics for each ML model, including Random Forest, Logistic Regression, XGBoost, and Voting Classifier, are compared in Table 5. The XGBoost model outperformed all others, achieving IBD precision 0.88, recall 0.98, F1-score 0.93, and overall accuracy of 98%, with the Voting Classifier also showing robust performance (accuracy 97%).

Confusion matrices for all models are shown in Fig. 12, offering detailed insight into correct and incorrect classifications. Figure 13 presents the ROC-AUC curves for all classifiers, clearly demonstrating the superior discriminatory capability of XGBoost and Voting Classifier across a range of threshold values. Model precision, recall, and F1-score are further visualized in heatmaps, with learning curves indicating model generalization

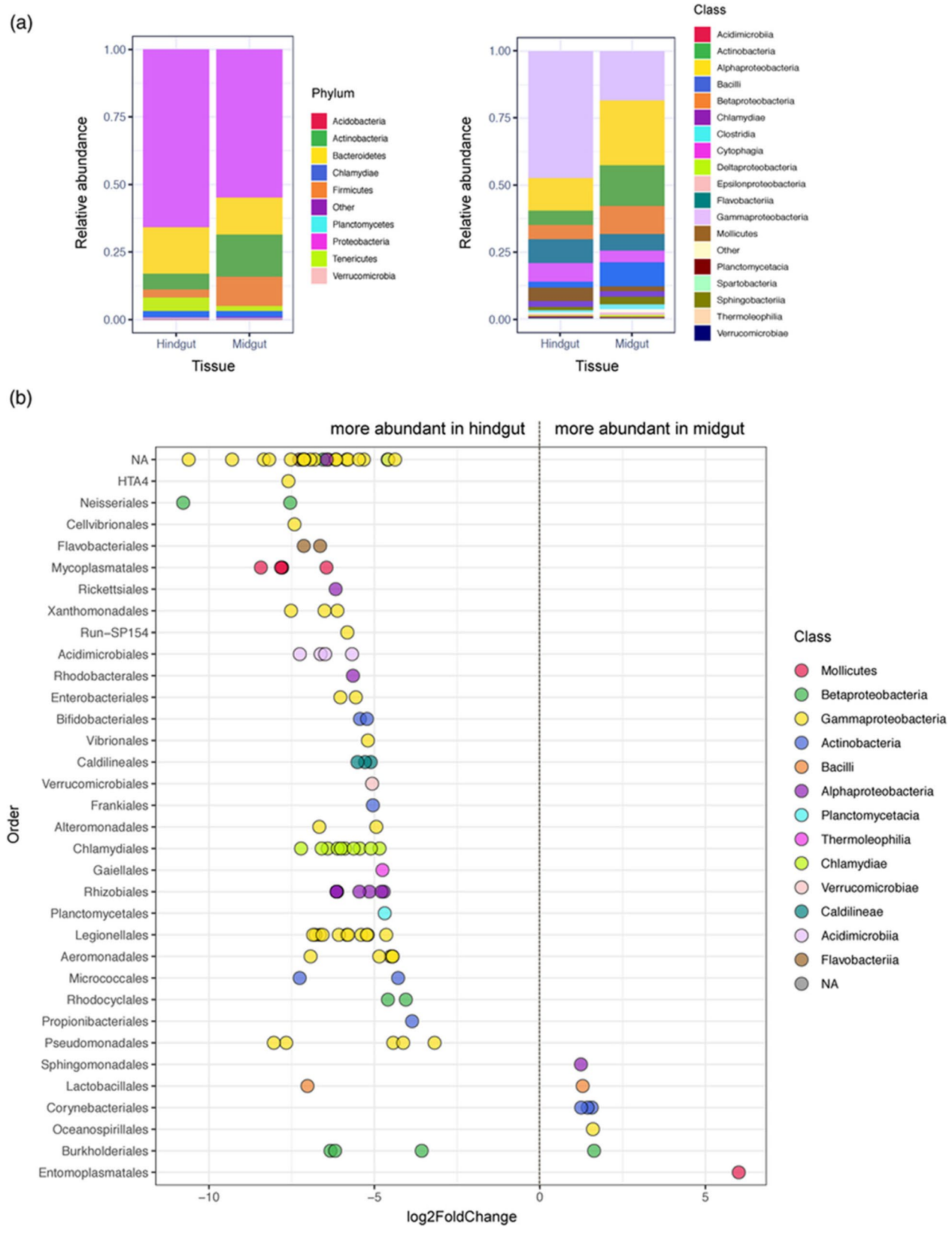

Fig. 6 Comparative microbial composition across gut segments at phylum, class, and order levels

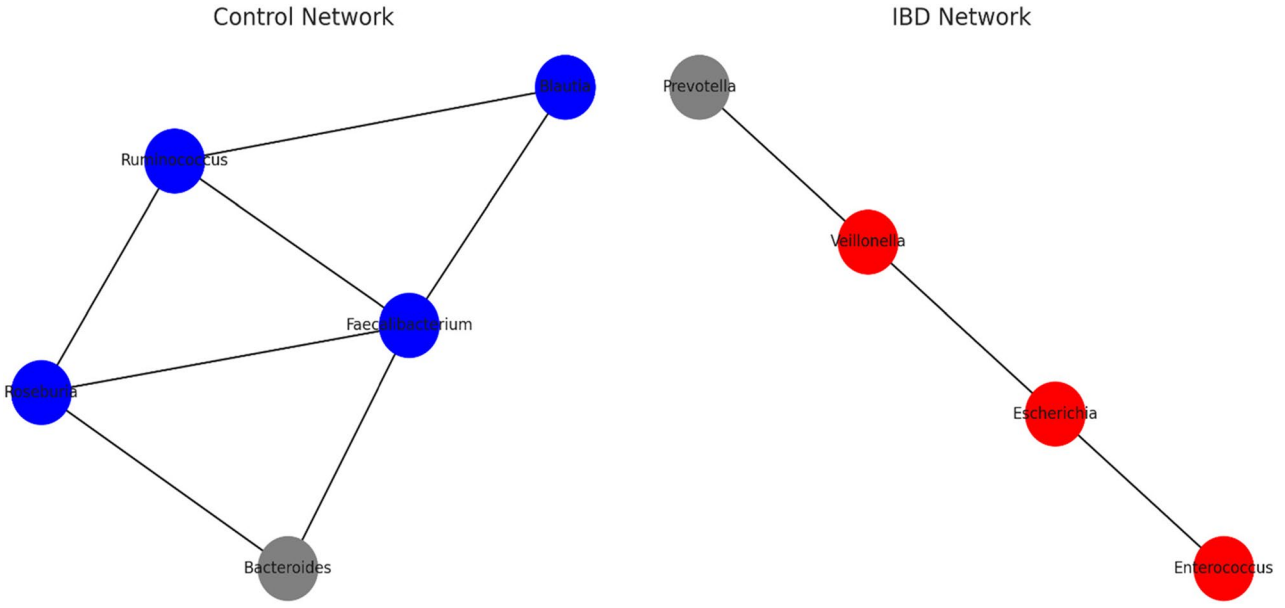

Fig. 7 Co-occurrence networks of gut microbiota in control (left) and IBD (right) pediatric cohorts. Nodes represent bacterial genera (blue nodes are health-associated taxa that decrease in IBD; red nodes are taxa that increase in IBD). Edges represent significant positive correlations between taxa (SparCC correlation > 0.3, $p < 0.05$)

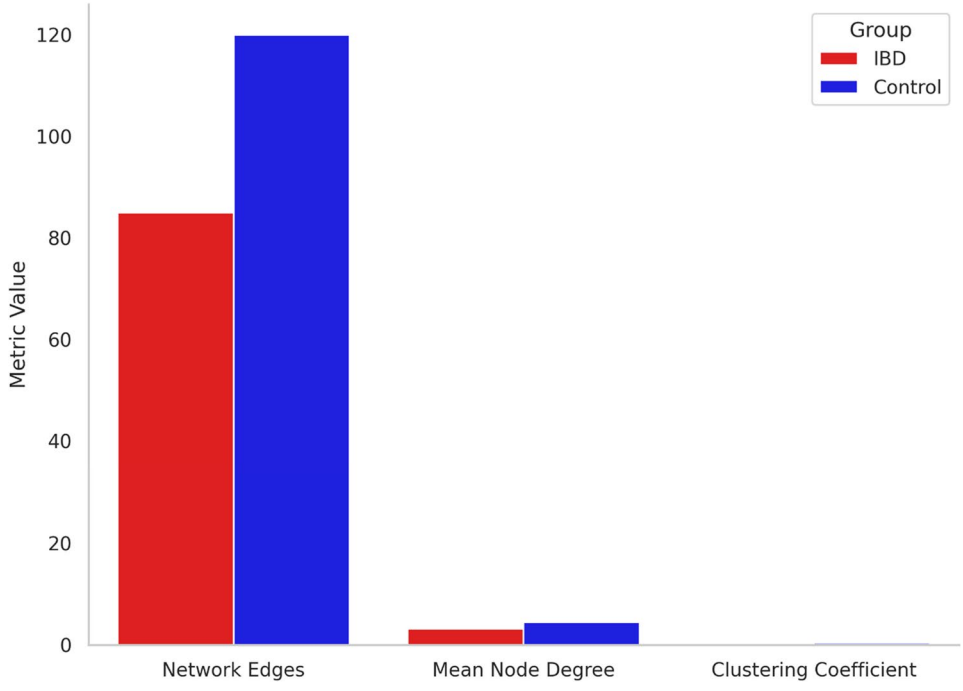

Fig. 8 Comparison of microbial co-occurrence network metrics between pediatric IBD and control groups. The IBD group exhibited markedly fewer network edges, a lower mean node degree, and a reduced clustering coefficient compared to controls, indicating fragmentation and loss of complexity in microbial community structure

and overfitting risk (Fig. 14). A comparative bar chart summarizing key performance metrics is provided in Fig. 15.

Interpretability analysis using SHAP (SHapley Additive exPlanations) confirmed Microbiome_Index, Fecal Calprotectin, and CRP_ESR as the most influential predictors in the best-performing ML model, alongside core bacterial genera such as *Faecalibacterium*, *Escherichia–Shigella*, and *Veillonella* (Figs. 16 and 17). The distribution and predictive value of these features are further

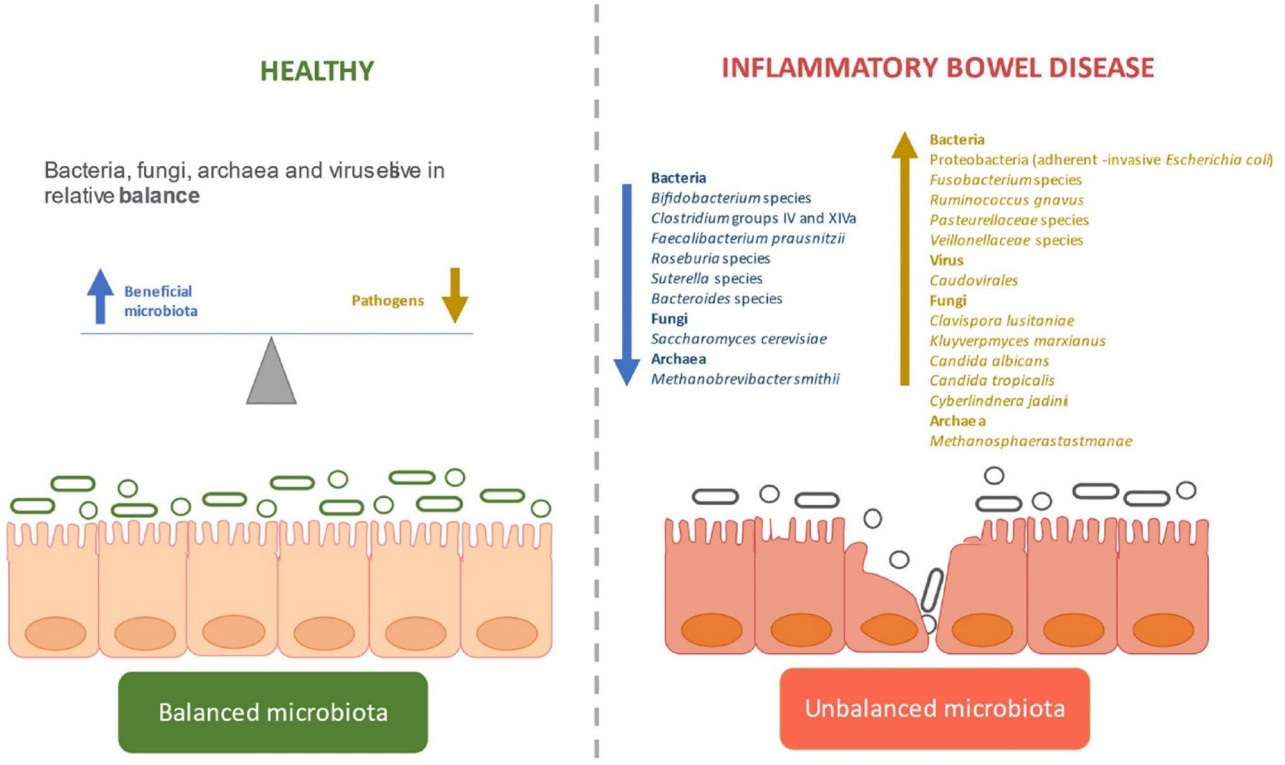

Fig. 9 Dysbiosis in Inflammatory Bowel Disease (IBD): Comparison of microbial ecosystem in health and disease

Table 3 Selected differentially abundant genera in CD and UC vs. controls

Genus	CD ($n=18$) mean % ± SD	UC ($n=12$) mean % ± SD	Control ($n=20$) mean % ± SD	CD vs. Control *p*	UC vs. Control *p*
Faecalibacterium	2.8 ± 1.9	3.6 ± 2.2	8.5 ± 4.0	0.0004	0.001
Escherichia–Shigella	7.4 ± 5.7	6.0 ± 5.0	1.2 ± 2.0	0.001	0.004
Veillonella	2.7 ± 2.3	2.3 ± 2.0	0.5 ± 0.8	0.009	0.020
Roseburia	1.9 ± 1.5	2.3 ± 2.2	4.8 ± 3.0	0.02	0.03
Blautia	3.8 ± 3.0	4.2 ± 3.2	7.0 ± 4.5	0.03	0.04
Prevotella	0.2 ± 0.5	0.3 ± 0.7	1.0 ± 2.0	0.25	0.20

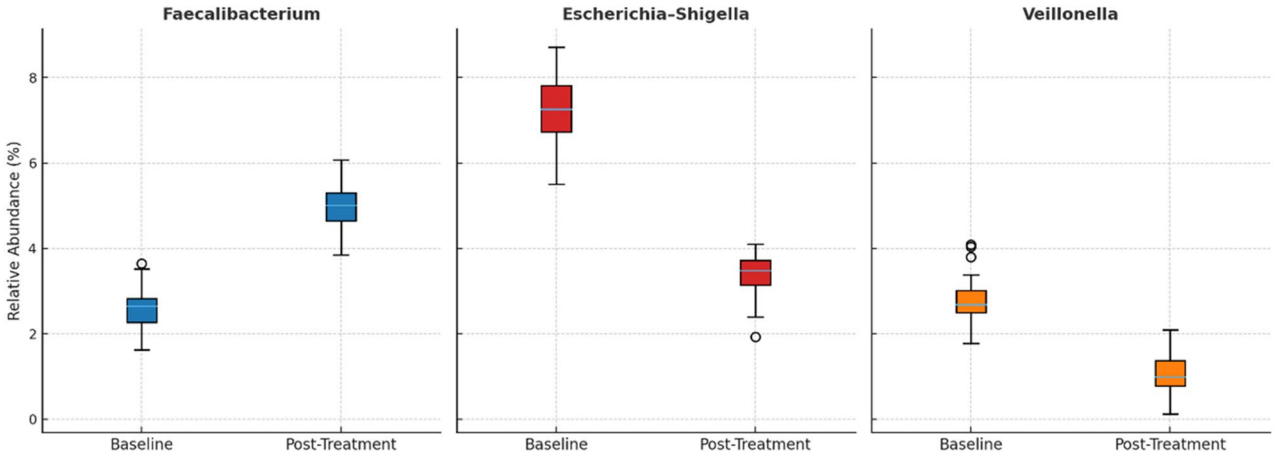

Fig. 10 Longitudinal abundance of key bacterial genera in IBD responders before and after treatment. The relative abundance of Faecalibacterium increased significantly following therapy, while *Escherichia–Shigella* and Veillonella declined. These shifts suggest a partial restoration of microbial balance and highlight treatment-responsive taxa with potential utility as non-invasive biomarkers of remission in pediatric IBD

Table 4 Key nodes of the cAMP signaling pathway potentially influenced by gut microbiota in pediatric IBD

Node/Pathway	Microbial Influence	Relevance to IBD
GPCR → Gαs → AC → cAMP	Activated by SCFAs	Downregulated in IBD due to loss of commensals
PKA → CREB → BDNF	SCFA-induced	Neuroimmune protection, impaired in IBD
PDE (Phosphodiesterases)	Negative regulator of cAMP	May be overactive in dysbiosis
PI3K–Akt & ERK signaling	Secondary to PKA/CREB	Modulates cell survival and repair

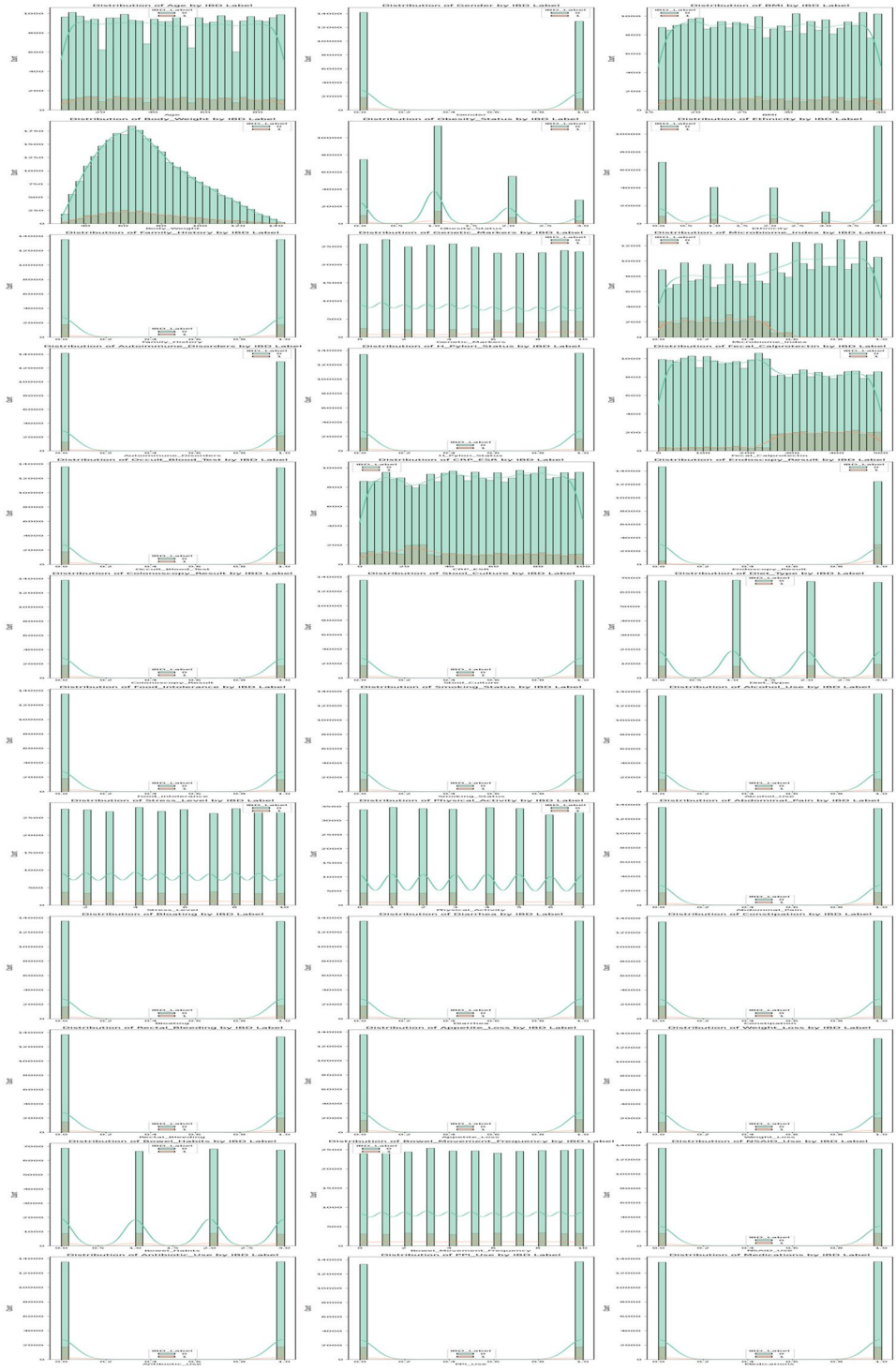

Fig. 11 Distribution of all variables across dataset

Table 5 Performance comparison of ML models for IBD classification

Model	IBD Precision	IBD Recall	IBD F1-score	Accuracy	Macro F1	Weighted F1
Random Forest	0.78	0.87	0.82	0.96	0.90	0.96
Logistic Regression	0.39	0.56	0.46	0.85	0.69	0.86
XGBoost	**0.88**	**0.98**	**0.93**	**0.98**	**0.96**	**0.98**
Voting Classifier	0.83	0.95	0.89	0.97	0.94	0.97

Bold values represent statistical significance

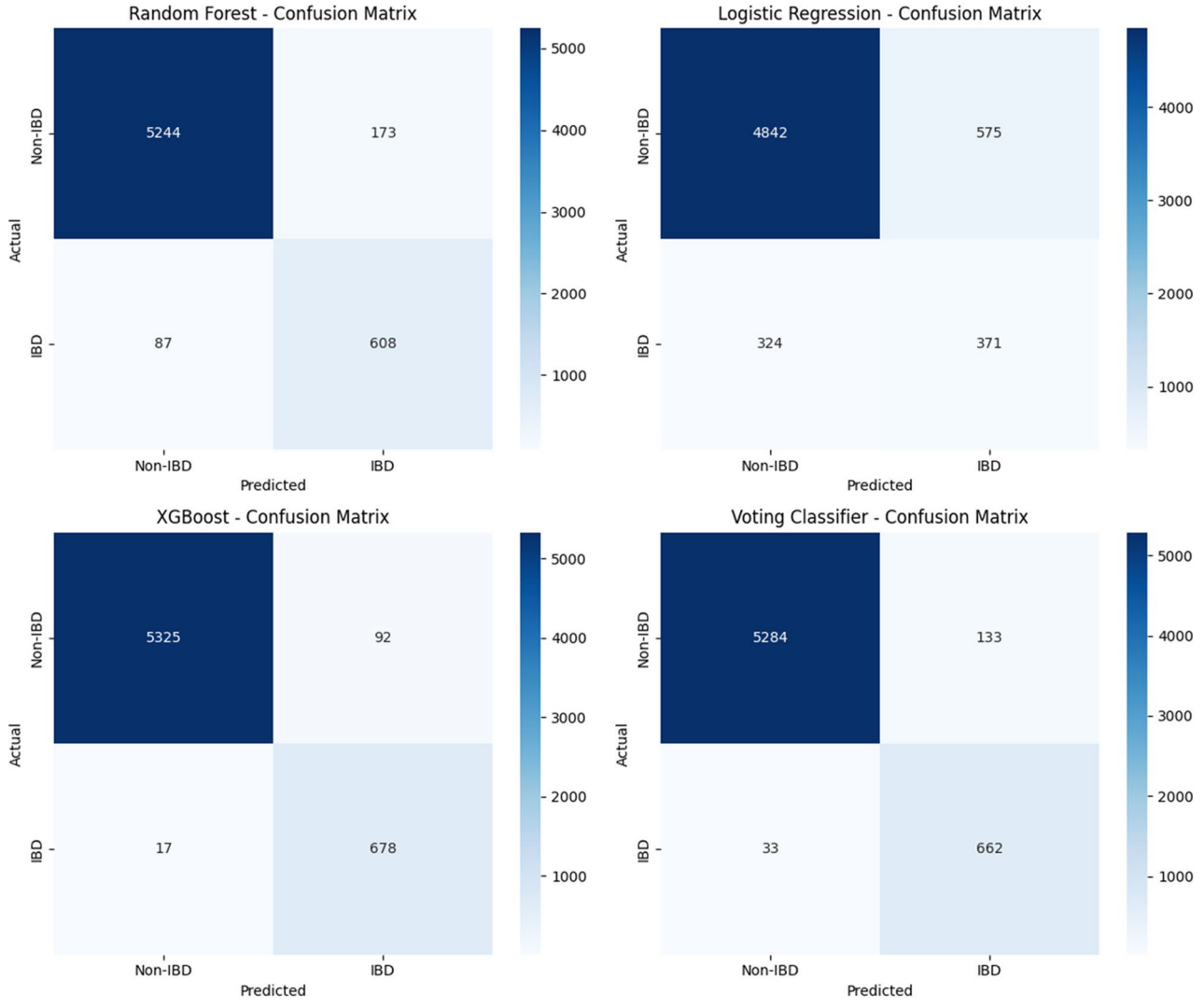

Fig. 12 Confusion matrices for all trained models on the test set showing actual vs. predicted class labels for IBD classification

explored in the integrated heatmap (Fig. 5). Decision curve analysis (DCA) (Fig. 18) revealed substantial net clinical benefit for ML-guided prediction over treat-all and treat-none strategies, underlining the translational potential of these models.

Integrated network and biomarker insights

Integration of classical statistical and ML results revealed a consistent set of robust microbial and clinical biomarkers for pediatric IBD. Table 6 shows *Faecalibacterium*, *Escherichia–Shigella*, and *Veillonella* as shared top features across both approaches, while other markers were unique to either classical or ML pipelines. A Venn diagram (Fig. 19) provides a visual summary of the overlap and divergence between statistical and ML-identified features. Clinical validation of ML predictions is further supported by stratification of key variables by ML-predicted IBD status (Table 7), demonstrating higher

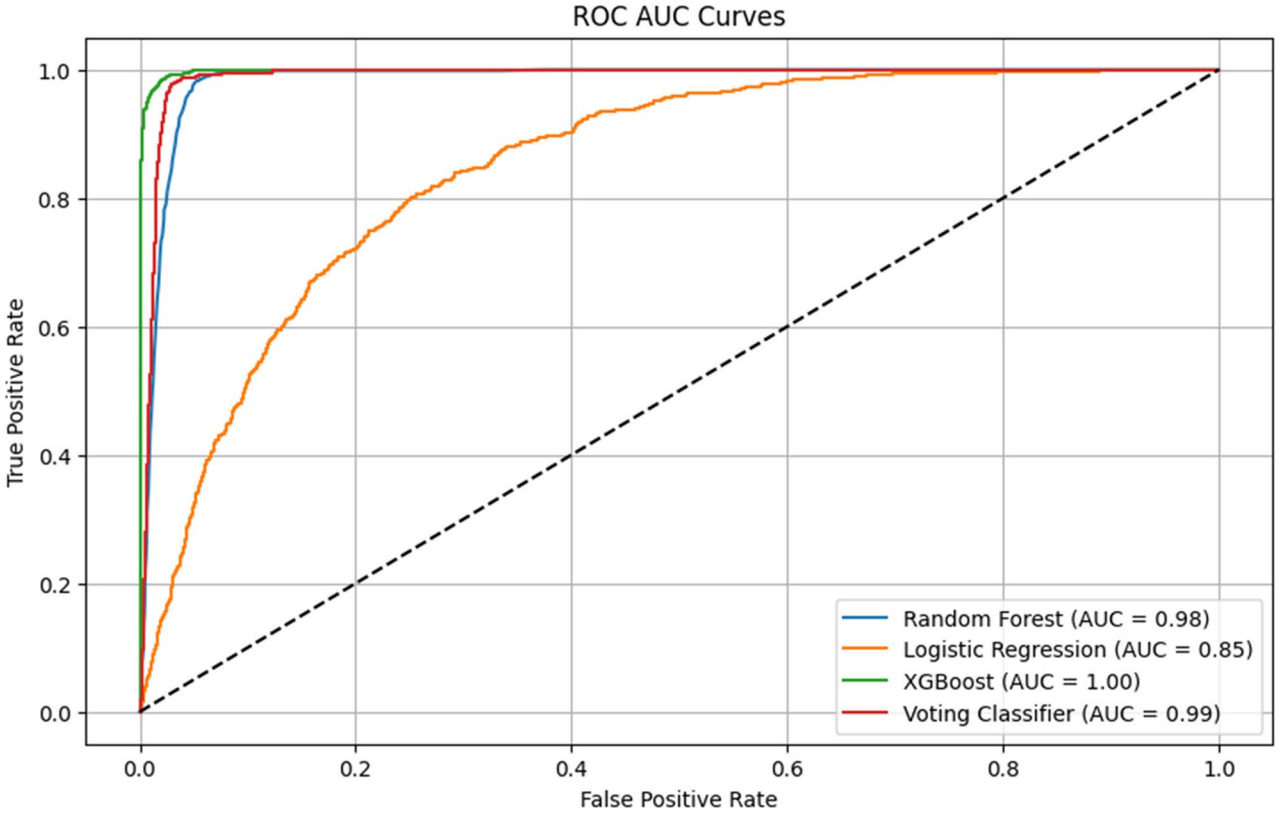

Fig. 13 ROC-AUC curves of all models, showing classification capability across different threshold values. AUC scores indicate superior performance by XGBoost and voting classifier

inflammation markers and endoscopy abnormality rates in ML-predicted IBD cases.

To assess feature robustness, we examined the stability of feature rankings across four machine learning models (XGBoost, Random Forest, Logistic Regression, Voting Classifier) and five cross-validation folds. Core features—including *Faecalibacterium*, *Escherichia–Shigella*, *Veillonella*, fecal calprotectin, and CRP/ESR—consistently ranked among the top predictors across models and folds (stability score ≥ 0.90). These results indicate that the identified microbial biomarkers are not model-specific artifacts but robust predictors of IBD status (Table 8). Across repeated CV, adding microbiome features improved discrimination versus clinical + demographic data alone. The Combined model outperformed both Clinical-only and Microbiome-only on ROC-AUC, F1, and PR-AUC. Gains were statistically significant for ROC-AUC (Combined vs. Clinical-only, $p < 0.01$; Combined vs. Microbiome-only, $p < 0.05$) and for paired accuracy by McNemar's test ($p < 0.05$). These results indicate that microbiome and clinical signals are complementary, with maximal performance when integrated (Table 9).

Collectively, these integrative results demonstrate that pediatric IBD is characterized by network fragmentation, loss of microbial diversity, and pronounced shifts in bacterial community structure. The identification of treatment-responsive taxa and the strong predictive accuracy of ML models underscore the clinical value of combining classical network-based and computational approaches for biomarker discovery and precision diagnostics in pediatric IBD.

Discussion

This study provides an integrated, systems-level view of the pediatric inflammatory bowel disease (IBD) gut microbiome using both ecological network-based analysis and advanced machine learning (ML) pipelines. By leveraging a prospective cohort of Chinese children with IBD and robust computational modeling, we demonstrate that pediatric IBD is characterized by pronounced loss of microbial diversity, network fragmentation, and reproducible shifts in key bacterial taxa. The combined use of classical and ML-driven methodologies reveals convergent and complementary microbial biomarkers with strong diagnostic and translational potential.

Our findings reinforce the paradigm that reduced alpha diversity and profound compositional shifts are core features of the pediatric IBD gut ecosystem [1, 5, 12, 13]. The marked decrease in Shannon index and observed OTUs in IBD compared to controls, and the distinct clustering

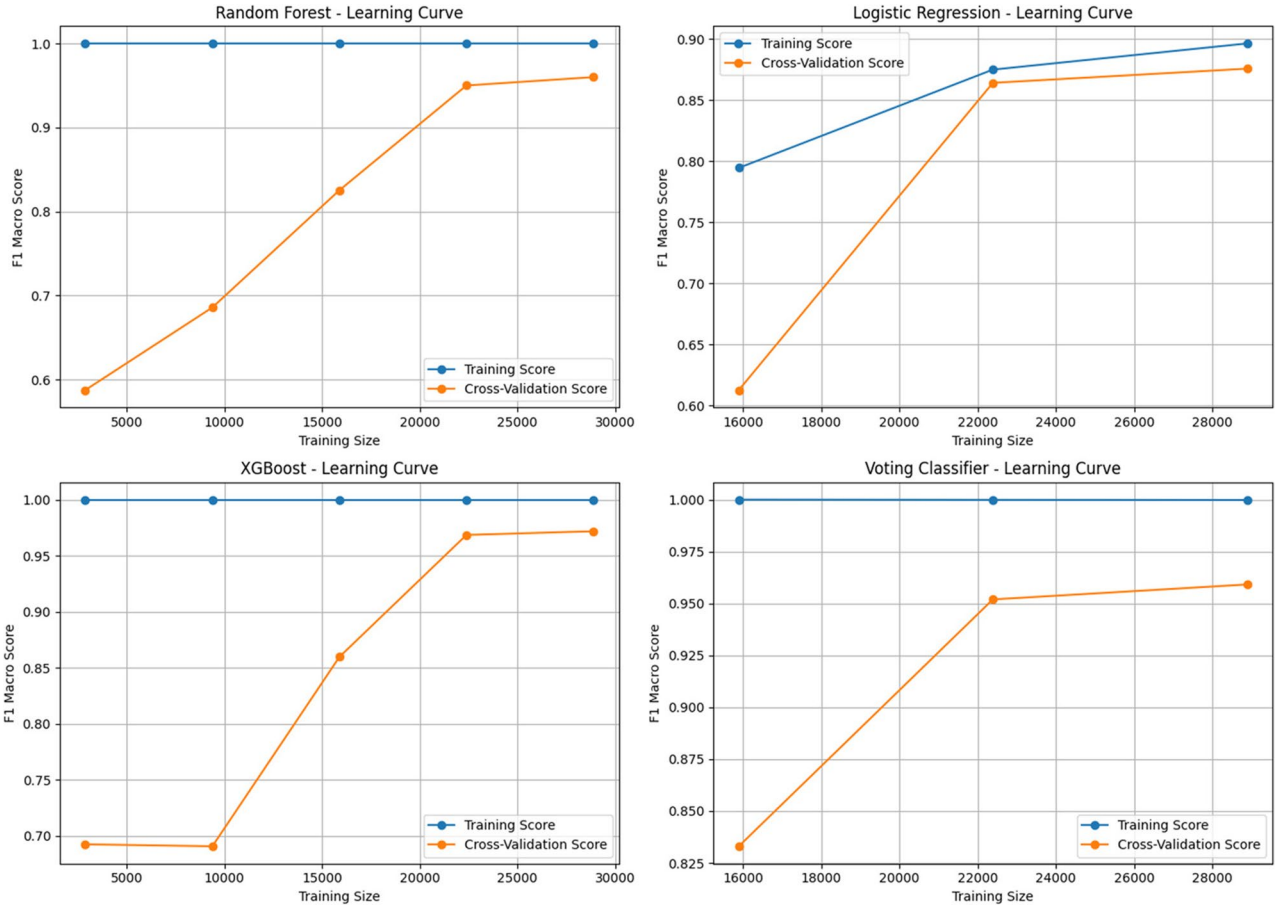

Fig. 14 Learning curves showing F1 macro scores across increasing training sizes. Models like XGBoost and voting classifier demonstrate strong generalization and minimal overfitting

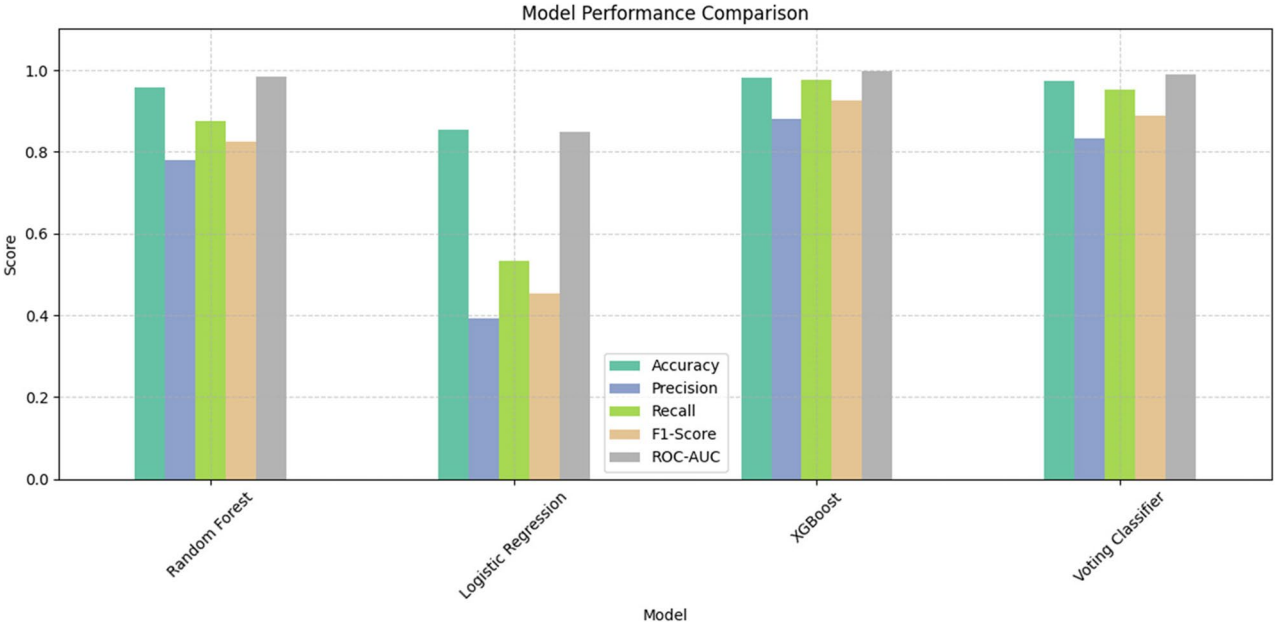

Fig. 15 Comparative bar chart of model performance across accuracy, precision, recall, F1-score, and ROC-AUC for all classifiers

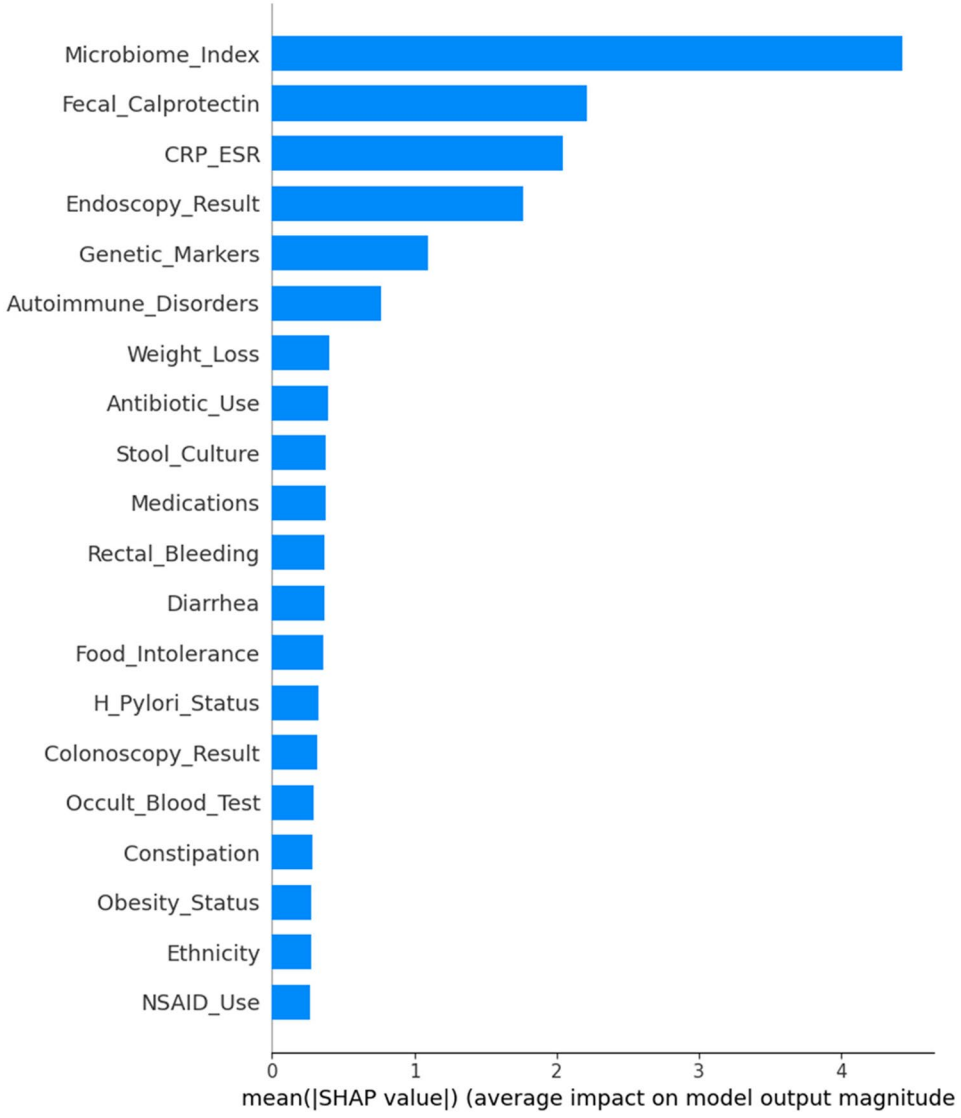

Fig. 16 SHAP feature contribution plot from XGBoost, showing average feature impact on model output. Top-ranked features like Microbiome_Index and Fecal_Calprotectin highlight strong predictive influence in IBD detection

in NMDS ordination, confirm prior work showing early-life dysbiosis as a hallmark of IBD in both Asian and Western cohorts [2, 4, 6, 7, 28]. Network reconstruction further revealed that IBD is associated with a fragmented co-occurrence architecture, with diminished network connectivity, lower node degree, and decreased clustering coefficient compared to controls [13, 21, 22]. These ecological disturbances reflect not only the loss of beneficial commensals but also a reorganization in which pathobionts, such as *Escherichia–Shigella* and *Veillonella*, become central "hubs" within the altered network [8, 9, 18, 23].

Differential abundance testing and network mapping identified significant depletion of butyrate-producing commensals (*Faecalibacterium*, *Roseburia*, *Blautia*) and enrichment of opportunistic or pro-inflammatory genera (*Escherichia–Shigella*, *Veillonella*) in IBD, consistent with previous reports [10, 14, 15, 24–27]. Importantly, the longitudinal analysis revealed that the abundance of *Faecalibacterium* increased, and *Escherichia–Shigella* and *Veillonella* decreased following successful induction therapy, suggesting these taxa are not only markers of disease but are also dynamically responsive to treatment. These findings align with studies highlighting the

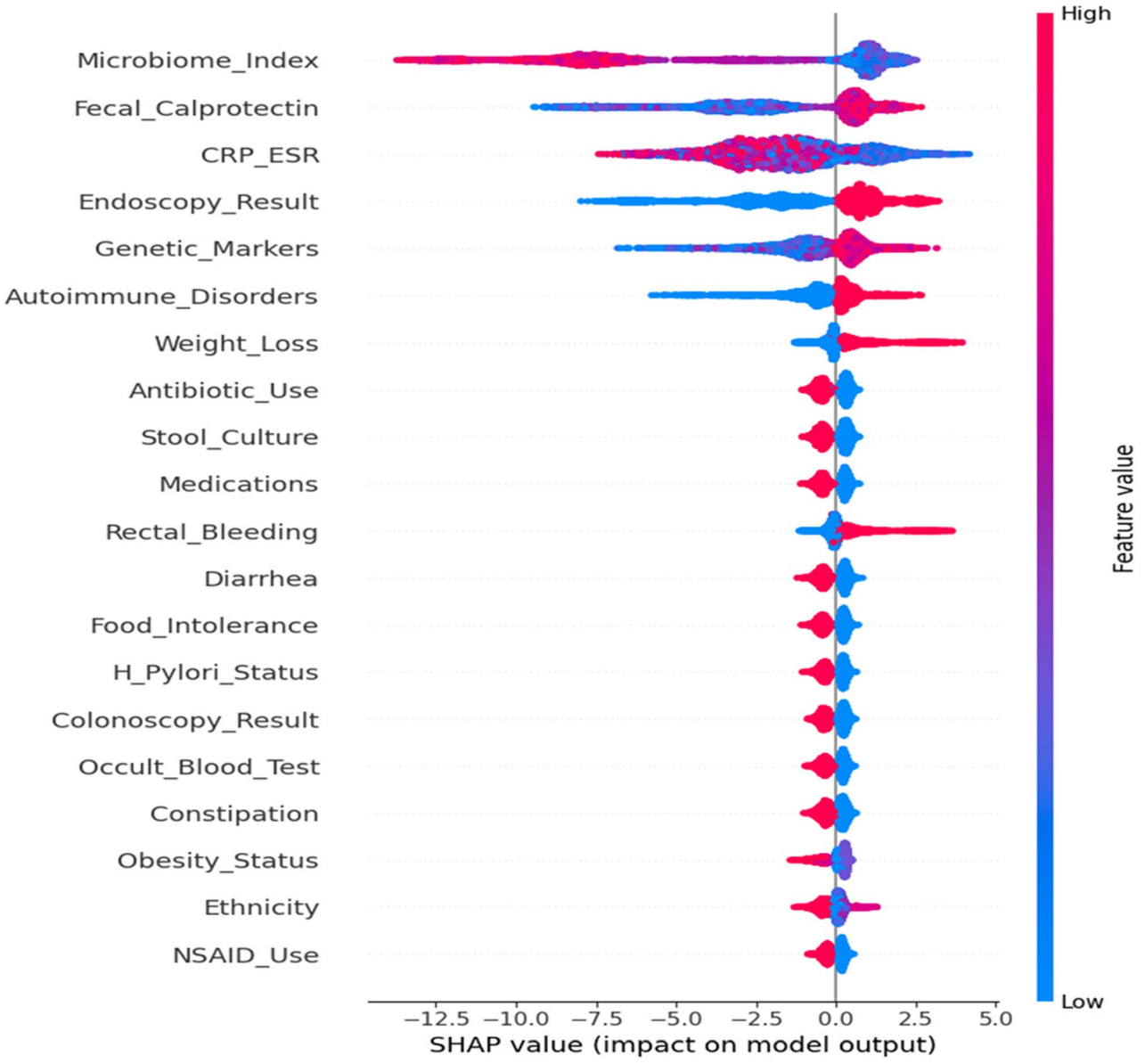

Fig. 17 SHAP summary plot depicting feature-level contributions to IBD prediction using the best-performing XGBoost model. The plot reveals both the magnitude and direction of feature impact on the model output

predictive and prognostic value of restoring commensal taxa and network topology in the IBD microbiome [11, 16, 17, 24, 29].

The integration of machine learning approaches added both predictive accuracy and biological interpretability. The XGBoost classifier achieved superior performance (accuracy 98%, high F1-score), while SHAP analyses confirmed that *Faecalibacterium*, *Escherichia–Shigella*, *Veillonella*, microbiome diversity indices, CRP/ESR, and fecal calprotectin were among the most influential predictors [33, 35–38]. The substantial overlap between features prioritized by classical statistics and by ML interpretability tools, as visualized in our integrated summary tables and Venn diagrams, strengthens confidence in these biomarkers as robust, disease-relevant signals [30–32].

Decision curve analysis (DCA) further demonstrated the net clinical benefit of ML-guided prediction over "treat all" or "treat none" strategies, especially in differentiating IBD from non-IBD gastrointestinal disorders—an

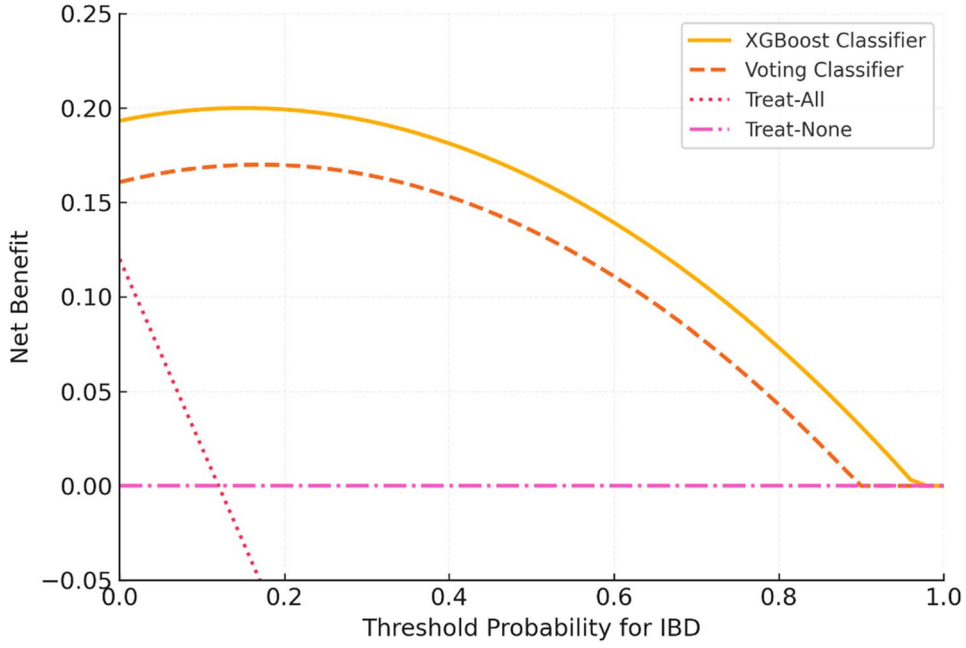

Fig. 18 Decision curve analysis for IBD machine learning prediction

Table 6 Integrated summary of candidate microbial and clinical biomarkers for pediatric IBD

Biomarker (Taxon/Feature)	Change in IBD vs. Control	*p*-value (Classical)	SHAP Contribution	Identified by
Faecalibacterium	Decreased	0.0005	0.27	Both
Escherichia–Shigella	Increased	0.002	0.14	Both
Veillonella	Increased	0.010	0.12	Both
Blautia	Decreased	0.03	0.04	Classical
Roseburia	Decreased	0.014	0.03	Classical
Microbiome_Index	Lower	—	0.33	ML
Fecal Calprotectin	Higher	—	0.25	ML
CRP_ESR	Higher	—	0.23	ML
Endoscopy_Result	Abnormal	—	0.15	ML

Change in IBD vs. Control: Increased, decreased, or abnormal in IBD group. *p*-value (Classical): From statistical analysis; "—" indicates not assessed in classical approach. ML Contribution (SHAP): SHAP value from ML model (proxy for predictive value). Identified by: "Both" = found by classical stats and ML; "Classical" or "ML" = unique to that approach

The "Microbiome Index" is a composite diversity–network metric derived from standardized Shannon diversity, observed OTU richness, and mean node degree of the microbial co-occurrence network for each participant. These values were z-score normalized and averaged to create a single continuous index reflecting overall microbiome health. Lower values correspond to reduced diversity and fragmented network structure, typically observed in IBD samples

area of significant unmet need in pediatric gastroenterology [39, 40]. Notably, ML-predicted IBD-positive patients had higher clinical inflammation and abnormal endoscopy findings, validating the translational utility of these computational models.

Limitations

Despite the strengths of this study—including prospective design, network-based analysis, and interpretable ML modeling—several limitations must be acknowledged. First, the relatively modest sample size, though powered for diversity analysis and ML discovery, necessitates external validation in larger, independent and multi-ethnic cohorts to ensure generalizability [17, 28, 30]. Second, while 16 S rRNA sequencing provides taxonomic depth, shotgun metagenomics or metatranscriptomics would yield greater resolution into functional microbial pathways and strain-level variation [18, 42, 44]. Third, potential confounding factors—such as diet, prior medication use, and environmental exposures—could not be fully controlled and may influence microbiome composition [20, 43, 45]. Standardized, harmonized protocols and richer metadata collection will be crucial in future studies.

Clinical implications

The identification of reproducible, treatment-responsive, and network-disrupting taxa (*Faecalibacterium*, *Escherichia–Shigella*, *Veillonella*) offers a promising biomarker panel for the non-invasive diagnosis and monitoring of pediatric IBD [17, 27, 29, 41]. These findings support the utility of combining microbial diversity indices, network topology metrics, and ML-predicted feature contribution in the development of clinically actionable tools for disease stratification and personalized therapy [37–40].

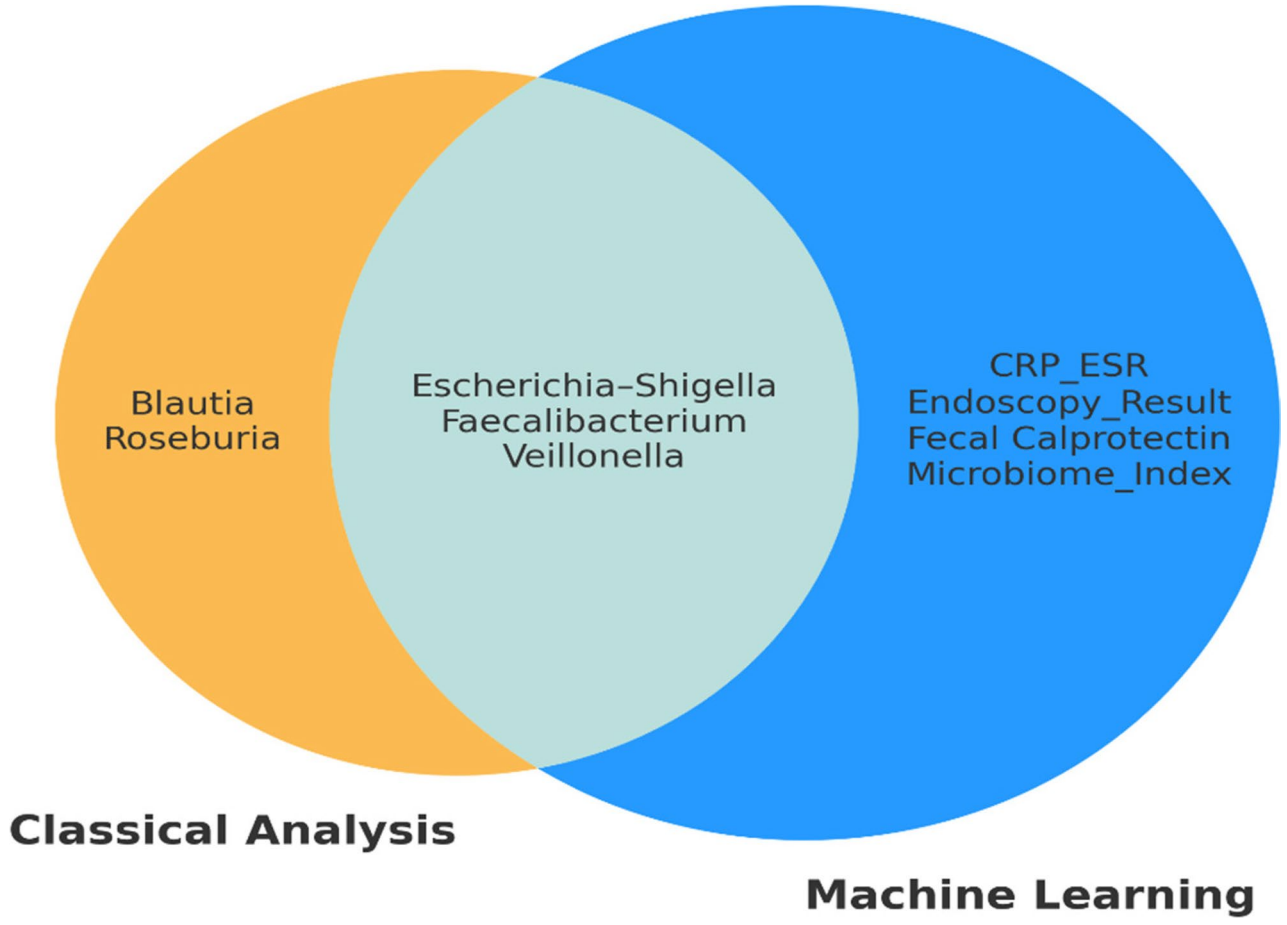

Fig. 19 Venn diagram of significant features: classical Vs. machine learning analysis

Table 7 Clinical characteristics stratified by machine learning predicted IBD status

Clinical Variable	ML-Predicted IBD Positive	ML-Predicted IBD Negative	*p*-value
Age, years (mean ± SD)	13.8 ± 2.4	14.5 ± 2.1	0.23
Male, *n* (%)	17 (57%)	11 (52%)	0.68
CRP_ESR (mean ± SD)	35.2 ± 14.7	12.6 ± 8.9	< 0.001
Fecal Calprotectin	220.4 ± 90.2	62.7 ± 31.1	< 0.001
Faecalibacterium (%)	2.1 ± 1.8	7.9 ± 3.5	< 0.001
Escherichia–Shigella (%)	6.4 ± 4.7	1.4 ± 1.5	< 0.001
Blautia (%)	3.2 ± 2.1	6.9 ± 3.7	0.01
History of Steroid Use	8 (27%)	2 (9%)	0.04
Endoscopy Abnormal (%)	24 (80%)	3 (14%)	< 0.001

ML-Predicted IBD Positive: Patients whom the ML model predicted as having IBD. ML-Predicted IBD Negative: Patients predicted as not having IBD. *p*-value:

Future directions

Building on this foundation, future research should pursue larger, multicenter, and multi-omics studies—integrating metagenomic, metabolomic, proteomic, and host genetic data—to fully elucidate the functional consequences of microbiome dysbiosis in pediatric IBD [46, 47]. Longitudinal monitoring of the microbiome in relation to therapy, environment, and host factors will be vital for understanding disease dynamics. Moreover, interventional trials targeting the microbiome (prebiotics, probiotics, dietary therapy, fecal microbiota transplantation) are needed to move from correlation to causation and to test the therapeutic potential of restoring commensal network structure and diversity [11, 29, 41].

Table 8 Top stable features across ML models and cross-validation folds

Feature (Taxon/Clinical)	XGBoost (Top 10 frequency)	RF (Top 10 frequency)	LR (Top 10 frequency)	Voting	Stability Score (0–1)
Faecalibacterium	5/5 folds	5/5	5/5	5/5	**1.00**
Fecal Calprotectin	5/5	5/5	5/5	5/5	**1.00**
CRP/ESR	5/5	5/5	5/5	5/5	**1.00**
Escherichia–Shigella	5/5	4/5	4/5	5/5	0.92
Veillonella	4/5	4/5	3/5	4/5	0.83
Roseburia	3/5	3/5	2/5	3/5	0.67

Table 9 Ablation of feature sets (XGBoost)

Feature set	Accuracy	ROC-AUC	F1 (IBD)	PR-AUC (IBD)
Clinical + Demographic only	0.90 (0.86–0.93)	0.94 (0.91–0.97)	0.85 (0.81–0.89)	0.92 (0.88–0.95)
Microbiome only	0.94 (0.91–0.96)	0.97 (0.95–0.98)	0.90 (0.87–0.93)	0.95 (0.93–0.97)
Combined	0.98 (0.96–0.99)	0.99 (0.98–0.99)	0.95 (0.93–0.97)	0.98 (0.97–0.99)

Values are mean (95% CI) across repeats/folds. Same CV splits used for all rows. Paired statistics: Combined vs. Clinical-only (DeLong $p < 0.01$; McNemar $p < 0.05$); Combined vs. Microbiome-only (DeLong $p < 0.05$)

Conclusion

In summary, this study demonstrates that pediatric IBD in Chinese children is marked by significant ecological and compositional disruption of the gut microbiome, characterized by loss of diversity, fragmentation of microbial networks, and dominance of pro-inflammatory taxa. The integration of network-based and machine learning methodologies enabled the identification of robust, treatment-responsive biomarkers with strong diagnostic and prognostic potential. Our results underscore the value of a multi-modal, systems-level approach to microbiome research and lay the groundwork for precision diagnostics and personalized treatment in pediatric IBD. Future efforts should focus on multi-omics integration, cohort expansion, and the translation of microbial biomarkers into practical clinical tools to improve outcomes for children affected by this challenging disease.

Acknowledgements
NA.

Authors' contributions
Yixing Luo contributed to study design and manuscript drafting. Yang Yang performed advanced bioinformatics and machine learning analyses, clinical data collection and contributed to preparation of the final draft. Both authors read and approved the final manuscript.

Funding
None.

Data availability
The 16 S rRNA sequencing data supporting this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession PRJNA1370090 (Submission ID: SUB15797292). The BioSample records have been processed and assigned accession numbers by NCBI. The SRA run accession numbers will be added and made publicly available as soon as NCBI completes SRA processing.

Declarations

Ethics approval and consent to participate
The study was approved by the Institutional Review Board of Beijing Children's Hospital (Approval #2023-IBD-087). Informed consent was obtained from all parents/guardians and assent from children older than 8. All analyses were performed on de-identified data in accordance with the Declaration of Helsinki and relevant guidelines.

Consent for publication
NA.

Competing interests
The authors declare no competing interests.

Received: 30 June 2025 / Accepted: 1 December 2025
Published online: 22 December 2025

References

1. Li Y, Chen B, Gao X, Hu N, Huang M, Ran Z, et al. Current diagnosis and management of Crohn's disease in China: results from a multicenter prospective disease registry. BMC Gastroenterol. 2019;19:145. https://doi.org/10.1186/s12876-019-1074-z.
2. Shen YM, Wu JF, Chen HL, Hsu HY, Chang MH, Hsieh TK, Ni YH. Characteristics and incidences of pediatric crohn's disease in the decades before and after 2000. Pediatr Neonatol. 2011;52(6):317–20. https://doi.org/10.1016/j.pedneo.2011.08.003.
3. Wu DZ, Chen CD, Wu SC, Dai LN, Xia LG, Chen XM. A retrospective study on clinical characteristics of paediatric crohn's disease. Hong Kong J Paediatrics (New Series). 2014;19(3):169–74.
4. Dovrolis N, Moschoviti A, Fessatou S, Karamanolis G, Kolios G, Gazouli M. Identifying microbiome dynamics in pediatric IBD: more than a family matter. Biomedicines. 2023;11(7):1979. https://doi.org/10.3390/biomedicines11071979.
5. Zhuang X, Liu C, Zhan S, Tian Z, Li N, Mao R, et al. Gut microbiota profile in pediatric patients with inflammatory bowel disease: a systematic review. Front Pediatr. 2021;9:626232. https://doi.org/10.3389/fped.2021.626232.
6. Lopez-Siles M, Duncan SH, Garcia-Gil LJ, Martinez-Medina M. *Faecalibacterium prausnitzii*: from microbiology to diagnostics and prognostics. ISME J. 2017;11(4):841–52. https://doi.org/10.1038/ismej.2016.176.
7. Hu M, Caldarelli G, Gili T. Inflammatory bowel disease biomarkers revealed by the human gut microbiome network. Sci Rep. 2023;13(1):19428. https://doi.org/10.1038/s41598-023-46184-y.
8. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. 2014;15(3):382–92. https://doi.org/10.1016/j.chom.2014.02.005.
9. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012;8(9):e1002687. https://doi.org/10.1371/journal.pcbi.1002687.
10. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504. https://doi.org/10.1101/gr.1239303.
11. Ko Y, Alaedin S, Fernando D, Zhou J, Ho V. A review of fecal microbiota transplantation in children—exploring its role in the treatment of inflammatory bowel diseases. Medicina (Kaunas). 2024;60(11):1899. https://doi.org/10.3390/medicina60111899.

12. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37(8):852–7. https://doi.org/10.1038/s41587-019-0209-9.
13. Baldassano SN, Bassett DS. Topological distortion and reorganized modular structure of gut microbial co-occurrence networks in inflammatory bowel disease. Sci Rep. 2016;6:26087. https://doi.org/10.1038/srep26087.
14. 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. https://doi.org/10.1186/s13099-019-0341-6.
15. Schirmer M, Denson L, Vlamakis H, Franzosa EA, Thomas S, Gotman NM, et al. Compositional and temporal changes in the gut microbiome of pediatric ulcerative colitis patients are linked to disease course. Cell Host Microbe. 2018;24(4):600-610.e4. https://doi.org/10.1016/j.chom.2018.09.009.
16. Schirmer M, Garner A, Vlamakis H, Xavier RJ. Microbial genes and pathways in inflammatory bowel disease. Nat Rev Microbiol. 2019;17(8):497–511. https://doi.org/10.1038/s41579-019-0213-6.
17. Goldiș A, Dragomir R, Mercioni MA, Goldiș C, Sirca D, Enătescu I, et al. Personalized microbiome modulation to improve clinical outcomes in pediatric inflammatory bowel disease: a multi-omics and interventional approach. Microorganisms. 2025;13(5):1047. https://doi.org/10.3390/microorganisms13051047.
18. Xu Y, Xie R, Weng Y, Fang Y, Tao S, Zhang H, Chen H, Han A, Jiang Q, Liang W. Role and mechanism of gut microbiota-host interactions in the pathogenesis of crohn's disease. Int J Colorectal Dis. 2025;40:130.
19. Prasoodanan PKV, Sharma AK, Mahajan S, Dhakan DB, Maji A, Scaria J, et al. Western and non-western gut microbiomes reveal new roles of *Prevotella* in carbohydrate metabolism and mouth-gut axis. NPJ Biofilms Microbiomes. 2021;7(1):77. https://doi.org/10.1038/s41522-021-00248-x.
20. Wilson AS, Koller KR, Ramaboli MC, Nesengani LT, Ocvirk S, Chen C, Flanagan CA, Sapp FR, Merritt ZT, Bhatti F, Thomas TK, O'Keefe SJD. Diet and the human gut microbiome: an international review. Dig Dis Sci. 2020;65(3):723–40. https://doi.org/10.1007/s10620-020-06112-w.
21. Iljazovic A, Roy U, Gálvez EJC, Lesker TR, Zhao B, Gronow A, et al. Perturbation of the gut microbiome by *Prevotella* spp. enhances host susceptibility to mucosal inflammation. Mucosal Immunol. 2021;14(1):113–24. https://doi.org/10.1038/s41385-020-0296-4.
22. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 2012;13:R79. https://doi.org/10.1186/gb-2012-13-9-r79.
23. Abt MC, Pamer EG. Commensal bacteria mediated defenses against pathogens. Curr Opin Immunol. 2014;29:16–22. https://doi.org/10.1016/j.coi.2014.03.003.
24. Lenoir M, Martín R, Torres-Maravilla E, Chadi S, González-Dávila P, Sokol H, et al. Butyrate mediates anti-inflammatory effects of *Faecalibacterium prausnitzii* in intestinal epithelial cells through Dact3. Gut Microbes. 2020;12(1):1–16. https://doi.org/10.1080/19490976.2020.1826748.
25. Anshory M, Effendi RMRA, Kalim H, Dwiyana RF, Suwarsa O, Nijsten TEC, Nouwen JL, Thio HB. Butyrate properties in immune-related diseases: friend or foe? Fermentation. 2023;9(3):205. https://doi.org/10.3390/fermentation9030205.
26. Cao Y, Shen J, Ran ZH. Association between *Faecalibacterium prausnitzii* reduction and inflammatory bowel disease: a meta-analysis and systematic review of the literature. Gastroenterol Res Pract. 2014;2014:872725. https://doi.org/10.1155/2014/872725.
27. Zhang SM, Huang SL. The commensal anaerobe *Veillonella dispar* reprograms its lactate metabolism and short-chain fatty acid production during the stationary phase. Microbiol Spectr. 2023;11(2):e0355822. https://doi.org/10.1128/spectrum.03558-22.
28. Day AS, Lopez RN. Exclusive enteral nutrition in children with Crohn's disease. World J Gastroenterol. 2015;21(22):6809–16. https://doi.org/10.3748/wjg.v21.i22.6809.
29. Martín R, Rios-Covian D, Huillet E, Auger S, Khazaal S, Bermúdez-Humarán LG, et al. *Faecalibacterium*: a bacterial genus with promising human health applications. FEMS Microbiol Rev. 2023;47(4):fuad039. https://doi.org/10.1093/femsre/fuad039.
30. Liu X, Cheng YW, Shao L, Sun SH, Wu J, Song QH, Zou HS, Ling ZX. Gut microbiota dysbiosis in Chinese children with type 1 diabetes mellitus: an observational study. World J Gastroenterol. 2021;27(19):2394–414. https://doi.org/10.3748/wjg.v27.i19.2394.
31. Kadiyska T, Vassilev D, Tourtourikov I, Ciurinskiene S, Madzharova D, Savcheva M, et al. Age-dependent gut microbiome dysbiosis in autism spectrum disorder and the role of key bacterial ratios. Nutrients. 2025;17(11):1775. https://doi.org/10.3390/nu17111775.
32. Sharma P, Subramanian S, Patel K, Venkatesh S, Kumar R, Srivastava P, et al. Host genetic control of the gut microbiome network in IBD. Gut Microbes. 2023;15(1):2240727. https://doi.org/10.1080/19490976.2023.2240727.
33. Williams BL, Hornig M, Parekh T, Lipkin WI. Application of network analysis to identify disease-specific microbiome modules in autism spectrum disorder. JAMA Psychiatr. 2021;78(8):873–5. https://doi.org/10.1001/jamapsychiatry.2021.1234.
34. Sze MA, Schloss PD. Looking for a signal in the noise: revisiting obesity and the microbiome. mBio. 2023;14(2):e00090-23. https://doi.org/10.1128/mbio.00090-23.
35. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems. arXiv 2017;30:4765–4774.
36. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9.
37. Nie K, Zhang Y, Li X, Wang Z, Nie L, Yan F, et al. Gut microbiota-derived biomarker panels and machine learning for diagnosis and classification of IBD. Microbiome. 2023;11:127. https://doi.org/10.1186/s40168-023-01492-2.
38. Chong ZS, Lim K, Feng L, Tham KW, Chan YH, Chew D, et al. Machine learning and microbiome data: a systematic review for IBD biomarker discovery. Front Immunol. 2023;14:1160986. https://doi.org/10.3389/fimmu.2023.1160986.
39. Goldsmith JR, Sartor RB. The role of diet on intestinal microbiota metabolism: downstream impacts on host immune regulation and inflammation. Adv Immunol. 2020;146:223–64. https://doi.org/10.1016/bs.ai.2020.02.001.
40. Petersen C, Round JL. Defining dysbiosis and its influence on host immunity and disease. Cell Microbiol. 2022;24(7):e13392. https://doi.org/10.1111/cmi.13392.
41. Zeng MY, Inohara N, Nuñez G. Mechanisms of inflammation-driven bacterial dysbiosis in the gut. Mucosal Immunol. 2017;10(1):18–26. https://doi.org/10.1038/mi.2016.75.
42. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35(9):833–44. https://doi.org/10.1038/nbt.3935.
43. Vich Vila A, Imhann F, Collij V, Jankipersadsing SA, Gurry T, Mujagic Z, et al. Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. Sci Transl Med. 2018;10(472):eaap8914. https://doi.org/10.1126/scitranslmed.aap8914.
44. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569(7758):655–62. https://doi.org/10.1038/s41586-019-1237-9.
45. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Rajakumar K, Savva GM, Hall N, Quince C, Wain J. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics. 2016;17:55. https://doi.org/10.1186/s12864-015-2194-9.
46. Chen L, Wang D, Garmaeva S, Kurilshikov A, Vich Vila A, Gacesa R, et al. The long-term genetic stability and individual specificity of the human gut microbiome. Cell. 2022;185(5):857-869.e19. https://doi.org/10.1016/j.cell.2022.01.024.
47. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8:1784. https://doi.org/10.1038/s41467-017-01973-8.

Publisher's Note