Disentangling Overfitting From Biological Signal: A Responsible AI Framework for IBS Analysis

Nathanya Q. Satriani
Carinthia University of Applied Sciences
Klagenfurt, Austria
Nathanya.Satriani@edu.fh-kaernten.ac.at

Abstract

Machine learning applied to Irritable Bowel Syndrome (IBS) frequently reports high diagnostic performance despite small cohorts and heterogeneous preprocessing. We present a four-stage responsible AI workflow applied to the MARS-IBS-2020 dataset ($n = 73$ at baseline), demonstrating that leakage ablation, harmonized preprocessing, and learning-curve diagnostics expose structural overfitting behind nominally high ROC-AUC values. A single evaluation yielded AUC 0.985 under temporal filtering, yet bootstrapped resampling ($N = 50$) reveals a mean AUC of only 0.63 with 95% CI spanning [0.10, 1.00], and memorisation persists across all sample sizes. SHAP attribution consistently surfaces bile acid and tryptophan metabolites as dominant predictors, aligning with established IBS biology. The study offers a transparent methodological template for evaluating small-cohort biomedical ML systems where traditional performance metrics are unreliable.

1 Introduction

Irritable Bowel Syndrome (IBS) affects 5–10% of the global population and carries a substantial quality-of-life burden [12]. Diagnosis remains symptom-based (Rome IV criteria), and IBS pathophysiology involves altered microbial composition, disrupted metabolic pathways, and gut–brain axis dysregulation [1]. Metabolomic studies highlight bile acid and tryptophan disturbances as hallmarks of the disease [15]. Machine learning has been proposed to decode these complex multi-omic signals, yet reproducibility remains a critical concern. High AUC values (>0.90) have been reported in datasets with fewer than 100 subjects, raising questions about methodological validity [14, 9]. Systematic reviews confirm that subject-level data leakage, non-standardized feature engineering, and absent external validation routinely inflate performance by 0.15–0.30 AUC points [2, 11].

This work applies a four-stage responsible AI framework to the MARS-IBS-2020 cohort [8, 10]. Through this case study, we demonstrate the utility of combining established practices into a cohesive workflow: **(1)** applying a Progressive Leakage Ablation protocol that quantifies performance inflation under increasingly strict validation; **(2)** learning-curve diagnostics revealing overfitting persistence across sample sizes; and **(3)** SHAP-based biological plausibility analysis as a stability measure when predictive metrics are unreliable.

2 Methodology

Figure 1 summarises the four-stage workflow from data harmonization to biological plausibility validation.

Dataset. The MARS-IBS-2020 dataset [8, 10] contains longitudinal fecal samples from 305 IBS subjects (across subtypes) and 139 healthy controls, with paired 16S rRNA sequencing and untargeted

The Third Austrian Symposium on AI and Vision (AIROV26).

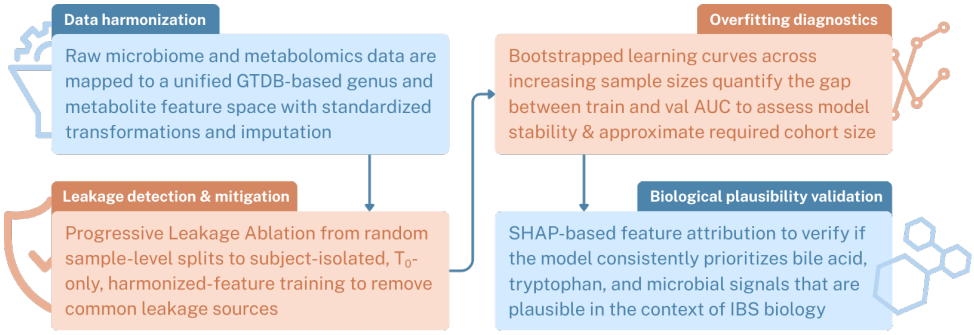

Figure 1: Four-stage responsible AI workflow: (1) data harmonization; (2) leakage mitigation; (3) overfitting diagnostics; (4) biological plausibility validation.

LC-MS metabolomics. Filtering for the first time-point (T_0) yields 73 unique subjects with complete multi-modal data (51 IBS cases, 22 healthy controls; median age 34 years, 79% female).

Feature engineering and harmonisation. Raw taxonomic strings were mapped to GTDB genus-level identifiers. Genera present in fewer than 10% of samples or with variance < 0.01 were removed; the remainder underwent Centred Log-Ratio (CLR) transformation. Metabolomics data were log-transformed, quantile-normalised, and imputed via KNN ($k = 5$). The final harmonised set comprised **53 features** (10 genera, 43 metabolites).

External validation cohorts. To assess generalisability, three independent public microbiome datasets were used: Franzosa ($N = 56$) [4], Jacobs ($N = 36$) [5], and iHMP ($N = 278$) [6]. Because these cohorts lacked aligned metabolomics data, they were used strictly to evaluate the microbiome-only variant of the model.

Validation framework. Performance was evaluated under four increasingly strict schemes:

1. **Sample-level split (leakage):** random 70/15/15 split allowing the same subject across folds.
2. **Subject-isolated split:** each subject appears in exactly one fold.
3. **Temporal filtering:** restriction to T_0 removes longitudinal leakage and establishes a baseline diagnostic capacity without the confounding temporal variance of disease progression, though this limits the exploration of temporal dynamics.
4. **Harmonised feature space:** model trained on the reduced 53-feature set.

An XGBoost classifier (learning rate 0.05, depth 5, subsample 0.8) was trained with 5-fold cross-validation and 50 bootstrap iterations. Bootstrapped learning curves ($n = 20$–73) were computed by re-fitting preprocessing and model parameters in each iteration. SHAP (tree explainer) was used for interpretability [7, 3].

3 Results

3.1 Progressive leakage ablation

Table 1 reports AUC across the four validation stages. A single evaluation suggested a paradoxical AUC *increase* from 0.936 to 0.985 as leakage was progressively eliminated. Bootstrapped resampling ($N = 50$), however, reveals that subject-isolated and temporal-filtered stages produce mean AUCs of only $\approx$0.63 with 95% CIs spanning nearly the full range (Table 1), confirming that single-run metrics in small cohorts are deeply unreliable. Only Stage 1 (with leakage) yields a stable estimate (0.944, CI [0.884, 0.986]).

Silhouette analysis on standardised features confirmed that class separability is near-zero in both the full dataset ($s = -0.009$, $n = 444$) and the T_0 subset ($s = -0.011$, $n = 73$), indicating the AUC paradox reflects memorisation of non-separable clusters rather than genuine discriminability.

Table 1: Progressive leakage ablation: single-run AUC vs. bootstrapped mean AUC with 95% confidence intervals ($N = 50$ iterations).

Stage	Validation scheme	Single-run	Bootstrap AUC (95% CI)
1	Sample-level split (leakage)	0.936	0.944 [0.884, 0.986]
2	Subject-isolated split	0.963	0.637 [0.228, 0.979]
3	Temporal filtering (T_0)	0.985	0.631 [0.103, 1.000]
4	Harmonised feature space	0.985	0.631 [0.103, 1.000]

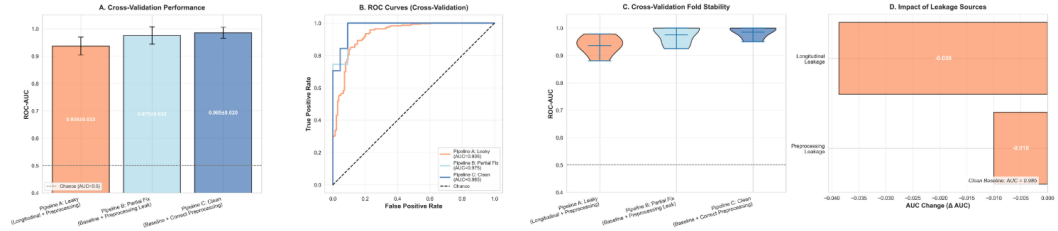

Figure 2: Progressive leakage ablation: single-run AUC appeared to increase (0.936→0.985) as leakage was removed; bootstrap analysis reveals this as a sampling artefact (Table 1).

To assess whether the overfitting pattern is model-specific, we compared XGBoost against simpler baselines under identical 5-fold stratified CV on the full dataset: Logistic Regression achieved AUC 0.725 (train 0.860, gap 0.134), Linear SVM scored AUC 0.703 (train 0.835, gap 0.131), and XGBoost reached AUC 0.793 (train 1.000, gap 0.207). All three models exhibited substantial train–test gaps, confirming that overfitting is a dataset-level phenomenon rather than an artefact of model complexity.

3.2 Learning-curve diagnostics

Bootstrapped learning curves (Figure 3) confirm structural overfitting. Training AUC remained at 1.0 across all sample sizes ($n = 20$ to 73), while validation AUC improved but the gap never closed. Visual extrapolation provides a heuristic suggesting that cohorts of at least $n >= 150$ may be required.

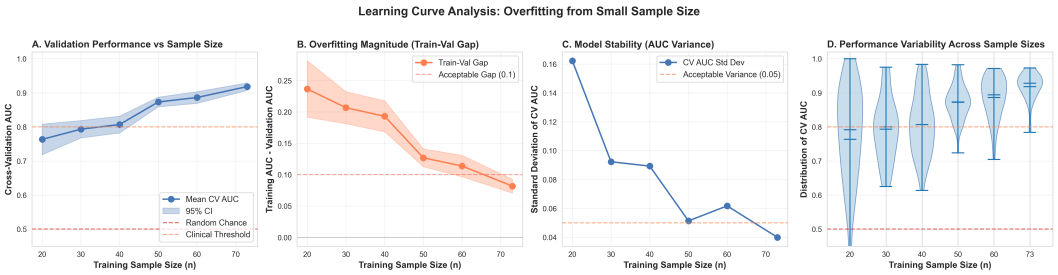

Figure 3: Bootstrapped learning curves: training AUC is high at all sample sizes while validation AUC improves; the gap persists, indicating continued overfitting.

3.3 Biological signal and external generalisation

Despite statistical limitations, SHAP attribution (Figure 4) consistently identified biologically plausible features: **bile acid metabolites** (glycocholic acid, glycodeoxycholic acid — 35% of top features), consistent with bile acid malabsorption in IBS-D [13]; **tryptophan metabolites** (tryptamine and serotonin precursors — 35%), reflecting gut–brain axis signalling; and **microbial genera** (*Faecalibacterium*, *Blautia* — 15%), matching known dysbiosis patterns. Bootstrap stability analysis ($N = 50$ iterations) confirmed that the top two features (*Prevotella* and hypoxanthine) appeared in the SHAP top-10 in 98% of iterations, and three further features exceeded 60% stability, supporting

the robustness of biological attribution even when AUC is unstable. External evaluation on three independent cohorts (Franzosa $N = 56$ [4], Jacobs $N = 36$ [5], iHMP $N = 278$ [6]) revealed that the microbiome-only model correctly identified all healthy controls but misclassified all IBD samples as IBS, indicating non-specific dysbiosis signal rather than IBS-specific discrimination.

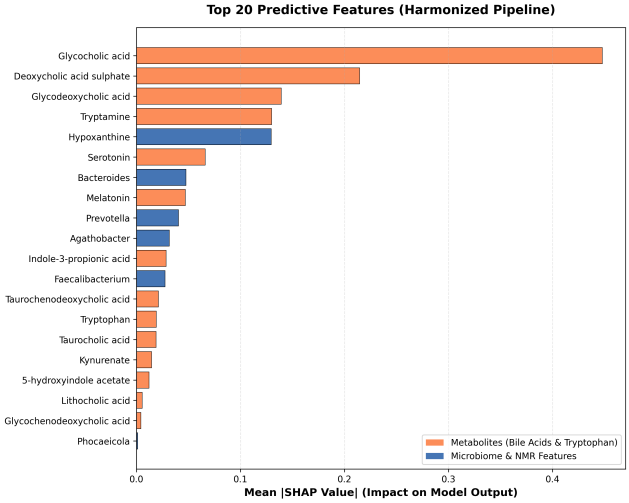

Figure 4: SHAP feature importance: bile acids dominate, followed by tryptophan metabolites and selected microbial genera.

4 Discussion and Conclusion

This study illustrates a fundamental tension in small-cohort biomedical ML: when $n < 100$, high AUC reflects memorisation, not generalisation. While a single evaluation suggested a paradoxical AUC increase under stricter validation, bootstrapping exposed this as a sampling artefact — the true mean AUC collapsed from 0.94 to 0.63 once subject leakage was removed, with 95% CIs spanning nearly the full range. Silhouette scores near zero ($s \approx -0.01$) confirm negligible class separability, and simpler baselines (Logistic Regression, Linear SVM) exhibited comparable overfitting gaps, demonstrating that the phenomenon is dataset-level rather than model-specific.

Critically, biological plausibility proved more stable than performance metrics under these conditions. SHAP attribution reproducibly surfaced bile acid and tryptophan pathways across validation schemes, providing independent corroboration aligned with established IBS mechanisms. This suggests that interpretability tools can serve as a *secondary validation signal* when hold-out metrics are unstable.

Three actionable findings emerge: **(1)** performance metrics alone are insufficient for small-cohort biomedical ML; leakage ablation and learning-curve analysis must accompany any AUC claim; **(2)** biological plausibility via feature attribution offers complementary stability evidence; **(3)** disease controls are essential during training to avoid confounding IBS-specific signals with general gut dysbiosis shared with IBD.

The main limitations are the small sample size ($n = 73$), absence of demographic fairness analysis, and the inability to include metabolomics features in external validation due to cross-cohort feature misalignment. Priorities for future work include multi-disease modelling frameworks that explicitly separate IBS from inflammatory disorders, larger harmonised cohorts with matched metabolomics, psychosocial covariates and longitudinal sampling, and fairness audits across demographic subgroups. The translational potential of fecal bile acid profiling as a complement to Rome IV criteria warrants prospective validation in IBS-D populations who may benefit from bile acid sequestrant therapy.

5 Supplementary Material

Code and figures are available at `https://github.com/nathanyaqueby/ibs-framework`

References

[1] Ioanna Aggeletopoulou, Christos Konstantakis, Stelios F Assimakopoulos, and Christos Triantos. The role of the gut microbiota in the treatment of inflammatory bowel diseases. *Microbial Pathogenesis*, 137:103774, 2019.

[2] A. S. Bhagavathula, K. Vidyasagar, and W. Tesfaye. Methodological quality of machine learning studies using microbiome data: a systematic review. *Frontiers in Microbiology*, 16:1516103, 2025.

[3] Olatomiwa O Bifarin. Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification. *Plos one*, 18(5):e0284315, 2023.

[4] Eric A Franzosa, Alexandra Sirota-Madi, Julian Avila-Pacheco, Nadine Fornelos, Henry J Haiser, Stefan Reinker, Tommi Vatanen, A Brantley Hall, Himel Mallick, Lauren J McIver, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature microbiology*, 4(2):293–305, 2019.

[5] Jonathan P Jacobs, Maryam Goudarzi, Namita Singh, Maomeng Tong, Ian H McHardy, Paul Ruegger, Miro Asadourian, Bo-Hyun Moon, Allyson Ayson, James Borneman, et al. A disease-associated microbial and metabolomics state in relatives of pediatric inflammatory bowel disease patients. *Cellular and molecular gastroenterology and hepatology*, 2(6):750–766, 2016.

[6] Jason Lloyd-Price, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, Nadim J Ajami, Kevin S Bonham, Colin J Brislawn, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, 2019.

[7] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[8] Ruben AT Mars, Yi Yang, Tonya Ward, Mo Houtti, Sambhawa Priya, Heather R Lekatz, Xiaojia Tang, Zhifu Sun, Krishna R Kalari, Tal Korem, et al. Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. *Cell*, 182(6):1460–1473, 2020.

[9] James T Morton, Jon Sanders, Robert A Quinn, Daniel McDonald, Antonio Gonzalez, Yoshiki Vázquez-Baeza, Jose A Navas-Molina, Se Jin Song, Jessica L Metcalf, Embriette R Hyde, et al. Balance trees reveal microbial niche differentiation. *MSystems*, 2(1):e00162–16, 2017.

[10] Efrat Muller, Yadid M Algavi, and Elhanan Borenstein. The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. *npj Biofilms and Microbiomes*, 8(1):79, 2022.

[11] Georgios Papoutsoglou, Sonia Tarazona, Marta B Lopes, Thomas Klammsteiner, Eliana Ibrahimi, Julia Eckenberger, Pierfrancesco Novielli, Alberto Tonda, Andrea Simeon, Rajesh Shigdel, et al. Machine learning approaches in microbiome research: challenges and best practices. *Frontiers in Microbiology*, 14:1261889, 2023.

[12] Sofia D Shaikh, Natalie Sun, Andrew Canakis, William Y Park, and Horst Christian Weber. Irritable bowel syndrome and the gut microbiome: a comprehensive review. *Journal of clinical medicine*, 12(7):2558, 2023.

[13] SA Slattery, O Niaz, Q Aziz, AC Ford, and AD Farmer. Systematic review with meta-analysis: the prevalence of bile acid malabsorption in the irritable bowel syndrome with diarrhoea. *Alimentary pharmacology & therapeutics*, 42(1):3–11, 2015.

[14] Marco Teixeira, Francisco Silva, Rui M Ferreira, Tania Pereira, Ceu Figueiredo, and Hélder P Oliveira. A review of machine learning methods for cancer characterization from microbiome data. *NPJ Precision Oncology*, 8(1):123, 2024.

[15] Leilei Yu, Yaru Liu, Shunhe Wang, Qingsong Zhang, Jianxin Zhao, Hao Zhang, Arjan Narbad, Fengwei Tian, Qixiao Zhai, and Wei Chen. Cholestasis: exploring the triangular relationship of gut microbiota-bile acid-cholestasis and the potential probiotic strategies. *Gut Microbes*, 15(1):2181930, 2023.