Grenze International Journal of Engineering and Technology, June Issue

Machine Learning Pipeline to Link Gut Microbiome Signature with Autism Spectrum Disorder

Afreena S[1], VE. Jayanthi[2] and Indhumathy[3]
[1-3]PSNA College of Engineering and Technology/Department of Biomedical Engineering, Dindigul, India
Email: Afreenasyed789@gmail.com, Hodbio@psnacet.edu.in, indhuharan@psnacet.edu.in

***Abstract*— Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition with growing evidence implicating the microbiota-gut-brain axis in its pathophysiology, highlighting the gut microbiome as a potential source of diagnostic biomarkers and therapeutic targets. Existing research is limited by cohort heterogeneity, inconsistent pre-processing pipelines, and poor cross-study reproducibility, which restrict the generalizability of reported microbial signatures. To address these gaps, the objective of this work is to develop a robust and interpretable machine learning pipeline. The pipeline is capable of identifying consistent gut microbiome signatures associated with ASD across independent datasets. This work integrates two publicly available cohorts GSE113690 (16S rRNA sequencing) and PRJNA516054 (shotgun metagenomics) harmonizes microbial features at the genus level. The harmonized features are applied for systematic pre-processing, feature extraction, and Random Forest classification, including permutation importance and SHAP analysis. The proposed model achieved an overall classification accuracy of 81%, demonstrating strong recall for ASD samples and identifying biologically relevant genera that consistently contributed to model predictions across datasets. These findings indicate that machine learning-driven integration of heterogeneous microbiome datasets can yield reproducible and interpretable microbial signatures, supporting the potential of gut microbiome profiling as a complementary tool for ASD diagnosis and advancing precision-based microbiome research in neurodevelopmental disorders.**

***Keywords*— Gut-Brain Axis, Microbiota, Neuro-inflammation, Stem Cell Therapy, Probiotics, Neurological Disorders.**

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition affecting approximately 1 in 100 children worldwide, with prevalence continuing to rise. ASD is clinically heterogeneous and is characterized by deficits in social interaction, communication challenges, and restricted or repetitive behaviors. Despite extensive research, the biological mechanisms underlying ASD remain incompletely understood, and diagnosis is primarily based on behavioral assessments rather than biological markers.

In recent years, growing attention has been directed toward the microbiota-gut-brain axis, a bidirectional communication network linking the gastrointestinal tract and the central nervous system. The gut microbiome influences brain development and function through immune modulation, metabolic pathways, and neuroactive compound production. Numerous studies have reported alterations in gut microbial composition in individuals with ASD, including changes in short-chain fatty acid–producing bacteria, immune-related taxa, and metabolic pathways.

Grenze ID: 01.GIJET.12.2.114_24

While individual studies have identified potential ASD-associated microbial signatures, results are often inconsistent across cohorts. These inconsistencies can be attributed to variations in sequencing platforms (16S rRNA versus metagenomics), geographic and dietary differences, age distributions, and analytical methodologies. Consequently, there is a need for integrative computational approaches that can handle heterogeneous datasets and extract generalizable patterns.

Machine learning (ML) techniques have emerged as powerful tools for analyzing high-dimensional microbiome data. Algorithms such as Random Forests can model complex non-linear relationships and identify discriminative microbial features. However, many existing ML-based ASD microbiome studies are limited to single cohorts, which restricts their generalizability. Moreover, the lack of interpretability in ML models has raised concerns regarding biological relevance.

In this work, address all these challenges by developing a machine learning pipeline that integrates multiple gut microbiome datasets to classify ASD and typically developing individuals. By combining 16S rRNA and metagenomic data at the genus level, applying rigorous feature filtering, and incorporating model interpretability through SHAP analysis, we aim to (i) improve cross-cohort classification performance and (ii) identify microbial genera that contribute most to ASD prediction. This work provides a comprehensive and interpretable framework for microbiome-based ASD research.

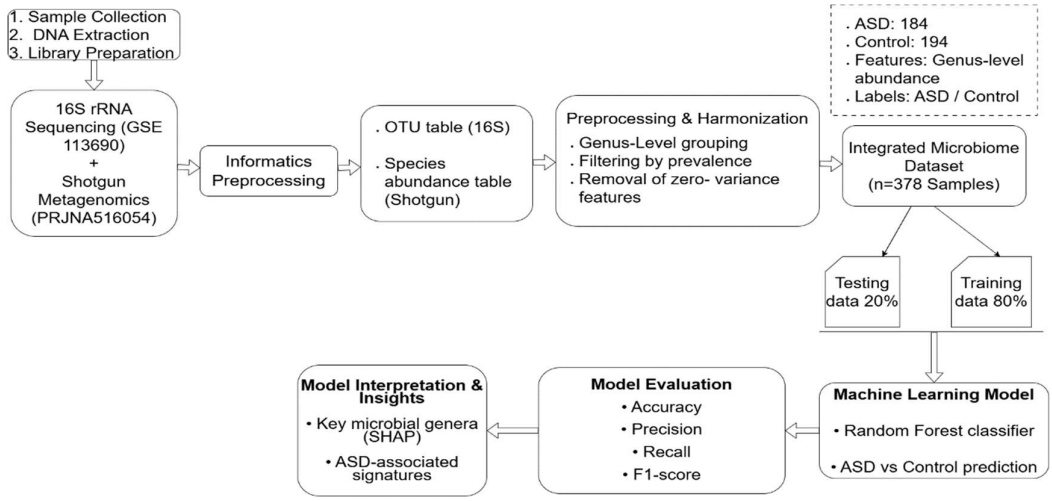

Figure 1. Block diagram

II. Materials and Methods

The work employs a computational pipeline to investigate the association between the gut microbiome and Autism Spectrum Disorder (ASD) using machine learning. Publicly available gut microbiome datasets generated using different sequencing platforms are integrated, harmonized at the genus level, and analyzed using a supervised classification approach. The workflow includes metadata processing, microbiome data preprocessing, dataset integration, feature filtering, normalization, machine learning model development, model evaluation, interpretation, and visualization. All the above mentioned steps are shown in the figure 2.1 and explanations are given in this section.

A. Metadata processing and dataset integration

Sample metadata from all datasets are curated to obtain diagnostic labels (ASD or Control). Sample identifiers are standardized by removing sequencing-specific suffixes and converting them into a consistent format to enable cross-dataset integration. Samples with missing or ambiguous diagnostic labels are retained but excluded from supervised model training.

Microbiome features from different sequencing platforms are harmonized at the genus level to ensure comparability. Taxonomic strings are parsed to extract genus-level annotations using standardized prefixes (e.g., g__). When genus-level information is unavailable, higher taxonomic levels are used or taxa are labeled as unclassified. Microbial abundance matrices are constructed for each dataset with samples as rows and genera as columns. Abundance values are converted into numeric format, and missing values are replaced with zeros.

Following preprocessing, datasets are merged by aligning shared and unique genera. The union of all detected genera is used as the feature space, and missing genera are assigned zero abundance values, enabling cross-cohort analysis while preserving dataset-specific microbial signatures.

B. Feature Filtering and Normalization

To reduce noise and improve model performance, feature filtering is applied. Genera present in fewer than 2% of samples are removed (low-prevalence filtering), and genera with zero variance across samples are excluded.
Microbial abundance values are then normalized to relative abundances by dividing each genus count by the total abundance per sample. A logarithmic transformation (log1p) is applied to stabilize variance and reduce the influence of highly abundant taxa.

C. Machine Learning, Evaluation, and Interpretation

A Random Forest (RF) classifier is used due to its robustness to high-dimensional data and ability to model non-linear relationships. Only samples labelled as ASD or Control are included in supervised learning. The dataset is split into training and testing sets using a stratified 80:20 ratio to maintain class balance. Hyperparameter optimization is performed using Randomized Search CV.
Model performance is evaluated using accuracy, precision, recall, and F1-score, with particular emphasis on recall for the ASD class.
Model interpretability is achieved using SHapley Additive exPlanations (SHAP), which quantify the contribution of each microbial genus to model predictions. Exploratory data analysis is conducted using Principal Component Analysis (PCA) to visualize sample clustering, along with heatmaps and hierarchical clustering to assess microbial abundance patterns.

D. Dataset Collection

Two publicly available gut microbiome datasets are used for this work. The GSE113690 dataset is obtained from the Gene Expression Omnibus (GEO) database and consists of 16S rRNA gene sequencing-based operational taxonomic unit (OTU) profiles derived from fecal samples of individuals with ASD and typically developing controls. Sample metadata, including diagnostic labels, are retrieved from the associated supplementary files.
The PRJNA516054 dataset is retrieved from the NCBI Sequence Read Archive (SRA) and comprises whole-metagenome sequencing (WMS) data from fecal samples of ASD and control subjects. Microbial relative abundance profiles and corresponding metadata are extracted from the supplementary materials of the original study.

III. Results and Discussion

A robust and interpretable machine learning pipeline is developed for identifying consistent gut microbiome signatures associated with ASD across independent datasets. The integrated analysis reveals meaningful microbial patterns and classification performance, as discussed below.

A. Dataset Integration and Feature Summary

Following preprocessing and taxonomic harmonization, the integrated dataset consists of 378 samples comprising ASD and typically developing controls. The merged feature space includes a wide range of microbial genera prior to filtering.
Application of prevalence-based and variance-based filtering reduces dimensionality while retaining biologically relevant taxa. The refined dataset provides a stable and informative feature set for downstream analysis.
The ability to integrate datasets generated from different sequencing platforms highlights the feasibility of genus-level harmonization for cross-cohort microbiome studies, despite known variability introduced by technical and population differences.

B. Classification Performance of the Machine Learning Model

A Random Forest classifier is trained on the integrated dataset to distinguish individuals with ASD from typically developing controls. Model performance is evaluated on a held-out test set using standard classification metrics.
The tuned model achieved an overall classification accuracy of 81%, indicating strong predictive performance across heterogeneous datasets. Detailed performance metrics are summarized below:
The performance metrics are defined as follows, where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively:

Accuracy = (TP + TN) / (TP + TN + FP + FN). (1)
ASD Recall = TP / (TP + FN) . (2)
ASD Precision = TP / (TP + FP). (3)
Control Recall = TN / (TN + FP). (4)
Control Precision = TN / (TN + FN). (5)
F1-score = 2 × (Precision × Recall) / (Precision + Recall). (6)
Weighted F1-score = Σ ($w_i \times F1_i$). (7)

- Accuracy: 81%
- ASD Recall: 91%
- ASD Precision: 75%
- Control Recall: 71%
- Control Precision: 90%
- Weighted F1-score: 0.81

Based on the above metrics, the model achieved an accuracy of 81%, ASD recall of 91%, ASD precision of 75%, control recall of 71%, control precision of 90%, and a weighted F1-score of 0.81.

Notably, the model demonstrated high recall for ASD samples, correctly identifying the majority of affected individuals. This is particularly important in clinical and screening contexts. where minimizing false negatives is critical.

Despite the inherent challenges of microbiome-based classification, including high dimensionality, sparsity, and batch effects, the model maintains stable performance. This suggests that the learned patterns reflect generalizable microbial signatures rather than dataset-specific noise.

C. Comparison with Cross-Dataset Expectations

The classification task involves integration of datasets generated using different sequencing platforms and populations, which introduces inherent variability and potential batch effects.

Previous studies have reported classification accuracies typically ranging from 60% to 80%, with reduced performance in cross-cohort settings. In this context, the achieved accuracy of 81% falls at the upper range of reported results.

Despite these challenges, the achieved accuracy of 81% falls at the upper range of performance reported in previous ASD microbiome studies, highlighting the robustness and generalizability of the proposed pipeline.

D. Feature Importance and Model Interpretability

SHapley Additive exPlanations (SHAP) analysis reveals the contribution of individual microbial genera to model predictions. The SHAP summary plot identifies a subset of genera that consistently influence ASD classification, with both increased and decreased abundances contributing to predictions. These findings indicate that the model captures biologically meaningful microbial patterns rather than noise-driven signals, as shown in Figure 2.

Rather than relying on a single dominant taxon, the model leverages combinatorial microbial patterns, reflecting the multifactorial nature of ASD-associated gut dysbiosis. These findings support the hypothesis that microbial dysbiosis contributes to ASD through complex, multi-pathway interactions rather than single-organism effects.

E. Principal Component Analysis

Principal Component Analysis (PCA) of log-transformed relative abundance data reveals partial separation between ASD and control samples.

Although overlap is observed due to inherent biological variability, the clustering patterns suggest underlying compositional differences between groups. These differences are more effectively captured by the supervised machine learning model Figure 3.

F. Heatmap and Clustering Analysis

Hierarchical clustering and heatmap visualization of highly variable genera highlight distinct abundance patterns between ASD and control samples.

Clusters of samples exhibit consistent genus-level profiles associated with diagnostic status, supporting the presence of altered microbial composition in ASD. These multivariate patterns reinforce the importance of combined microbial signatures in disease classification Figure 4.

G. Summary of Key Findings

Integration of heterogeneous microbiome datasets at the genus level is effective for cross-cohort analysis. The proposed machine learning model achieves 81% classification accuracy, with high ASD recall (91%) indicating

strong sensitivity. SHAP analysis identifies key microbial contributors, while PCA and heatmap analyses support the presence of compositional differences between ASD and control groups.

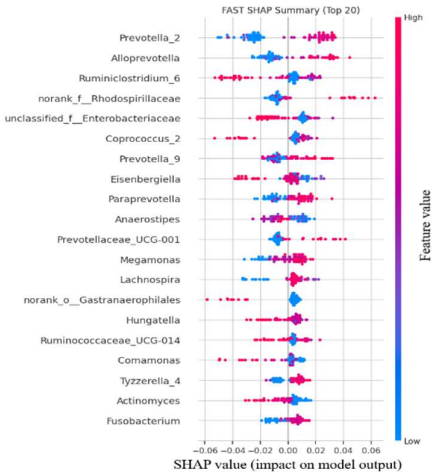

Figure 2. SHAP summary plot showing the top 20 microbial genera contributing to ASD classification. Each point represents a sample, with color indicating feature value (low to high) and position representing the impact on model output.

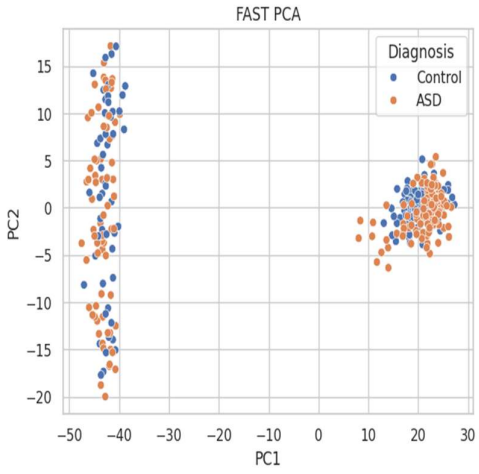

Figure 3. Principal Component Analysis (PCA) plot of microbiome samples based on log-transformed relative abundance data. Each point represents a sample, colored by diagnostic group (ASD vs Control), showing partial separation between classes.

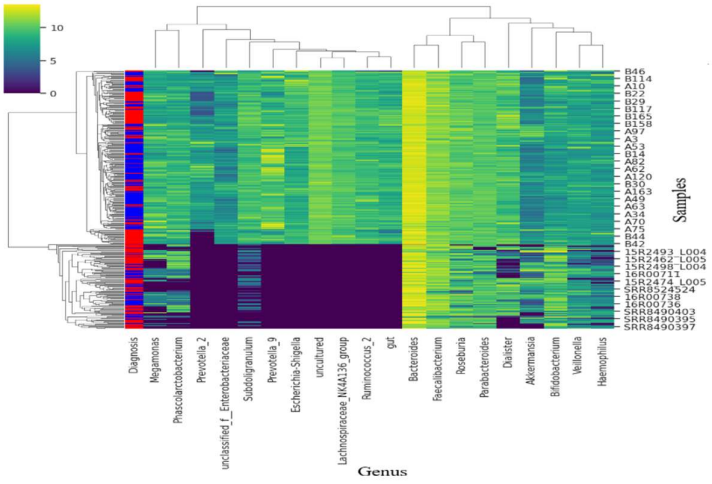

Figure 4. Heatmap of microbial genus abundance with hierarchical clustering of samples and features. Color intensity represents normalized abundance levels, highlighting distinct microbial patterns between ASD and control groups.

IV. Conclusion

In this work, we aimed to investigate the association between the gut microbiome and Autism Spectrum Disorder (ASD) using an integrated multi-cohort analytical framework. To achieve this, a machine learning–based pipeline was developed and implemented, leveraging a Random Forest classifier to model complex, non-linear relationships within microbiome data. Publicly available datasets generated using 16S rRNA gene sequencing and whole-metagenome sequencing were integrated at the genus level, enabling cross-platform analysis despite differences in sequencing technologies and cohort characteristics.

The findings indicate that ASD classification is driven by combinatorial patterns of microbial genera rather than individual taxa, supporting the hypothesis that ASD is associated with community-level gut microbial alterations linked to the microbiota–gut–brain axis. Exploratory analyses further revealed compositional differences between ASD and control groups, although partial overlap reflects the inherent heterogeneity of the disorder.

The proposed model achieved a classification accuracy of 81%, with high sensitivity for ASD samples (91%), demonstrating robust and generalizable predictive performance across heterogeneous datasets.

Despite these promising results, several limitations should be acknowledged. Genus-level harmonization may obscure species-level variations, and inconsistencies in metadata across datasets may affect model training. Additionally, important clinical and environmental factors such as diet, medication, and symptom severity were not incorporated into the analysis.

Future work will focus on validating the model across larger and more diverse cohorts, integrating multi-omics data such as metabolomics, and incorporating advanced machine learning approaches to further improve predictive performance and biological interpretability. These efforts will enhance the applicability of microbiome-based approaches in ASD research and precision neurodevelopmental studies.

ACKNOWLEDGMENT

The authors gratefully acknowledge the computational resources and institutional support provided by the PSNA College of Engineering and Technology. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors declare no conflict of interest.

REFERENCES

[1] A. R. L. Toledo et al., "Gut–Brain Axis as a Pathological and Therapeutic Target for Neurodegenerative Disorders," Int. J. Mol. Sci., vol. 23, no. 3, p. 1184, 2022.
[2] J. F. Cryan et al., "The microbiota–gut–brain axis," Physiol. Rev., vol. 99, no. 4, pp. 1877–2013, 2019.
[3] Y. He et al., "Microbiota-gut-brain axis in health and neurological disease: Interactions between gut microbiota and the nervous system," J. Cell. Mol. Med., vol. 28, no. 18, 2024.
[4] R. Zhong et al., "The Oral–Gut–Brain Axis: The Influence of Microbes as a Link of Periodontitis With Ischemic Stroke," CNS Neurosci. Ther., vol. 30, no. 12, 2024.
[5] J. Sharon et al., "The central nervous system and the gut microbiome," Cell, vol. 167, no. 4, pp. 915–932, 2016.
[6] T. G. Dinan and J. F. Cryan, "Gut–brain axis and neuropsychiatric disorders," Curr. Opin. Clin. Nutr. Metab. Care, vol. 20, no. 6, pp. 515–519, 2017.
[7] M. E. Mitchell et al., "Gastrointestinal, Behaviour and Anxiety Outcomes in Autistic Children Following Synbiotics and Gut-Directed Hypnotherapy," J. Autism Dev. Disord., 2024.
[8] E. K. Kang et al., "Akkermansia muciniphila improves cognitive dysfunction by regulating BDNF and serotonin pathways," Microbiome, vol. 12, 2024.
[9] M. N. Ekwudo, C. Gubert, and A. J. Hannan, "The microbiota–gut–brain axis in Huntington's disease: pathogenic mechanisms and therapeutic targets," FEBS J., vol. 292, no. 6, pp. 1282–1315, 2025.
[10] S. Noronha et al., "High-fat diet, microbiome-gut-brain axis signaling, and anxiety-like behavior in male rats," Biol. Res., vol. 57, p. 23, 2024.
[11] F. Donoso et al., "Inflammation, Lifestyle Factors, and the Microbiome–Gut–Brain Axis: Relevance to Depression and Antidepressant Action," Clin. Pharmacol. Ther., vol. 113, pp. 246–259, 2022.
[12] G. S. Clarke et al., "The gut-brain axis in appetite, satiety, food intake, and eating behavior," Pharmacol. Res. Perspect., vol. 12, no. 5, 2024.
[13] Z. Qu et al., "Feeding the microbiota–gut–brain axis: Nucleotides and their role in early life," Food Frontiers, vol. 4, 2023.
[14] B. Maitra and O. B. Akan, "Molecular Quantum Communication Channel in the Gut–Brain Axis Synapse," IEEE Trans. Mol. Biol. Multi-Scale Commun., vol. 10, no. 4, pp. 604–612, Dec. 2024.
[15] E. M. Ekici and Ö. M. Çelik, "Night eating behavior, gastrointestinal symptoms, and psychological well-being," J. Eat. Disord., vol. 13, 2025.
[16] C. J. Bates, "Targeting the Gut to Treat the Brain," IEEE Pulse, vol. 12, no. 4, pp. 2–5, 2021.
[17] M. A. Schächtle and S. P. Rosshart, "The Microbiota–Gut–Brain Axis in Health and Disease and Its Implications for Translational Research," Front. Cell. Neurosci., vol. 15, p. 698172, 2021.
[18] H. Jiang et al., "Altered fecal microbiota composition in patients with major depressive disorder," Brain Behav. Immun., vol. 48, pp. 186–194, 2015.
[19] M. E. Evrensel and M. E. Ceylan, "The Gut–Brain Axis: The Missing Link in Depression," Clin. Psychopharmacol. Neurosci., vol. 13, no. 3, pp. 239–244, 2015.
[20] N. Prince et al., "Mouse strain-specific responses along the gut–brain axis upon fecal microbiota transplantation from children with autism," Gut Microbes, vol. 17, no. 1, 2025.
[21] L. Diaz-Marugan et al., "Microbiota, diet, and the gut–brain axis in multiple sclerosis and stroke," Eur. J. Immunol., vol. 53, no. 11, 2023.
[22] K. J. O'Riordan et al., "The gut microbiota–immune–brain axis: Therapeutic implications," Cell Rep. Med., vol. 6, no. 3, p. 101982, 2025.

[23] J. A. Foster and K.-A. McVey Neufeld, “Gut–brain axis: How the microbiome influences anxiety and depression,” Trends Neurosci., vol. 36, no. 5, pp. 305–312, 2013.
[24] S. N. Overton et al., “Serotonin Sensing Technologies to Promote Understanding of the Gut–Brain Axis,” IEEE Sens. Lett., vol. 8, no. 6, 2024.
[25] X. Zhao et al., “High-tannin food enhances spatial memory via the microbiota–gut–brain axis,” Microbiome, vol. 12, 2024.
[26] T. G. Dinan, K. J. O’Riordan, K. Sandhu, V. Peterson, and J. F. Cryan, “The microbiota–gut–brain axis,” Physiol. Rev., vol. 99, no. 4, pp. 1877–2013, 2019.
[27] M. Bates, "Gut Feeling: Researchers Are Discovering a Link Between Depression and Gut Bacteria," in IEEE Pulse, vol. 8, no. 6, pp. 15-18, Nov.-Dec. 2017.
[28] Shao, L., Cai, G., Fu, J. et al. Gut microbial ‘TNFα-sphingolipids-steroid hormones’ axis in children with autism spectrum disorder: an insight from meta-omics analysis. J Transl Med 22, 1165 (2024)
[29] M. Bates, “Gut Feeling: Researchers Are Discovering a Link Between Depression and Gut Bacteria,” IEEE Pulse, vol. 8, no. 6, pp. 15–18, 2017.
[30] X. Wang et al., “The Bidirectional Signal Communication of Microbiota–Gut–Brain Axis in Hypertension,” Int. J. Hypertens., vol. 2021, 8174789, 2021.
[31] M. Sanders et al., “Probiotics and prebiotics in intestinal health and disease: From biology to the clinic,” Nat. Rev. Gastroenterol. Hepatol., vol. 16, no. 10, pp. 605–616, 2019.
[32] A. Evrensel et al., “The Gut–Brain Axis: The Missing Link in Depression,” Clin. Psychopharmacol. Neurosci., vol. 13, no. 3, pp. 239–244, 2015.
[33] G. Lee et al., “IUPHAR review: Microbiota–gut–brain axis and its role in neuropsychiatric disorders,” Pharmacol. Res., vol. 216, 2025.
[34] B. R. Prince, “The Gut–Brain–Microbiota Connection and Its Role in Autism Spectrum Disorders,” Nutrients, vol. 17, no. 7, 2025.
[35] A. J. Schneider et al., “A Role for the Microbiota–Gut–Brain Axis in Avoidant/Restrictive Food Intake Disorder,” Int. J. Eat. Disord., vol. 57, no. 12, 2024.