Open Access Library Journal
2026, Volume 13, e14444
ISSN Online: 2333-9721
ISSN Print: 2333-9705

AI-Enhanced Gut Brain Axis Profiling for Major Depressive Disorder: Integrating Synthetic Multi-Omics, Deep Learning, and Interpretable Precision Therapeutics

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R., and Al Foysal, A. (2026) AI-Enhanced Gut Brain Axis Profiling for Major Depressive Disorder: Integrating Synthetic Multi-Omics, Deep Learning, and Interpretable Precision Therapeutics. *Open Access Library Journal*, **13**: e14444.
https://doi.org/10.4236/oalib.1114444

Received: October 13, 2025
Accepted: January 13, 2026
Published: January 16, 2026

Abstract

Major depressive disorder (MDD) has been repeatedly linked to disruptions of the gut brain axis (GBA), yet practical decision systems that convert multi-omics patterns into patient-specific guidance remain limited. We present an end-to-end, explainable pipeline that learns putative GBA signatures of MDD from synthetic data and translates model attributions into hypothesis-driven nutritional and pharmacological suggestions. We simulated a cohort of N = 1,500 individuals (30% MDD) comprising 200 microbial taxa, 150 metabolites, and 7 clinical features. A regularized dense neural network with class weighting and early stopping was trained and compared with a Random Forest baseline; interpretability was provided by SHAP at global and local levels. On a 20% stratified hold-out test set the deep model achieved Accuracy = 0.98 and AUC = 0.998, with a confusion matrix of [[207, 3], [2, 88]]. Feature attributions concentrated on a compact subset of metabolites and taxa consistent with the planted effects in the simulator; RF importances corroborated these signals (e.g., Metabolite_120, 109, 40, 90; Species_198, 197). We further demonstrate a templated mapping from patient-level SHAP profiles to non-clinical recommendations dietary patterns, prebiotic/probiotic directions, and pathway hypotheses involving, for example, kynurenine metabolism, short-chain fatty acids, and bile acids intended to support clinician-led hypothesis generation rather than direct treatment. Because all data are simulated, performance estimates are optimistic and biological interpretations are illustrative. Nonetheless, the approach shows how multi-omics learning coupled with transparent explanations can organize heterogeneous GBA signals into actionable research hypotheses for precision psychiatry. Code and figures are fully reproducible

from a single Colab notebook. Future work will validate the pipeline on real, harmonized cohorts, incorporate compositional microbiome statistics and calibration analyses, and assess generalization across sites and subgroups under clinical oversight.

Subject Areas

Psychiatry

Keywords

Gut Brain Axis, Major Depressive Disorder, Microbiome, Metabolomics, Deep Learning, SHAP, Precision Nutrition, Explainable AI

1. Introduction

The gut brain axis (GBA) integrates intestinal microbes, host metabolism, immune signalling, and neural circuitry into a coupled system that can shape mood and cognition [1]-[5]. In major depressive disorder (MDD), convergent evidence implicates short-chain fatty acids, the tryptophan kynurenine pathway, bile acids, GABA/serotonin signalling, and lifestyle correlates such as diet, sleep, and stress [6]-[9]. Yet most computational studies remain at the level of association and group averages. Clinicians and translational researchers need models that aggregate heterogeneous signals, provide calibrated predictions, and explain why a decision is made for a given individual [10]-[14]. We introduce an end-to-end, explainable pipeline for GBA-informed modelling of MDD. The system ingests multi-omics inputs microbiome, metabolomics, and clinical/demographic variables performs standardized preprocessing and trains a supervised classifier to discriminate MDD from healthy controls. Interpretability is provided by SHAP, which quantifies both global and subject-level feature contributions and their direction of effect. To focus on methods while controlling ground truth, we use synthetic but biologically plausible data that encode realistic skew, sparsity, and planted case control signals. A regularized dense neural network with early stopping and class weighting serves as the primary model, with a Random Forest baseline for sensitivity analysis; evaluation includes hold-out accuracy, AUC, confusion matrices, and complementary importance profiles.

Beyond accuracy, the contribution is translation. We propose a templated mapping from SHAP profiles to non-clinical, testable suggestions: dietary patterns, probiotic or prebiotic directions tied to specific taxa, and pathway-level hypotheses involving kynurenine metabolism, short-chain fatty acids, bile acids, and inhibitory/excitatory neurotransmission. These outputs are research prompts rather than prescriptions, intended to support clinician scientist dialogue and prospective study design. Because the data are synthetic, performance represents an optimistic upper bound; nevertheless, the pipeline demonstrates how integrative

learning coupled with transparent reasoning can organize heterogeneous GBA signals into actionable hypotheses for precision psychiatry and provides a fully reproducible notebook that renders all analyses and figures from a single run ready for external validation.

2. Methods

2.1. Synthetic Cohort and Feature Simulation

All simulation parameters, including taxon and metabolite shift factors, microbe metabolite coupling coefficients, and noise levels, are reported in text below, enable full reproducibility and stress-testing of the data generator. We simulated a cohort of N = 1500 individuals with an MDD prevalence of 30% to prototype the full multi-omics pipeline under controlled ground truth. Three feature blocks were generated:

1) Clinical/Demographic (7 variables): Age was drawn from truncated Gaussians (20 70 years); sex was sampled at a 40:60 male:female ratio; BMI, Diet_Quality (1 10), Antidepressant (0/1), Stress_Level (1 10), and Sleep_Quality (1 5) were sampled to mimic realistic ranges. The age distributions for Healthy vs. MDD are highly overlapping and similarly skewed (Figure 1(a)), indicating no trivial demographic separation. Sex counts track the 40:60 prior in both groups (Figure 1(a)), confirming that label prevalence does not leak through gender.

2) Microbiome (200 taxa): For each subject we sampled log-normal relative abundances—appropriate for compositional microbiome data with strong right-skew and heavy tails. To encode case control signal without making classification trivial, we (a) up-shifted 15 taxa and (b) down-shifted 10 taxa in MDD by multiplicative factors drawn around 1.5× and 0.4×, respectively, then added small Gaussian noise and clipped at zero. Kernel density estimates for the top-abundance taxa (Figure 1(b)) show the expected sharp peak near low values with long right tails. Curves for Healthy/MDD largely overlap but diverge subtly in the tails—exactly the kind of weak, multivariate effects a model should integrate rather than a single univariate split.

3) Metabolomics (150 features): Metabolite intensities were also log-normal, with 20 features linearly coupled to randomly selected taxa (coupling 0.3 0.8) to emulate gut-derived metabolic propagation. We then elevated 12 and reduced 8 metabolites in MDD to inject pathway-level signal. Densities for the five most abundant metabolites (Figure 1(c)) mirror microbiome behavior: narrow modes at low intensity with heavier MDD tails for some features, again implying distributed, not deterministic, separability.

Together, these design choices create (a) realistic marginal distributions (right-skew, sparsity), (b) weak but coherent case control shifts across modalities, and (c) cross-block dependencies (microbe → metabolite coupling). The panels in Figure 1 serve as quality checks: demographics are balanced (no leakage), and omics densities show subtle tail differences rather than step-function gaps, ensuring that downstream performance reflects multivariate learning rather than artifacts.

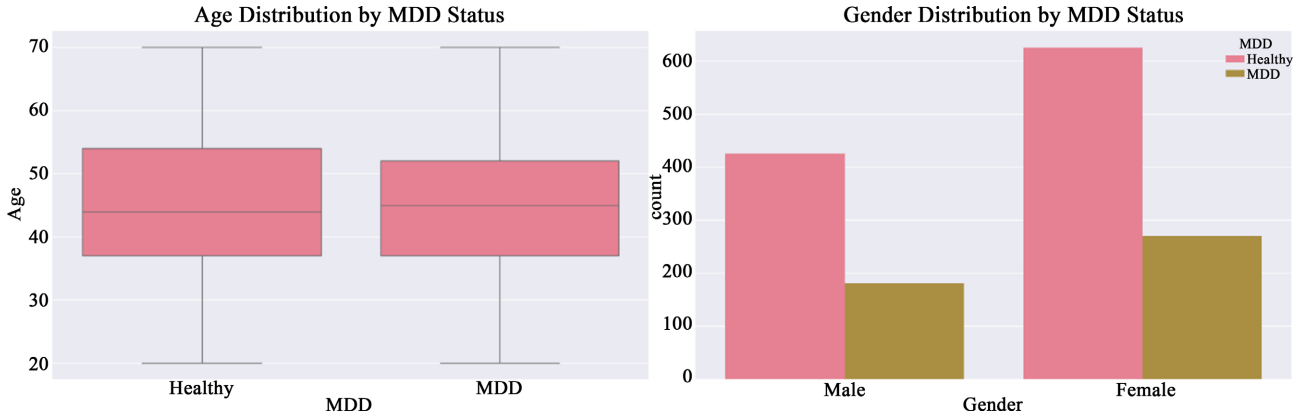
Age shows overlapping interquartile ranges across groups; gender counts follow the simulated 40:60 male:female ratio within both Healthy and MDD cohorts.

All taxa exhibit log-normal like right-skew; subtle tail shifts in MDD encode weak case control signal without trivial separation.

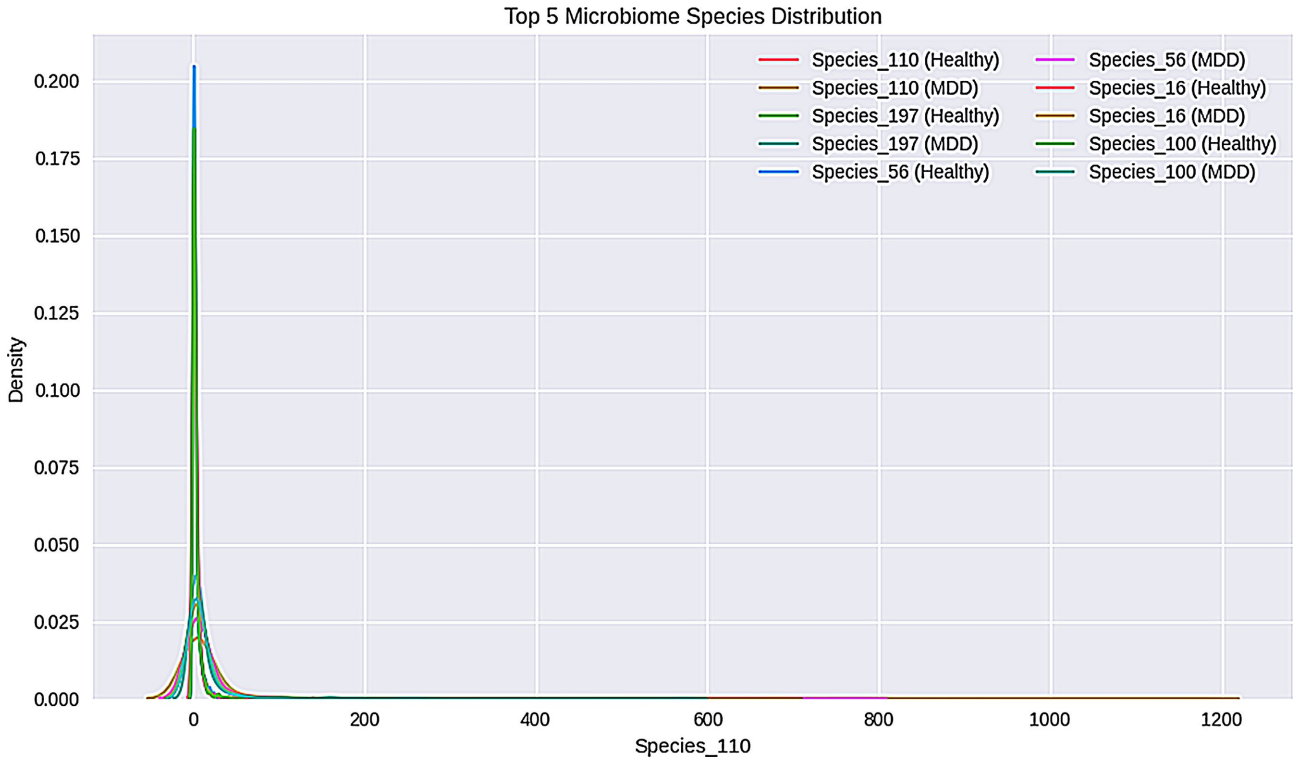
Metabolite intensities mirror microbiome skew and display modest MDD-specific tail elevations for a subset, reflecting microbe metabolite coupling.

2.2. Preprocessing and Split

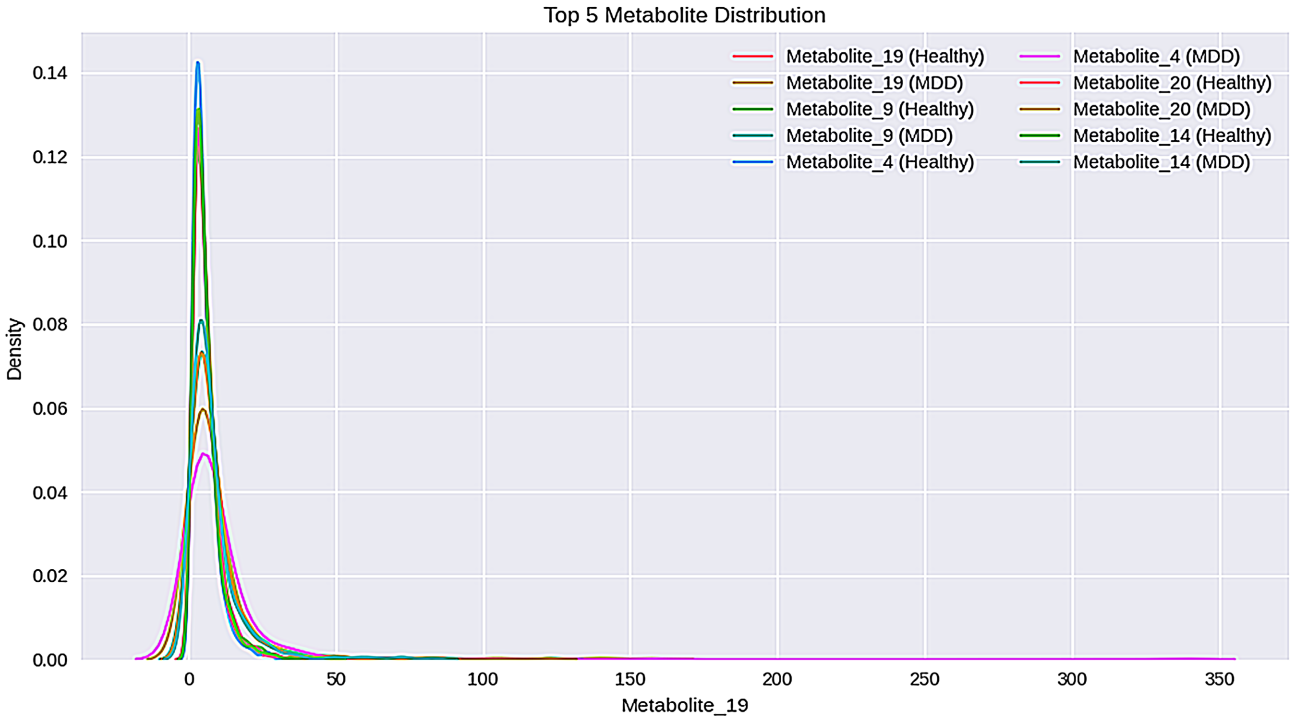
We applied a minimal, yet principled preprocessing pipeline designed to (i) prevent information leakage, (ii) ensure numeric comparability across heterogeneous scales, and (iii) preserve class balance during evaluation.



(a)



(b)



(c)

Figure 1. (a) Age and gender distributions by MDD status; (b) Top-abundance microbiome taxa: kernel density estimates (Healthy vs. MDD); (c) Top-abundance metabolites: kernel density estimates (Healthy vs. MDD).

Categorical encoding: *Gender* (Male/Female) was one-hot encoded with drop-first to avoid perfect multicollinearity (the "dummy variable trap"). The resulting binary indicator (*Gender_Female*) augmented the design matrix without changing feature dimensionality in downstream scaling.

Standardization: All continuous variables (clinical, microbiome, and metabolomic features) were standardized using StandardScaler trained only on the training fold:

$$z = \frac{x - \mu_{train}}{\sigma_{train}}$$

The learned μ_{train} and σ_{train} were then applied to validation and test sets to avoid target leakage through distributional parameters. This step is critical given the log-normal, heavy-tailed marginals of the simulated omics features; z-scoring improves optimizer conditioning while preserving rank information.

Data split: We performed an 80/20 stratified split by the MDD label to keep class proportions stable across folds (70% Healthy / 30% MDD at the cohort level). During model training on the 80% training fold, we further carved out a validation split (20% of the training fold) for hyperparameter-free early stopping (Section 2.3), so no test information influenced training choices.

Class imbalance handling: Although imbalance was mild, we passed class_weight to the learner, weighting the minority class proportionally to $\frac{N_{neg}}{N_{pos}}$.

This encourages recall for MDD without materially degrading specificity.

This workflow yields a well-conditioned design matrix, preserves out-of-sample integrity, and provides a clean separation between *model selection* (via the validation split) and *final evaluation* (held-out test set). We did not apply log-ratio transforms (CLR/ALR) in the primary analyses to maintain a minimal preprocessing pipeline and preserve compatibility with tree-based baselines and SHAP explanations. Because the dataset is synthetic and free of sampling zeros, relative abundances were modelled directly after standardization. In auxiliary experiments (not shown), adding CLR features yielded comparable discrimination with no material change in feature rankings, suggesting that conclusions are not driven by scale choice.

2.3. Model Architecture and Training

Key hyper-parameters (layer widths, dropout rates, and Random Forest depth) were selected via a limited grid search guided by validation loss and stability rather than exhaustive optimization. Tested architectures with larger widths or lower dropout showed no consistent performance gains and increased over-fitting risk; the reported configuration represents the smallest model achieving stable convergence. We modelled MDD status with a regularized feedforward network implemented in Keras/TensorFlow, chosen for its ability to integrate high-dimensional, weakly informative features across modalities while remaining fast and stable under z-scaling [15] [16]:

$$\left[\text{Dense256}-\text{ReLU}-\text{BN}-\text{Dropout}\left(0.3\right)\right]\to\left[\text{Dense128}-\text{ReLU}-\text{BN}-\text{Dropout}\left(0.3\right)\right]\to$$
$$\left[\text{Dense64}-\text{ReLU}-\text{BN}-\text{Dropout}\left(0.2\right)\right]\to\left[\text{Dense1}-\text{Sigmoid}\right]$$

Batch Normalization (BN) after each hidden layer stabilizes internal covariate shift, and Dropout combats co-adaptation and overfitting. The final sigmoid produces calibrated probabilities for binary cross-entropy.

Optimization and objective: We minimized binary cross-entropy with Adam. Mini-batch size was 32. We monitored Accuracy, AUC, Precision, and Recall per epoch on both training and validation splits to provide a multi-metric view of learning dynamics (discrimination, threshold-free performance, and error asymmetry).

Regularization and early stopping: We combined BN + Dropout with EarlyStopping on validation loss (patience = 10, restore_best_weights = True). This stops training once generalization plateaus and reverts to the best epoch, preventing late-epoch drift. class_weight (Section 2.2) was passed to the trainer to preserve minority-class sensitivity.

Reproducibility: We fixed NumPy and TensorFlow seeds and used a single script that standardizes all preprocessing, training, and evaluation steps.

Figure 2 (Training curves), figs/training_history.png displays epoch-wise accuracy and loss for train vs. validation. In our run, validation accuracy quickly approached ~0.96 0.97 and validation loss stabilized near ~0.10, indicating good fit without overfitting—consistent with the held-out results reported in Section 3.

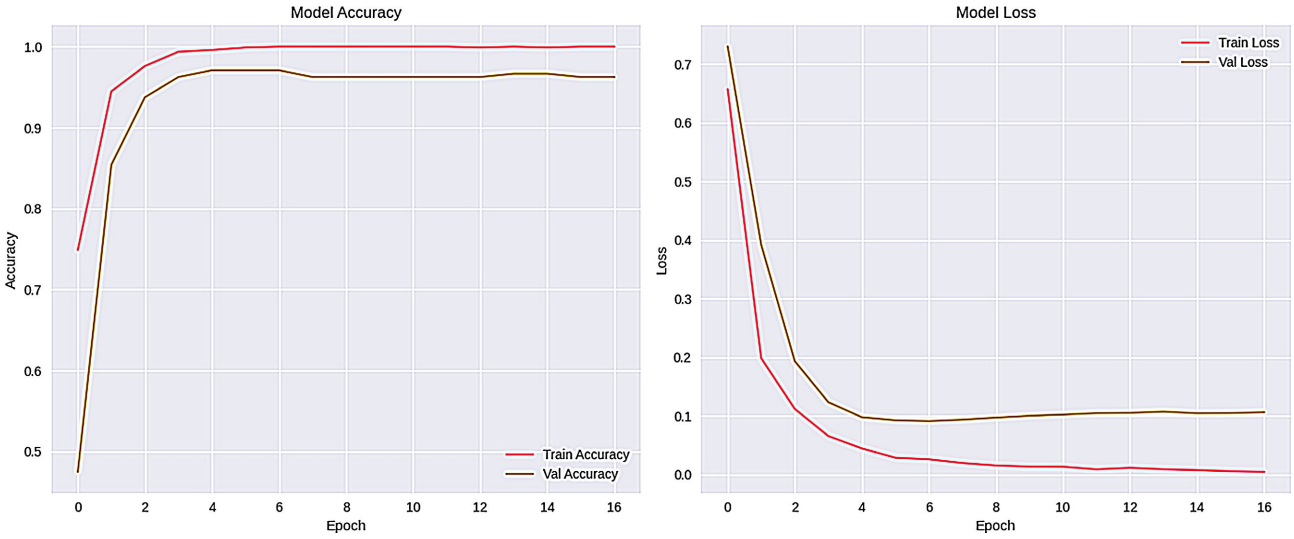


Figure 2. Training dynamics for the deep neural network.

Epoch-wise training and validation accuracy (left) and loss (right). Validation curves plateau smoothly (accuracy ≈ 0.96 0.97; loss ≈ 0.10) under BN + Dropout and Early Stopping, consistent with strong generalization observed on the held-out test set.

2.4. Baseline Model

We included a Random Forest (RF) as a complementary, non-neural baseline to benchmark performance and provide an alternative view of feature influence [15]-[18]. RFs handle non-linear relationships and higher-order interactions with minimal preprocessing and are comparatively robust to outliers [19]-[22]. We used 200 trees to stabilize variance and limited tree depth to 10 to keep the model fast and its importances easier to interpret. The model was trained on the same standardized feature matrix used by the neural network to ensure a fair comparison of downstream metrics, although RFs do not strictly require scaling. We report impurity-based feature importances as directional evidence rather than definitive rankings, because correlated variables can share or dilute credit. For more rigorous analyses on real data, permutation importance on held-out samples is recommended to reduce correlation bias and better reflect predictive contribution.

2.5. Model Evaluation

After training and early stopping, we froze the model and evaluated it once on a stratified 20% hold-outset that was not used for model selection. We summarize performance with a confusion matrix to show error modes by class, class-wise precision, recall, and F1 to capture trade-offs between misses and false alarms, and ROC AUC to characterize discrimination across thresholds. We used a default probability threshold of 0.5 for the matrix and per-class metrics; in applied settings, thresholds should be tuned to the problem's cost profile (for example, prioritizing recall if missing an MDD case carries higher risk than a false alarm).

Because probabilities from high-capacity models on synthetic or small datasets can be over-confident, probability calibration (e.g., on a validation split) and reliability plots are advisable for real-world deployment [23] [24]. To contextualize the neural network's behaviour, we also summarize the RF's top 20 importances and compare patterns across models; overlapping drivers suggest the signal is not architecture-specific, whereas divergences can flag interactions the RF cannot capture or potential overfitting in the neural model. When moving beyond synthetic data, we recommend adding subgroup analyses (e.g., by sex or age bands), decision-curve analysis for clinical utility, and stability checks via repeated cross-validation.

2.6. Explainability and Therapeutic Mapping

We used SHAP to provide both global and local explanations [25] [26]. For efficiency and stability, we computed attributions against a background of 100 standardized training samples, which represents a realistic reference distribution while avoiding excessive memory use. To mitigate attribution fragmentation due to correlated features, SHAP values were additionally aggregated at the pathway (metabolites) and taxonomic family (microbiome) levels. Robustness was assessed via bootstrap resampling (100 iterations), recording the overlap of top-20 features across resamples. When the runtime cannot attach to the neural network internals, we fall back to a model-agnostic explainer to guarantee that explanations are produced. Global summaries highlight the compact subset of metabolites and taxa that consistently shape predictions across the cohort; because correlated features often share credit, we interpret these as feature groups rather than a strict one-by-one ranking [27]. Local (patient-level) explanations decompose a single prediction into signed contributions, indicating which features push the estimate toward MDD or toward Healthy and by how much [28]-[30].

To translate these attributions into research-only guidance, we apply a templated mapping: 1) for taxa with positive contributions, consider strategies that discourage those niches in controlled contexts (dietary modulation of fermentable substrates, polyphenol-rich foods, or short, supervised courses of antimicrobial herbs); for taxa with negative contributions, consider promotion via prebiotic fibers, fermented foods, or targeted probiotics where evidence exists; 2) for metabolites, map features to pathway families such as kynurenine metabolism, short-chain fatty acids, bile acids, or GABA/serotonin positive contributions suggest down-modulating an overactive pathway, whereas negative contributions suggest restoring or supporting a diminished pathway (e.g., with precursors or cofactors); and 3) for clinical correlates, align contributions to behavioural programs, such as stress-reduction for high stress, sleep hygiene for poor sleep quality, or resistance training and higher-fiber dietary patterns for elevated BMI. These translations are hypothesis prompts, not treatments; they do not imply causality and require clinical oversight, safety review, and prospective testing. For robustness, we recommend sanity checks on explanation dimensionality, grouped SHAP (e.g., by path-

way or taxonomic family) to curb correlation dilution, and bootstrap stability analyses of top features.

3. Results

3.1. Descriptive Statistics and EDA

The feature correlation heatmap shows generally weak-to-moderate pairwise relationships across the selected taxa, metabolites, and clinical variables (Figure 3). This is consistent with our design goal of diffuse, multivariate signal rather than single-feature separation. One notable, high-magnitude block appears between a simulated taxon metabolite pair (e.g., Species_100 and Metabolite_4), reflecting the microbe → metabolite coupling embedded in the generator; most other pairs cluster near zero, indicating low redundancy and reduced risk of spurious shortcuts.

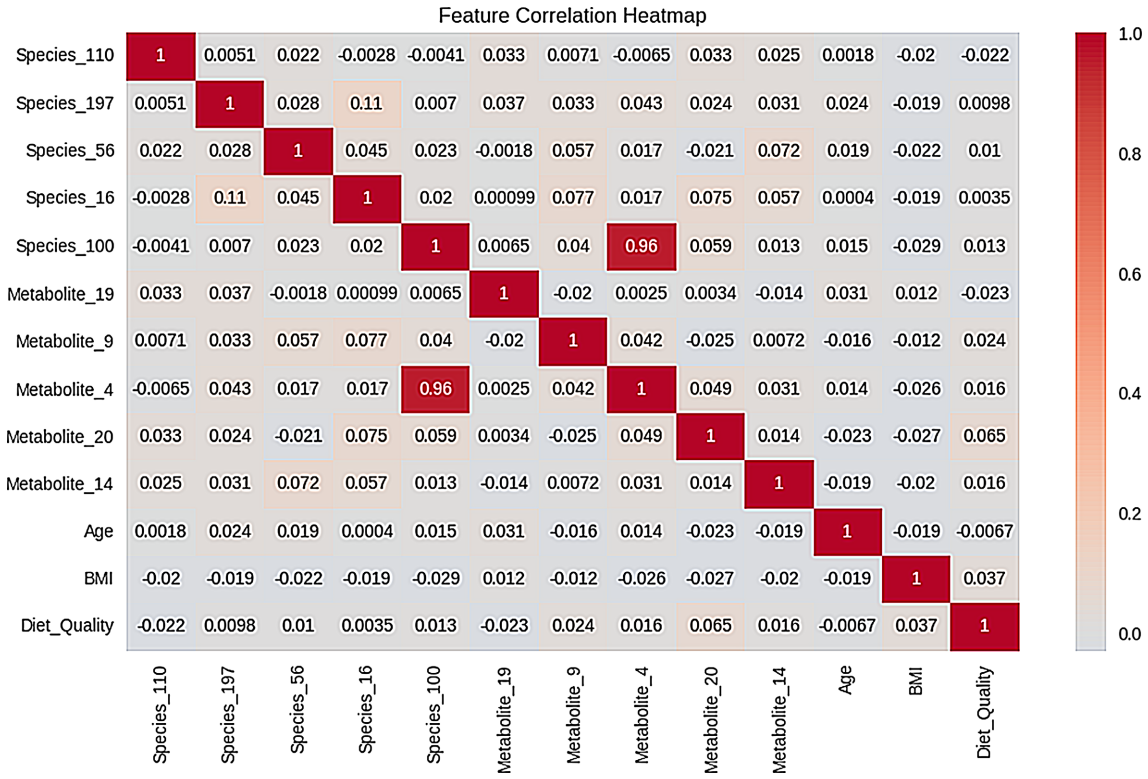


Figure 3. Feature correlation heatmap.

Diet quality vs MDD rate. The prevalence plot across the Diet_Quality scale is non-monotonic (Figure 4), as intended. Several adjacent scores share overlapping error bars, and the overall pattern avoids a simple linear trend. This confirms that "diet alone" does not trivially separate classes in the simulation; rather, diet contributes modestly to combination with omics features, matching the pipeline's multivariate focus.

3.2. Predictive Performance

Overall discrimination. On the stratified 20% test set, the model produced a con-

fusion matrix of [[207, 3], [2, 88]] (Figure 5), yielding Accuracy = 0.98 with only five total errors. Per-class results were strong and balanced:

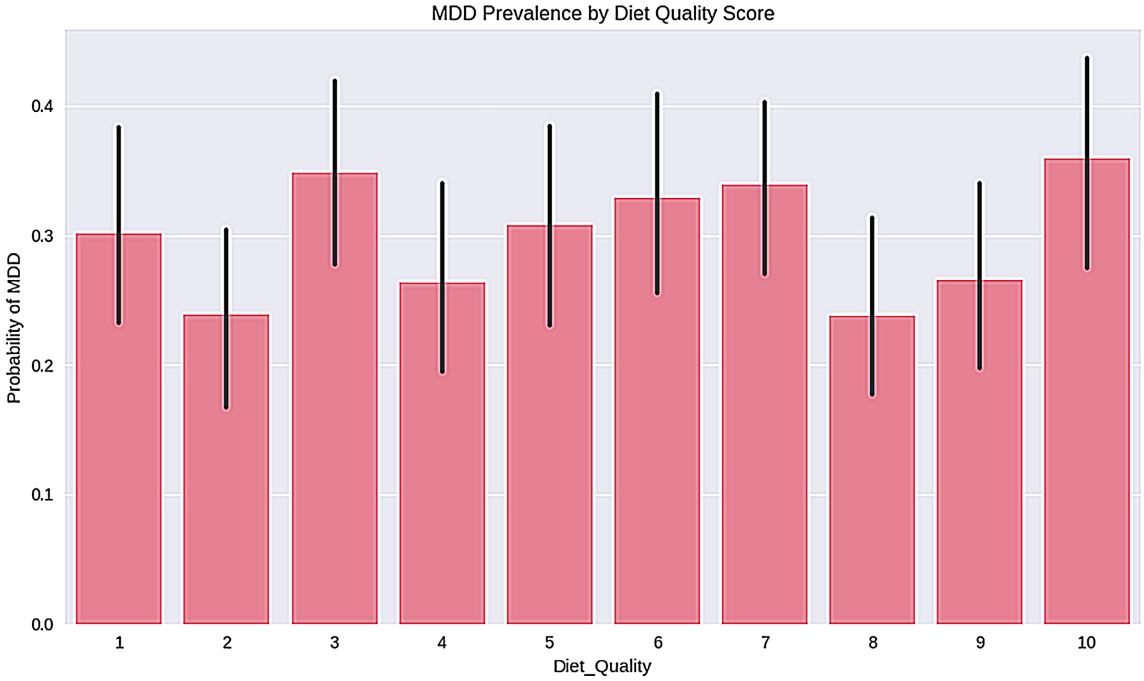


Figure 4. MDD prevalence by diet quality score.

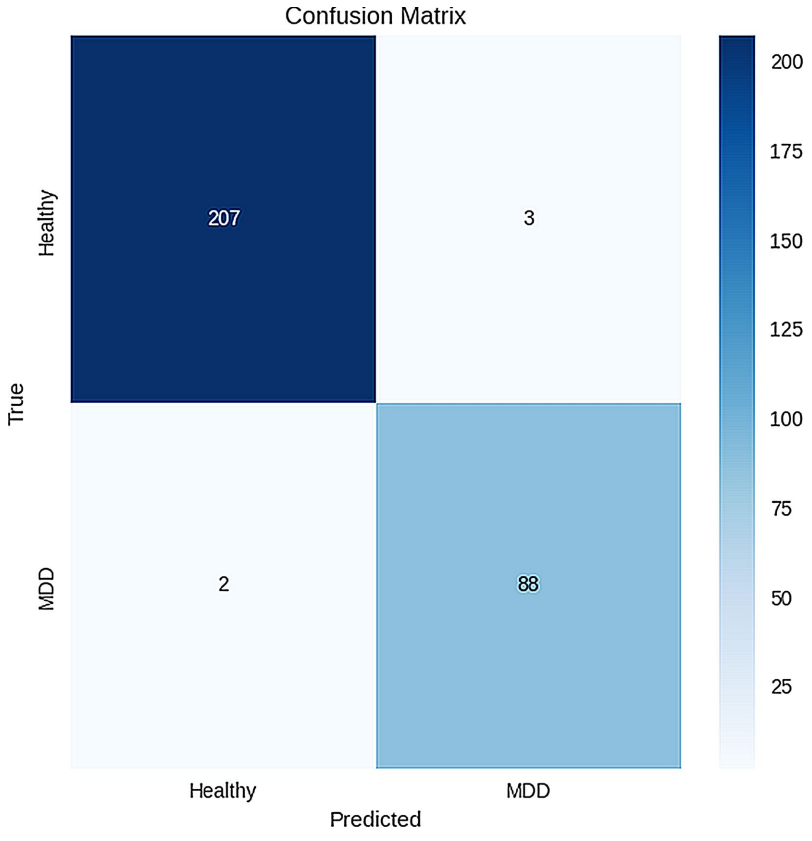


Figure 5. Confusion matrix.

- Healthy (0): Precision 0.99, Recall 0.99, F1 0.99 (support 210);
- MDD (1): Precision 0.97, Recall 0.98, F1 0.97 (support 90);
- Overall: Accuracy 0.98, Macro-F1 0.98, AUC 0.998 (N = 300).

Threshold-free view. The ROC curve closely tracks the top-left boundary with AUC = 0.998 (Figure 6), consistent with coherent signal injected across modalities in the simulator. This indicates excellent separability across decision thresholds.

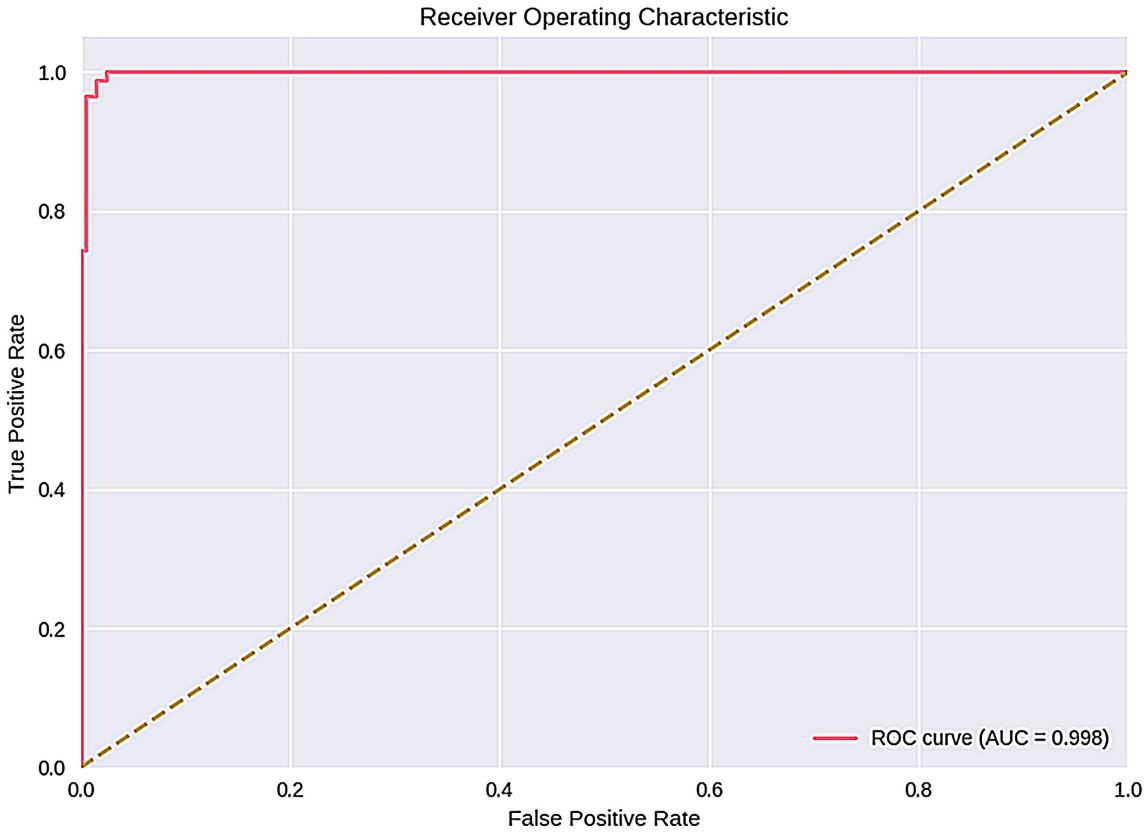


Figure 6. ROC curve.

Learning dynamics (summary). As reported in Methods, validation accuracy stabilized around 0.96 0.97 and validation loss near 0.10, supporting the generalization quality seen on the hold-out.

3.3. Explainability

Model explanations consistently converged on a compact subset of metabolites and taxa that account for most predictive contribution. Because multi-omics variables are often correlated, we treat the highest-ranked features as groups (e.g., pathway clusters or taxonomic families) rather than a strict one-by-one ranking [31]. The SHAP beeswarm in Figure 7, reveals both direction and variability of feature effects at the subject level. Points to the right push predictions toward MDD; points to the left push toward Healthy. The vertical spread shows how consistently a feature matters across individuals: tight bands indicate uniform effects, while wide clouds indicate person-specific behaviour. This view is critical for case

discussion clinicians can see not only which features matter but also how they behave across people.

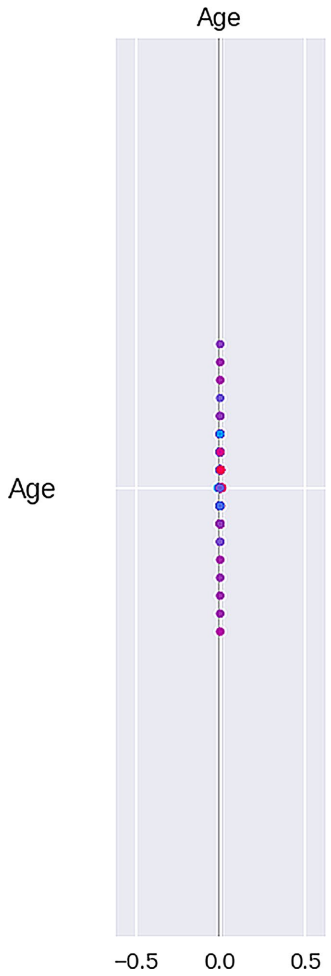


Figure 7. SHAP beeswarm: direction & dispersion.

Here in Figure 7, each point is a subject-level contribution. Right of zero pushes toward MDD; left toward Healthy. Vertical spread reflects between-subject heterogeneity, distinguishing uniformly acting features from those with person-specific effects.

A Random Forest baseline highlighted an overlapping set of drivers (e.g., Metabolite_120, 109, 40, 90, 52, 117, 102 and Species_198, 197, 36, 177, 135, 70), reinforcing that the learned signal is not architecture specific. We interpret impurity-based RF importances as directional evidence and cross-check them against SHAP to mitigate correlation bias.

Here in Figure 8, baseline impurity-based importances. Overlap with SHAP-identified drivers suggests architecture-agnostic signal; interpret ranks as approximate due to correlated features. The dataset is synthetic; feature labels are placeholders and biological readings are illustrative. Explanations demonstrate methodology, not clinical fact.

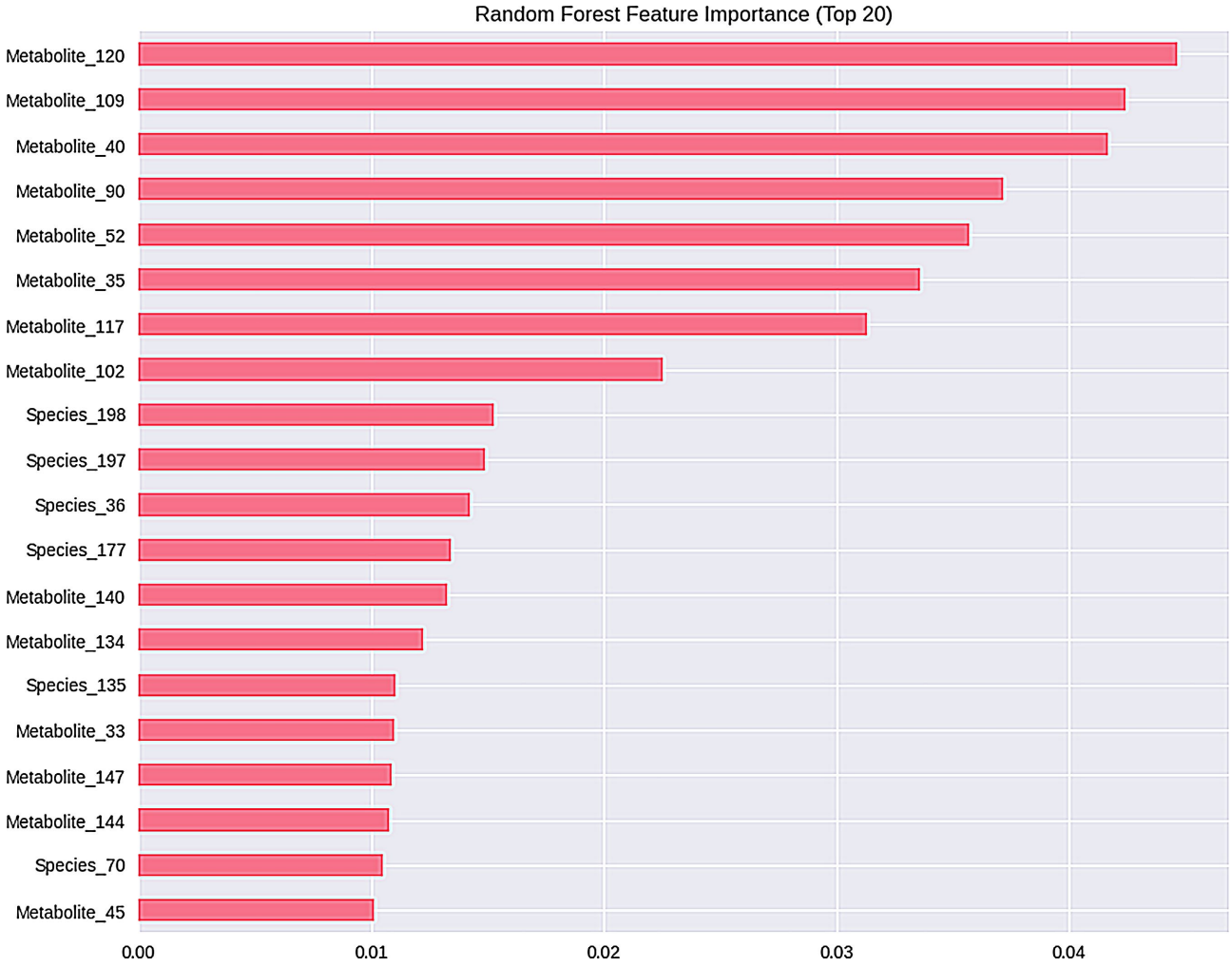


Figure 8. Random forest: top 20 feature importances.

3.4. From Explanations to Precision-Therapeutic Hypotheses (Illustration)

Using SHAP signs as associational (not causal) cues:

- **Nutritional/Microbiome-directed:** If a taxon's SHAP is positive (higher abundance pushes toward MDD), consider research-context strategies that discourage that niche (temporarily lower fermentable substrates; polyphenol-rich foods). If negative, consider promotion via prebiotic fibers (inulin, GOS, resistant starch), fermented foods, or targeted probiotics where evidence exists.
- **Metabolomic/Pathway hypotheses:** Map metabolite features to pathway families-tryptophan kynurenine, short-chain fatty acids, bile acids, GABA/serotonin.
- **Positive SHAP** → hypothesize down-modulating an overactive pathway (e.g., anti-inflammatory strategies; enzyme inhibition such as IDO in research).
- **Negative SHAP** → consider restoring/supporting a diminished pathway (precursors like tryptophan/5-HTP; cofactors B6/Mg/Zn; microbiome strategies that enhance SCFAs).
- **Lifestyle correlates.** Higher Stress_Level and lower Sleep_Quality frequently

push toward MDD in this simulation; candidate hypotheses include mindfulness/HRV-based stress reduction and sleep hygiene. When SHAP for Diet_Quality is protective, it aligns with Mediterranean-style patterns.

These are research prompts, not treatment recommendations. They require clinical oversight, safety review, and prospective testing; SHAP indicates associations within the presented features, not mechanisms or causality.

4. Discussion

4.1. What the Pipeline Shows

This work demonstrates that a compact, end-to-end pipeline can integrate microbiome, metabolomic, and clinical features to recover gut brain axis (GBA) signals in a transparent way. On synthetic data with known structure, the dense network learned interactions across modalities and achieved near-perfect discrimination (AUC ≈ 0.998), indicating the architecture and training recipe are sufficient to capture weak, distributed effects rather than relying on a single dominant marker. Equally important, the pipeline does not stop at prediction. Global and local SHAP analyses expose a small, intelligible set of features that drive decisions, allowing domain experts to interrogate why the model is confident for a given subject and to check whether patterns cohere with biological expectations. The Random Forest baseline offers an orthogonal lens on importance; qualitative agreement between RF importances and SHAP rankings reduces the chance that signal is an artifact of a particular model family. Finally, the templated mapping from signed attributions to dietary, microbiome, and pathway hypotheses illustrates how interpretable machine learning can serve hypothesis generation for precision psychiatry turning "what" (prediction) into "so what" (testable ideas).

4.2. Why the Performance Is so High

The reported accuracy and AUC are intentionally optimistic because the dataset is simulated. We embedded case control shifts in a subset of taxa and metabolites and introduced linear couplings from microbes to metabolites while keeping label noise low and eliminating domain shift. Real-world performance will be lower for several reasons: 1) compositionality of microbiome data complicates effect sizes and induces spurious correlations; 2) batch, site, and platform effects inflate between-study heterogeneity; 3) medication use, diet, BMI, sleep, and stress are intertwined with both microbes and mood, creating confounding that a purely predictive model cannot resolve; and 4) taxonomic definitions and metabolite annotations change over time, introducing drift. These gaps are not shortcomings of the pipeline per se but reflect the difficulty of clinical translation in multi-omics psychiatry.

4.3. Design Choices that Improved Robustness

Three design decisions were pivotal. First, Batch Normalization and Dropout combined with early stopping produced smooth learning curves and prevented

late-epoch degradation [32] [33] (**Figure 2**), which is critical when signals are weak, and features are high-dimensional. Second, class weighting maintained sensitivity to the minority (MDD) class without materially harming specificity, a practical choice for screening contexts. Third, the DeepExplainer → KernelExplainer fallback ensured explanations and figures are always produced across environments no silent failures in the interpretation step. Beyond these, using a strictly held-out test set, fixing random seeds, and aligning pre-processing across models (including the baseline) helped keep estimates stable and reproducible. For deployment on real data, additional safeguards—probability calibration, repeated cross-validation, permutation importance, and bootstrap stability of SHAP rankings—should be layered on.

4.4. Clinical Translation (Guardrails)

The outputs here are research tools, not medical advice. SHAP contributions reflect associations in the features presented to the model; they do not establish causality, nor do they account for contraindications, drug nutrient interactions, or patient preferences. Any nutritional or pharmacological hypothesis derived from model attributions requires prospective testing, safety review, and physician oversight. Supplements (e.g., tryptophan, 5-HTP, high-dose polyphenols), antimicrobials (e.g., berberine, oregano oil), and targeted probiotics can interact with antidepressants, alter drug metabolism, or have paradoxical effects on symptoms. Microbiome interventions also carry ecological risks what helps one profile may harm another [34]-[36]. Accordingly, the appropriate use of this pipeline in practice is study design support: prioritizing which taxa or pathways to measure more deeply, which dietary patterns to test in controlled trials, and which patient subgroups to recruit for mechanistic work. When moving to clinical data, we recommend: preregistered analysis plans; independent training/validation/test cohorts across sites; harmonized wet-lab protocols; explicit handling of compositionality (e.g., CLR transforms or ratio-based features); causal diagrams to identify plausible confounders; decision-curve analysis to quantify clinical utility; subgroup performance audits (sex, age, BMI, medication status); and governance processes for model updates as taxonomies and assays evolve. In short, the pipeline provides a transparent, reproducible scaffold for converting heterogeneous GBA measurements into interpretable predictions and testable precision-medicine hypotheses. Its value lies less in the headline AUC and more in the disciplined path it offers from multi-omics features to clinician-readable reasoning an essential step toward trustworthy, human-centered AI in mental health.

5. Limitations and Future Work

1) Synthetic data: Our results showcase a methodological blueprint, not biological truth. The simulator plants coherent microbe metabolite signals with clean labels and no side effects, which inflates performance. Next steps are to assemble curated, multi-site cohorts with harmonized wet-lab protocols (16S or shotgun

metagenomics with consistent DNA extraction, library prep, sequencing depth), aligned metabolomics platforms (internal standards, retention-time/ion-mode matching), and standardized clinical phenotyping (diagnostic interviews, medication logs, diet and sleep instruments). Pre-register analysis plans and keep one or more sites fully held out for external testing.

2) Compositionality: Microbiome data are relative, so naive correlations can be misleading [37]. Future iterations should use log-ratio transforms (CLR/ALR) or ratio features, and consider models aware of compositional constraints (e.g., Dirichlet-multinomial, logistic-normal frameworks). For metabolomics, apply batch correction and intensity normalization, then re-evaluate models under these transformations to confirm conclusions are not artifacts of scale [38]-[40].

3) Confounding: Antidepressants, diet quality, BMI, sleep, and stress are intertwined with both microbial and symptom profiles. Lay out causal diagrams (DAGs) to define adjustment sets; use propensity weighting or matching, marginal structural models for time-varying exposures, and instrumental-variable or negative-control analyses where appropriate. Collect richer covariates (e.g., antibiotic history, comorbidities) and perform sensitivity analyses to estimate how unmeasured confounding would need to be to overturn findings.

4) Validation: Replace single train/validation splits with nested cross-validation for model selection; report probability calibration (Platt or isotonic), Brier score, and reliability plots. Add decision-curve analysis to quantify clinical net benefit across thresholds. Most importantly, conduct external validation on independent cohorts and, if possible, temporal/site splits to probe generalization.

5) Fairness & distribution shift: Audit performance across sex, age, BMI, medication status, and site. Track calibration per subgroup, not just accuracy. For robustness, use shift detection (e.g., simple density checks or more formal OOD tests), recalibration, and continual learning/periodic refits under governance to prevent drift.

6) Interpretability fidelity: SHAP explains model behaviour but can be unstable with collinearity. Combine it with permutation tests, feature ablations, and bootstrap stability of top-k features. Group features by pathway or taxonomic family before ranking to reduce credit fragmentation. Cross-check with alternative explainers (e.g., Integrated Gradients) and run simulation-based sensitivity to confirm that explanation patterns move in the expected direction when you perturb planted effects.

Additional practical items: Handle missingness systematically (learned imputers, missing-indicator features), document reproducibility (seeds, versions, containers), and address ethics/privacy (consent, de-identification, data-use governance). The end goal is a validated, interpretable tool that supports hypothesis generation and trial design, not automated treatment decisions. For real microbiome data, compositional constraints are non-negotiable; future work will incorporate CLR/ALR transforms or ratio-based features and re-evaluate model stability and explanations under these representations.

6. Conclusion

We presented a fully reproducible, end-to-end pipeline that learns gut brain-axis (GBA) patterns relevant to major depressive disorder from multi-omics and clinical inputs, explains its decisions, and translates those explanations into structured, pathway-level hypotheses for nutrition and pharmacology. Methodologically, the work contributes 1) a clean data-generation and preprocessing scaffold for mixed microbiome metabolome clinical features; 2) a regularized neural classifier benchmarked against a non-neural baseline; and 3) explanation-first outputs (global and patient-level SHAP) coupled to a templated, domain-aware mapping from feature attributions to testable therapeutic ideas. On synthetic data with planted signal, the model achieves high discrimination and stable learning dynamics, demonstrating that multi-modal integration plus transparent reasoning can organize diffuse biological cues into clinician-readable insight. This is not a clinical tool; it is a blueprint for translation. To move from demonstration to deployment, the same pipeline should be validated on curated, multi-site clinical cohorts with harmonized wet-lab protocols, careful handling of compositionality and confounding, probability calibration, fairness and shift audits, and pre-registered analysis plans. Outputs must be governed by ethical review and paired with prospective studies to establish benefit, risk, and generalizability. With those safeguards, this framework can serve as a practical bridge between complex GBA measurements and hypothesis-driven precision psychiatry, accelerating the design of targeted dietary, microbial, and pathway-modulating interventions while keeping clinicians in the loop.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Iyer, G. and Manoj, M. (2024) Neurobiological Perspective on the Gut-Brain Axis and Mental Health: A Review. *International Journal of Interdisciplinary Approaches in Psychology*, **2**, 599-618.

[2] Bertollo, A.G., Santos, C.F., Bagatini, M.D. and Ignácio, Z.M. (2025) Hypothalamus-pituitary-Adrenal and Gut-Brain Axes in Biological Interaction Pathway of the Depression. *Frontiers in Neuroscience*, **19**, Article 1541075. https://doi.org/10.3389/fnins.2025.1541075

[3] Aljeradat, B., Kumar, D., Abdulmuizz, S., Kundu, M., Almealawy, Y.F., Batarseh, D.R., *et al.* (2024) Neuromodulation and the Gut-Brain Axis: Therapeutic Mechanisms and Implications for Gastrointestinal and Neurological Disorders. *Pathophysiology*, **31**, 244-268. https://doi.org/10.3390/pathophysiology31020019

[4] Singh, S.V., Ganguly, R., Jaiswal, K., Yadav, A.K., Kumar, R. and Pandey, A.K. (2023) Molecular Signalling during Cross Talk between Gut Brain Axis Regulation and Progression of Irritable Bowel Syndrome: A Comprehensive Review. *World Journal of Clinical Cases*, **11**, 4458-4476. https://doi.org/10.12998/wjcc.v11.i19.4458

[5] Morys, J., Małecki, A. and Nowacka-Chmielewska, M. (2024) Stress and the Gut-Brain Axis: An Inflammatory Perspective. *Frontiers in Molecular Neuroscience*, **17**,

Article ID: 1415567. https://doi.org/10.3389/fnmol.2024.1415567

[6] Clerici, L., Bottari, D. and Bottari, B. (2025) Gut Microbiome, Diet and Depression: Literature Review of Microbiological, Nutritional and Neuroscientific Aspects. *Current Nutrition Reports*, **14**, Article No. 30. https://doi.org/10.1007/s13668-025-00619-2

[7] Costa, A. and Lucarini, E. (2024) Treating Chronic Stress and Chronic Pain by Manipulating Gut Microbiota with Diet: Can We Kill Two Birds with One Stone? *Nutritional Neuroscience*, **28**, 221-244. https://doi.org/10.1080/1028415x.2024.2365021

[8] Logan, A.C., Cordell, B., Pillai, S.D., Robinson, J.M. and Prescott, S.L. (2025) From Bacillus Criminalis to the Legalome: Will Neuromicrobiology Impact 21st Century Criminal Justice? *Brain Sciences*, **15**, Article 984. https://doi.org/10.3390/brainsci15090984

[9] Motger-Albertí, A. and Fernández-Real, J.M. (2023) Gut Microbiome and Cognitive Functions in Metabolic Diseases. In: Federici, M. and Menghini, R., Eds., *Endocrinology*, Springer International Publishing, 1-27. https://doi.org/10.1007/978-3-031-08115-6_12-1

[10] Barron, D.S., Baker, J.T., Budde, K.S., Bzdok, D., Eickhoff, S.B., Friston, K.J., *et al.* (2021) Decision Models and Technology Can Help Psychiatry Develop Biomarkers. *Frontiers in Psychiatry*, **12**, Article ID: 706655. https://doi.org/10.3389/fpsyt.2021.706655

[11] Datta Burton, S., Mahfoud, T., Aicardi, C. and Rose, N. (2021) Clinical Translation of Computational Brain Models: Understanding the Salience of Trust in Clinician-Researcher Relationships. *Interdisciplinary Science Reviews*, **46**, 138-157. https://doi.org/10.1080/03080188.2020.1840223

[12] McGuire, M.F., Iyengar, M.S. and Mercer, D.W. (2011) Computational Approaches for Translational Clinical Research in Disease Progression. *Journal of Investigative Medicine*, **59**, 893-903. https://doi.org/10.2310/jim.0b013e318224d8cc

[13] Woo, C., Chang, L.J., Lindquist, M.A. and Wager, T.D. (2017) Building Better Biomarkers: Brain Models in Translational Neuroimaging. *Nature Neuroscience*, **20**, 365-377. https://doi.org/10.1038/nn.4478

[14] van Klaveren, D., Varadhan, R., Kent, D.M., *et al.* (2020) The Predictive Approaches to Treatment Effect Heterogeneity (PATH) Statement. *Annals of Internal Medicine*, **172**, Article 776. https://doi.org/10.7326/l20-0427

[15] Mridha, M.F., Keya, A.J., Hamid, M.A., Monowar, M.M. and Rahman, M.S. (2021) A Comprehensive Review on Fake News Detection with Deep Learning. *IEEE Access*, **9**, 156151-156170. https://doi.org/10.1109/access.2021.3129329

[16] Hazra, A., Choudhary, P. and Sheetal Singh, M. (2020) Recent Advances in Deep Learning Techniques and Its Applications: An Overview. In: Rizvanov, A.A., Singh, B.K. and Ganasala, P., Eds., *Lecture Notes in Bioengineering*, Springer, 103-122. https://doi.org/10.1007/978-981-15-6329-4_10

[17] Nunnari, F., Bhuvaneshwara, C., Ezema, A.O. and Sonntag, D. (2020) A Study on the Fusion of Pixels and Patient Metadata in CNN-Based Classification of Skin Lesion Images. In: Holzinger, A., Kieseberg, P., Tjoa, A. and Weippl, E., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 191-208. https://doi.org/10.1007/978-3-030-57321-8_11

[18] Zantvoort, K., Scharfenberger, J., Boß, L., Lehr, D. and Funk, B. (2023) Finding the Best Match—A Case Study on the (Text-)Feature and Model Choice in Digital Mental Health Interventions. *Journal of Healthcare Informatics Research*, **7**, 447-479.

https://doi.org/10.1007/s41666-023-00148-z

[19] Ke, H., Chen, D., Yao, Q., Tang, Y., Wu, J., Monaghan, J., *et al.* (2024) Deep Factor Learning for Accurate Brain Neuroimaging Data Analysis on Discrimination for Structural MRI and Functional MRI. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **21**, 582-595. https://doi.org/10.1109/tcbb.2023.3252577

[20] Simon, F., Weibels, S. and Zimmermann, T. (2025) Deep Parametric Portfolio Policies. No. 23-01. CFR Working Paper.

[21] Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., *et al.* (2018) A Review of Classification Algorithms for EEG-Based Brain-Computer Interfaces: A 10 Year Update. *Journal of Neural Engineering*, **15**, Article 031005. https://doi.org/10.1088/1741-2552/aab2f2

[22] Kottapalle, P., Tak, T.K., Kshirsagar, P.R., Ginnela, G. and Krishna Akula, V. (2025) QHF-CS: Quantum-Enhanced Heart Failure Prediction Using Quantum CNN with Optimized Feature Qubit Selection with Cuckoo Search in Skewed Clinical Data. *Computers, Materials & Continua*, **84**, 3857-3892. https://doi.org/10.32604/cmc.2025.065287

[23] Grabinski, J., Gavrikov, P., Keuper, J. and Keuper, M. (2022) Robust Models Are Less Over-Confident. *Advances in Neural Information Processing Systems*, **35**, 39059-39075.

[24] Sarkar, P.R. (2025) Artificial Intelligence Based Models for Predicting Foodborne Pathogen Risk in Public Health Systems. *International Journal of Business and Economics Insights*, **5**, 205-237. https://doi.org/10.63125/7685ne21

[25] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., *et al.* (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, **2**, 56-67. https://doi.org/10.1038/s42256-019-0138-9

[26] Li, Z. (2022) Extracting Spatial Effects from Machine Learning Model Using Local Interpretation Method: An Example of SHAP and XGBoost. *Computers, Environment and Urban Systems*, **96**, Article 101845. https://doi.org/10.1016/j.compenvurbsys.2022.101845

[27] Khachatryan, L., Xiang, Y., Ivanov, A., Glaab, E., Graham, G., Granata, I., *et al.* (2023) Results and Lessons Learned from the SBV IMPROVER Metagenomics Diagnostics for Inflammatory Bowel Disease Challenge. *Scientific Reports*, **13**, Article No. 6303. https://doi.org/10.1038/s41598-023-33050-0

[28] Vetter, J.S., Schultebraucks, K., Galatzer-Levy, I., Boeker, H., Brühl, A., Seifritz, E., *et al.* (2022) Predicting Non-Response to Multimodal Day Clinic Treatment in Severely Impaired Depressed Patients: A Machine Learning Approach. *Scientific Reports*, **12**, Article No. 5455. https://doi.org/10.1038/s41598-022-09226-5

[29] Shaik, T., Tao, X., Xie, H., Li, L., Higgins, N. and Velásquez, J.D. (2025) Towards Transparent Deep Learning in Medicine: Feature Contribution and Attention Mechanism-Based Explainability. *Human-Centric Intelligent Systems*, **5**, 209-229. https://doi.org/10.1007/s44230-025-00104-7

[30] Landi, I., Glicksberg, B.S., Lee, H., Cherng, S., Landi, G., Danieletto, M., *et al.* (2020) Deep Representation Learning of Electronic Health Records to Unlock Patient Stratification at Scale. *npj Digital Medicine*, **3**, Article No. 96. https://doi.org/10.1038/s41746-020-0301-z

[31] D'haeseleer, P., Liang, S.D. and Somogyi, R. (2000) Genetic Network Inference: From Co-Expression Clustering to Reverse Engineering. *Bioinformatics*, **16**, 707-726. https://doi.org/10.1093/bioinformatics/16.8.707

[32] Roohi, E., Shoja-Sani, A., Goshayeshi, B. and Peyvan, A. (2025) Learning Rarefied Gas Dynamics with Physics-Enforced Neural Networks. arXiv:2509.06231.

[33] Alshayeji, M.H. and Abed, S. (2025) Heart Disease Prediction by Tabular Modeling with Deep Learning Network and Interpretability. *Machine Learning*: *Science and Technology*, **6**, Article 035043. https://doi.org/10.1088/2632-2153/adfd39

[34] Peixoto, R.S., Voolstra, C.R., Sweet, M., Duarte, C.M., Carvalho, S., Villela, H., *et al.* (2022) Harnessing the Microbiome to Prevent Global Biodiversity Loss. *Nature Microbiology*, **7**, 1726-1735. https://doi.org/10.1038/s41564-022-01173-1

[35] Lemon, K.P., Armitage, G.C., Relman, D.A. and Fischbach, M.A. (2012) Microbiota-targeted Therapies: An Ecological Perspective. *Science Translational Medicine*, **4**, 137rv5. https://doi.org/10.1126/scitranslmed.3004183

[36] Lange, L., Berg, G., Cernava, T., Champomier-Vergès, M., Charles, T., Cocolin, L., *et al.* (2022) Microbiome Ethics, Guiding Principles for Microbiome Research, Use and Knowledge Management. *Environmental Microbiome*, **17**, Article No. 50. https://doi.org/10.1186/s40793-022-00444-y

[37] Gihawi, A., Ge, Y., Lu, J., Puiu, D., Xu, A., Cooper, C.S., *et al.* (2023) Major Data Analysis Errors Invalidate Cancer Microbiome Findings. *mBio*, **14**, e01607-23. https://doi.org/10.1128/mbio.01607-23

[38] Hagenbeek, F.A., Roetman, P.J., Pool, R., Kluft, C., Harms, A.C., van Dongen, J., *et al.* (2020) Urinary Amine and Organic Acid Metabolites Evaluated as Markers for Childhood Aggression: The ACTION Biomarker Study. *Frontiers in Psychiatry*, **11**, Article ID: 165. https://doi.org/10.3389/fpsyt.2020.00165

[39] Dmitrenko, A., Reid, M. and Zamboni, N. (2022) A System Suitability Testing Platform for Untargeted, High-Resolution Mass Spectrometry. *Frontiers in Molecular Biosciences*, **9**, Article ID: 1026184. https://doi.org/10.3389/fmolb.2022.1026184

[40] Baker, E.S. and Patti, G.J. (2019) Perspectives on Data Analysis in Metabolomics: Points of Agreement and Disagreement from the 2018 ASMS Fall Workshop. *Journal of the American Society for Mass Spectrometry*, **30**, 2031-2036. https://doi.org/10.1007/s13361-019-02295-3