IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Machine Learning-Driven Classification of Type 2 Diabetes Using Gut Microbiome Profiles for Personalized Therapeutics

Tara Pratapa, Ayan Dharod

Indus International School, Kondakal Village, Shankarpalle, Hyderabad, Telangana, 501203, India; tarapratapa7@gmail.com

ABSTRACT: The purpose of the study is to investigate Type 2 Diabetes Mellitus (T2DM) progression and risk through investigating gut microbiome biomarkers, providing insights into disease status and potential synbiotic and fecal microbiota transplantation interventions. Gut dysbiosis serves as a key indicator for T2DM detection, supporting effective intervention strategies by outlining potential targets for up/down-regulation. Our preliminary research aimed to compile gut microbiome biomarkers documented to be associated with disease states, forming the foundation for constructing a model capable of disease state analysis. Gut microbiome composition data of individual patients from sequenced metagenomic files were compiled, encompassing the abundance of over 4,500 microbial taxa across profiles. Multiple models were tested to evaluate the greatest efficacy given the nature of the data used, from which XGBoost was selected and optimized. Despite 95.8% sparsity in our microbiome data alongside high local intrinsic dimensionality (LID≈12), a 0.86 AUC score and 0.80 testing accuracy were achieved. SHAP values enabled the analysis of key bacterial features, alongside a correlational network for biological validation through the visualization of co-occurring T2DM-associated bacterial taxa. This supported the hypothesis that machine learning can provide critical insights into the nature of dysbiosis, precisely classifying disease states and modeling biologically accurate relationships. Our findings demonstrate the biological accuracy of the model, along with its ability to identify under-researched species significantly correlated with disease states. This suggested their potential as stable and reliable biomarkers. While further data is required to enhance model performance, this work represents a crucial step toward revolutionizing targeted diagnosis, patient care, and intervention strategies.

KEYWORDS: Medical and Health Sciences, Epidemiology, Machine Learning, Gut Microbiome, Type 2 Diabetes.

Introduction

Type 2 Diabetes Mellitus (T2DM) is a metabolic disorder characterized by the excess production of glucose, caused by the inadequate secretion of insulin by Beta-cells, resulting in hyperglycemic states.[1] Data from the 2022 NCD Risk Factor Collaboration reports that approximately 828 million individuals live with diabetes globally, with over 95% having type 2 diabetes mellitus. Type 2 diabetes not only affects older populations but has seen a two- to threefold increase among individuals below the age of 40.[2] Even more concerning is the fact that almost one in two adults with diabetes is undiagnosed, remaining unaware that they suffer from the disease.[3] This reflects a concerning gap in access to timely diagnosis, thereby enabling early intervention and treatment for the disease, despite its worldwide prevalence.

With millions affected and at risk of severe complications, patients would greatly benefit from the application of precision medicine that accounts for the inherently multifactorial nature of this disease.[4] For example, clinical trials reveal higher remission rates for T2DM patients taking personalized dietary interventions compared to generalized nutrition advice, highlighting the effectiveness of customized strategies to combat T2DM.[5] The employment of other such personalized therapeutic avenues, including fecal microbiota transplantation (FMT) and synbiotics, displays great efficacy in improving patient health outcomes over time, influencing microbial diversity.[6]

Gut dysbiosis is a notable marker for the T2DM disease state. The human gut contains a diverse range of bacterial taxa, and research has demonstrated its key role in metabolic homeostasis. Disruptions in this balanced composition can drive the progression of various diseases, including metabolic syndrome disorders like T2DM.[7] The availability of large datasets enables machine learning algorithms to identify key patterns in microbiome compositions of patient samples in the same group. This makes computational methods well-suited for processing the heterogeneity of gut microbiome data through the employment of techniques like feature identification, selection, and post-intervention strategy optimization.[8]

The scope of the application of machine learning algorithms to handle diverse gut microbiome datasets led to the investigation of their efficacy in analyzing disease states and suggesting intervention strategies based on this analysis. This project examines the gut microbiome as a source of biomarkers for detecting T2DM, and leverages computational approaches to aid its early detection and intervention. The main objective is improving methods for assessing a person's risk of diabetes through biomarkers in the gut, further enabling personalized interventions for the given patient(s).

DOI: 10.36838/IJHSR814.36

Methods

Data Collection:

Publicly available human gut microbiome datasets containing both healthy and T2DM cohorts were identified through a comprehensive literature search. Studies were included only if they provided relative abundance data from stool-derived bacterial species based on sequencing analyses. In order to ensure proper standardization of data, clearly defined inclusion and exclusion criteria were constructed, as outlined in Table 1.

Table 1: Exclusion and inclusion criteria were used to filter gut microbiome data samples from studies, including subjects based on age (1-90 years), BMI (15-45), and sex, and excluding based on chronic diseases, medications, smoking, pregnancy, and treatments.

Criteria	*Type*	*Requirement*
Age	Inclusion	Sample is from a subject between 1-90 years; samples younger than 1 year or older than 90 years were excluded
BMI	Inclusion	Sample is from a subject 15-45 BMI value; samples outside this range were excluded
Sex	Inclusion	Male and female samples were included
Other chronic disease	Exclusion	Presence of any other chronic disease apart from diabetes resulted in exclusion of the given subject's sample
Medication	Exclusion	Samples from subjects on any type of medication linked to diabetes were excluded
Smoking	Exclusion	Samples from subjects who were smokers were excluded
Pregnancy	Exclusion	Any pregnant subjects were automatically excluded
Treatment for infectious disease / with antibiotics for past three months	Exclusion	Samples who had recently contracted an infectious disease and were treated for it, including those on antibiotics for the past three months, were excluded
Treatment with steroids or immunomodulatory drugs	Exclusion	Samples with any type of treatment involving steroids or immunomodulatory drugs were excluded
Cognitive dysfunction or diagnosed mental health conditions	Exclusion	Samples with a diagnosed mental health condition and/or any form of cognitive dysfunction were excluded

The inclusion criteria listed in Table 1 reflect the demographic diversity of the samples. Additionally, datasets were chosen to enable geographic diversity, including representation across Europe, North America, and Asia. A final set of six datasets (PRJEB1786, PRJNA289586, PRJNA299502, PRJNA389481, PRJNA422434, PRJNA448494) was used for training the machine learning model. Given the anonymity of samples and the absence of any personally identifiable information, risks of data leakage for misuse were eliminated.

Two primary sequencing methodologies were used across the selected datasets: whole-genome shotgun (WGS) metagenomics and 16S rRNA amplicon sequencing. Post-DNA extraction,[11] libraries were prepared and barcoded for multiplexed sequencing on Illumina platforms.[12,13] Host reads were removed using Bowtie2,[14] and species abundances were derived via MetaPhlAn,[15,16] producing final sample-level abundance tables. 16S rRNA amplicon sequencing targets the bacterial 16S gene to infer community composition at the species level.[17,18] In such workflows, the 16S rRNA gene was amplified by PCR,[19] followed by barcoded Illumina sequencing.[20] Reads were quality-filtered using DADA2 and Deblur,[21] demultiplexed,[22] and taxonomically assigned using the SILVA reference database to generate amplicon sequence variants (ASVs) and abundance matrices.[23]

Sample reads were compiled, with columns representing bacterial species, while rows represented samples. A given abundance value could be understood based on the column it belonged to (which bacterial species the abundance was of) and the row it was present in (which species had this abundance level). Finally, binary encoding was done for each sample based on the class it belongs to: '0' (healthy) and '1' (diabetic).

Separation was made between true zeroes and unmeasured values; true zeroes were retained while unmeasured values were removed. Samples with >10% of bacterial species not measured were excluded. Additionally, those bacterial species present in <5% of samples were eliminated, reducing noise and overfitting. Missing abundance values were imputed by taking the median abundance of each bacterial species within the given group. Imputed values were flagged for later identification, and a final filtration was done of any sample or feature with over 20% imputed entries. This elimination ensured minimal fabricated data, preserving biological validity and natural variation.

Post-data cleaning, the final pre-processing steps were performed to eliminate any remaining discrepancies. Feature harmonization helped resolve duplicate taxa created by inconsistent naming across studies. Once complete, a sample-by-feature matrix was generated, with the remaining metadata stored separately. A train-test split was done in the 80:20 ratio, respectively. Class balance was ensured across splits to prevent model bias towards any majority class.

Constructing the Machine Learning Model:

Random forest (RF) model was selected due to the data's high dimensionality and sparsity, with features greatly exceeding the number of patient samples. Such a model can mitigate the overfitting commonly seen in single decision trees trained on the given biological data's nature. As an ensemble method, it reduces variance by averaging predictions from multiple, diverse trees and thereby reducing model variance.[24]

The implemented random forest model yielded an accuracy of 0.82. However, there was significant overfitting and low biological validity, as found through discrepancies between top SHAP-derived feature importances and expected biological relevance. The features used to train the RF model include: number of samples in which a species was detected, mean relative abundance, median abundance, and standard deviation of abundance. The data's high sparsity, with many zero and near-zero values, combined with high dimensionality, worsened model overfitting. Too many predictors relative to samples exacerbated this. Derived statistical measures that the random forest could not fully mitigate. While RF constructs each tree on a random subset of features and samples, each tree can individually overfit noisy, high-dimensional input that was present in the dataset, and generalization is impaired.

After evaluating the RF model, simpler regression-based approaches were tested to compare performance and interpretability. Recognizing that training on derived statistics (mean, median, and standard deviation) increased input noise and overfitting, subsequent models were built using solely abundance values to improve signal-to-noise ratio and biological interpretability. XGBoost, LASSO, Elastic Net, and Ensemble regression models were evaluated. Each regression model was selected to enhance generalizability by learning from features

across patients while managing sparse data through selective feature weighting. This balanced bias and variance control for improved stability and interpretability of features. LASSO and Elastic Net applied built-in regularization constraints to minimize overfitting in such high-dimensional data. The ensemble regression model was implemented to combine multiple weak learners, enhancing predictive robustness while also reducing variance. Performance results of the models based on standard ML metrics are shown in Figure 1.

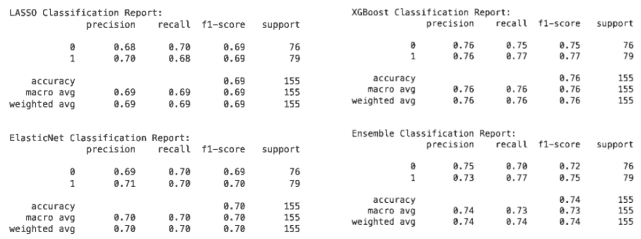

Figure 1: Classification performance across regression-based models shows that tree-based and ensemble approaches outperform linear models, with all achieving balanced precision, recall, and F1-scores.

XGBoost, a gradient-boosted ensemble, was specifically selected because of its ability to sequentially correct prior errors and natively handle sparse data, addressing both bias and variance. The model's sequential tree-building process improved generalization over RF, improving the biological plausibility of selected features. XGBoost achieved the best performance compared to the other models with a testing accuracy of 0.7613. Its iterative error correction and regularized approach sequentially corrects prior mistakes, reduces model complexity, and accommodates sparsity with built-in mechanisms.

Only 458 columns contained >10% non-zero values, prompting tests of higher density thresholds, where datasets were reconstructed using limits to filter features with >30% and >50% non-zero values. However, model accuracies decreased when trained on these filtered datasets. This approach inadvertently excluded biologically important yet sparse features that showed significant variation between diabetic and healthy groups. The omission of these features also degraded SHAP-reported feature importance profiles, ultimately reducing biological accuracy.

Age and BMI values for each patient were collected, but had not been incorporated into the model initially. These variables were subsequently included, as they were hypothesized to be indicative of metabolic status and potentially improve generalization by providing physiologically relevant context. However, with a significant amount of BMI data available only as categorical ranges rather than precise values, its inclusion led to a slight reduction in model accuracy, suggesting limited predictive benefit under the current data conditions. Therefore, the final model only included age as a demographic feature from the sample metadata.

Table 2: Precision and recall analysis post-dataset alterations: original data (precision 0.76, recall 0.75); after columns with >30% non-zero values removed (0.73, 0.70); >50% non-zero values columns removed (0.75, 0.71); with age and BMI covariates (0.73, 0.76).

Dataset	Precision (0, 1)	Recall (0, 1)
Original	0.76, 0.76	0.75, 0.77
30 percent (Columns)	0.73, 0.72	0.70, 0.75
50 percent (Columns)	0.75, 0.73	0.71, 0.77
With age and BMI	0.73, 0.71	0.68, 0.76

When calculating SHAP values, there remained a lack of biological accuracy and scope for improving generalization through refined feature selection. To address this, two complementary approaches were implemented: Gradient Boosting Machine (GBM) and SelectKBest using f_classif for feature selection. Since the model continued to demonstrate overfitting, these methods were applied to ensure that the selected features enhanced generalizability. Applying both techniques helped mitigate the effects of high dimensionality by ensuring that trees utilized consistently differentiating or statistically significant features for prediction, thereby improving model performance.

GBM provided a measure of feature importance, quantifying the contribution of each feature to model accuracy during training. SelectKBest ranked features based on the ANOVA F-test for classification, selecting the top K features with the highest scores. SelectKBest was applied once on the training subset obtained from an 80:20 train–test split. Feature selection was not performed for each fold. The feature selector was fitted exclusively on the training data, and the resulting feature transformation was then applied to the test set.

Following implementation and SHAP-based feature evaluation, there was a notable increase in biological accuracy. More of the identified species corresponded to documented gut microbiome biomarkers associated with diabetes, accompanied by a measurable improvement in predictive accuracy. Among the top 15 SHAP values, 7 overlapping features were observed between the GBM and SelectKBest results, reinforcing the robustness of the final feature set.

The validation methodology employed in this study utilized GridSearchCV with the model during hyperparameter tuning, supporting the identification of optimal parameter combinations.[25] Specifically, a k-fold cross-validation method was conducted on each combination.[26] The parameter set with the best average score was chosen as the optimal model hyperparameters, ensuring performance was generalised across train-test data splits.

For GBM, changes, including adjusting the number of estimators (129) and maximum depth (4), and refining ultra-sparse feature thresholds, increased model accuracy from 0.73 to 0.78. For SelectKBest, changes, including tuning the number of selected features (K=1480) and updating the sparsity threshold, improved accuracy from 0.76 to 0.80. These parameter refinements ensured more stable feature selection and better representation of biologically relevant features.

Interventions:

Probiotics contain live bacteria with beneficial effects, while prebiotics are nutrients that promote the proliferation of these

DOI: 10.36838/IJHSR814.36

beneficial microorganisms. Together, they synergistically aim to enhance the growth of targeted species like Bifidobacterium and Lactobacillus, improving gut health. In combination with the addition of substrates such as lactose, which support bacterial growth, they form a synbiotic. This combination results in a greater overall effect, specifically designed to support the newly introduced beneficial bacteria and improve gut microbiota balance.[27]

The analysis of patient-derived microbiome samples identified key microbial taxa contributing to gut dysbiosis and potential progression toward disease states. Quantitative assessment of compositional profiles provided insight into taxa whose relative abundances deviated significantly from those characteristic of a healthy microbiome. These deviations were used to inform strategies for the administration of synbiotics driven by computationally-derived hypotheses for targeted modulation strategies.

Taxa exhibiting abundances exceeding or falling below established healthy thresholds were prioritized for individualized intervention. The magnitude of deviation from these thresholds informed the formulation of patient-specific synbiotic compositions, comprising agents that either promote or suppress the growth of particular bacterial taxa. This precision-guided framework generates candidate intervention hypotheses, requiring experimental validation before therapeutic application.

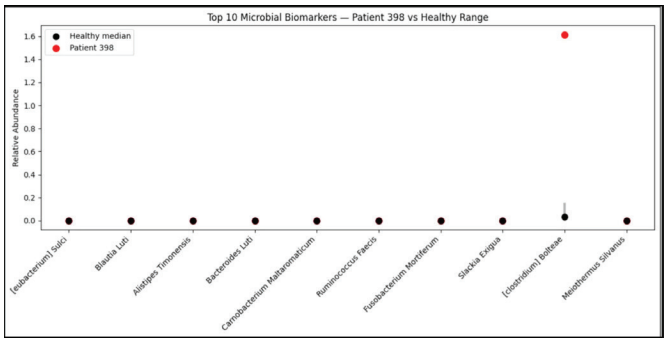

Figure 2: Comparative abundance of the top 10 microbial biomarkers reveals pronounced deviations in specific taxa between a diabetic patient and the healthy reference range, underscoring candidate species for targeted microbiome modulation.

Individual patient analyses for specific microbiome features are illustrated in the figures. Key indicators of dysbiosis can be identified by examining bacterial taxa within patient samples that significantly contribute to the disease state. In the depiction above, diabetic patients exhibit species-level abundances that differ from the median values observed in healthy controls. These deviations can be mapped to bacterial species within a repository containing evidence-based synbiotics known to upregulate or downregulate specific taxa. In the first patient, targeted intervention may be beneficial due to the observation of elevated *Clostridium bolteae* levels. The synbiotic combination *Clostridium butyricum* with chitooligosaccharides (COS) has been tested and shown to inhibit pathogenic Clostridium species, improving intestinal barrier integrity and reducing inflammation.[28] This demonstrates how individualized microbiome analysis can be integrated with a synbiotic intervention repository to enable microbiome-targeted intervention. This represents a hypothesis warranting in vitro or clinical validation, not a confirmed therapeutic recommendation.

Fecal microbiota transplantation (FMT) refers to the transfer of a healthy donor's stool (believed to have a balanced gut microbiome) to a diseased recipient (with an unbalanced microbiome composition).[29] The aim of this intervention is to restore the microbial balance of up- or down-regulated gut elements in the diseased recipient, reversing dysbiosis.[30] Healthy samples from the initial data sheet were taken as potential donors, with 50 diabetic samples randomly selected as recipients of FMT. During preprocessing, only bacterial species present in at least 10% of samples were retained to remove extremely rare features.

Across the healthy samples, key measures of central tendency were calculated, including mean, median, standard deviation, and the 10th and 90th percentiles. These became the benchmarks of each bacterial species as it would exist in a healthy, balanced gut microbiome, enabling comparison of excess and deficit species in the diabetic patient. For each Type 2 Diabetic patient (recipient), three key metrics were calculated: deviation from the mean, Z-score, and percentile position. Given the high variance in diabetic sample species abundance compared across the healthy donors, values of '0' and '100' were given if the values were lower than the minimum or exceeded the maximum, respectively, handling outliers during percentile ranking.

Donor-recipient mapping began by identifying taxa in each diabetic patient requiring modulation, using Z-score and percentile ranks to classify over- or under-represented species. A conditional framework combined these criteria to label each taxon as high, low, or normal in abundance, prioritizing significant deviations as biologically relevant. Figure 3 highlights the step-by-step process through which a verdict was defined on whether modulation was needed for a given bacterial species.

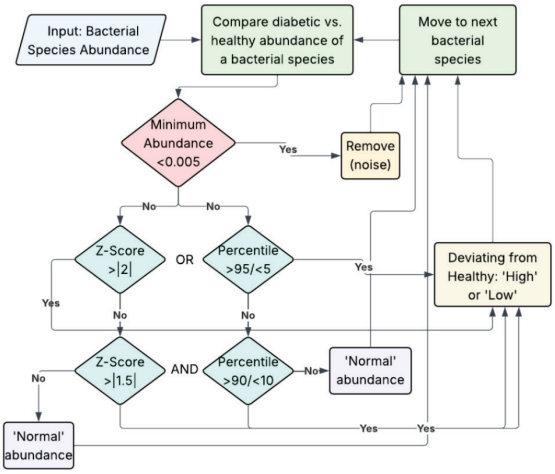

Figure 3: A flowchart illustrates the step-by-step evaluation process for each bacterial species' abundance in gut microbiome samples, determining the need for regulation based on Z-score, percentiles, and deviation from healthy norms.

A donor-species matrix was crafted by calculating each taxon's abundance across all healthy samples. percentile ranks for each bacterial species across potential donors. This helped identify the placement of different donors for a given bacterial species presence in the context of the remaining samples,

which would aid in ranking candidates for the given diabetic recipient. For each recipient, only dysregulated bacterial species were considered. Percentile values were converted into directional percentiles: higher percentiles favored donors capable of upregulating deficient taxa, while inverted percentiles favored donors suited to downregulate excessive ones.

These directional percentiles were weighted by the absolute value of the Z-score for each taxon, emphasizing those species that were deviating more significantly. A weighted average was calculated to produce donor-specific computationally predicted compatibility scores, where increased scores indicated better donor-recipient matches. Accordingly, the weighted averages were ranked in descending order. The donor ID was retrieved for the corresponding score, enabling a final list of the top donors for a given recipient.

Results and Discussion

Model Optimization and Performance:

The final testing accuracy achieved for the model was 80.00%. Alongside accuracy, other metrics were also measured to evaluate the model's performance, as shown in Table 3.

Table 3: Model metrics evaluating performance of XGBoost classifier: healthy class (precision 0.80, recall 0.79, F1 0.79, support 76); diabetic class (0.80, 0.81, 0.81, support 95); averages (0.80).

Class	Precision	Recall	F1-Score	Support
Healthy ('0')	0.80	0.79	0.79	76
Diabetes ('1')	0.80	0.81	0.81	79
Average	0.80	0.80	0.80	155 (*sum*)

A confusion matrix of the XGBoost model predicted 64 true positives, 60 true negatives, 16 false positives, and 15 false negatives. Figure 4 depicts the ROC/AUC curve plotted for the XGBoost model. The XGBoost model achieved an AUC value of 0.856, indicating strong discriminative ability with high true-positive rates and low false-positive rates.

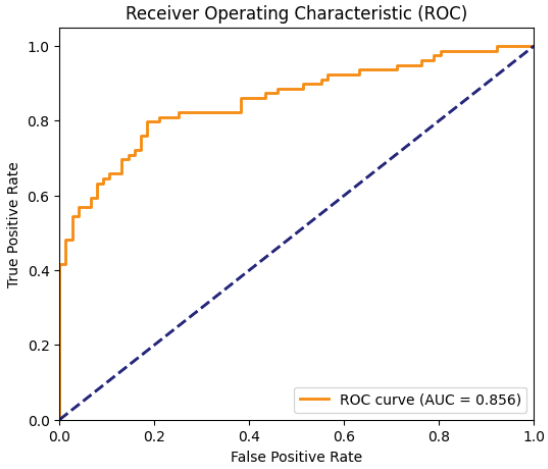

Figure 4: ROC curve for the XGBoost model predicting type 2 diabetes from gut microbiome data (AUC=0.86), demonstrating strong discriminative ability.

A precision-recall curve demonstrated a strong model performance with an average precision of 89.66%. The strong ranking ability demonstrated across a wide range of recall values highlights the balanced performance of classification. The consistency of the model is further highlighted by the absence of sharp drops, which would otherwise indicate overfitting or instability.

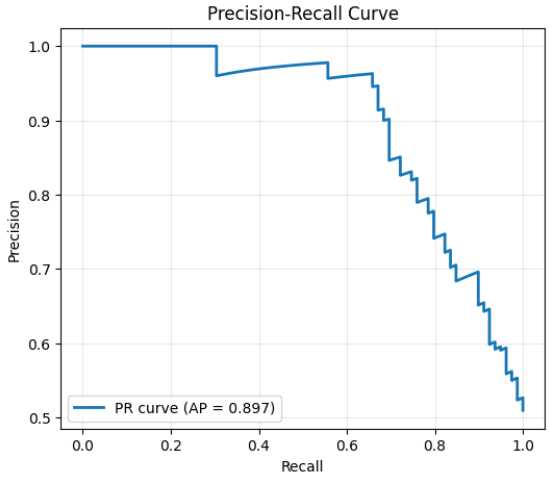

Figure 5: Precision-recall curve of the XGBoost model, with an 89.66% average precision.

Building on the inferences made on the importance of feature selection, a chart was plotted to deduce the relationship between an increasing number of features and the resulting model accuracy. The chart was built post-filtering the ultra-sparse features, with <2000 features of the original ≈4500 features for the SelectKBest method. Iterations were conducted at a step value of 20, measuring the model accuracy with an increasing number of features. Figure 10 depicts this relationship. An evident finding from the chart is the absence of monotonic improvement of the model correlating with the number of features included. As can be seen, the highest accuracy value was achieved not with the complete set of non-ultra-sparse features, but in fact from a lower number of selected features (≈1440). The curve can be interpreted as showing how fewer features, in such a case, can help declutter data for enhanced model performance. Selecting only the most informative features supports generalization and the denoising of data, since the inclusion of more redundant bacteria increases dimensionality while not adding signal. Therefore, alongside simply ultra-sparse feature elimination, purposeful feature selection using computationally-backed methods is critical when working with data of the nature of microbiome samples.

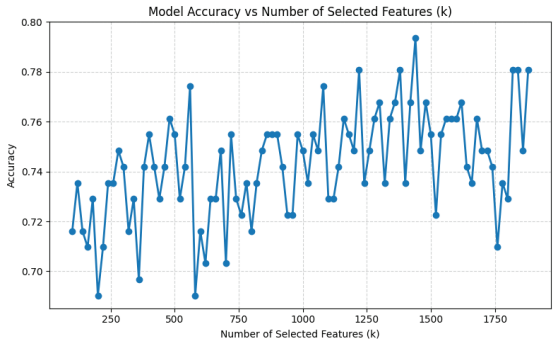

Figure 6: Model accuracy of XGBoost classifier across varying numbers of selected features (k=100 to 1750), showing peak performance at ≈1440 features.

DOI: 10.36838/IJHSR814.36

Model Explainability and Feature Robustness:

SHAP values describe the magnitude of contribution of a given feature, alongside the directional effect of a given feature's values on the predicted risk of both an individual sample and the overall model. A global SHAP feature importance plot helps rank different features (bacterial species) by their mean SHAP contribution, enabling comparison of each feature's overall influence irrespective of direction. This supports the identification of the most influential microbiome species, helping validate feature selection and interpretability. Figure 6 details such a plot for the XGBoost model. The presence of age as the feature with the largest contribution is consistent with the strong age-linked microbiome patterns expected in T2DM. Other gut microbial species shown have prior associations with key biological processes involved in diabetes, including metabolism, inflammation, and insulin resistance, highlighting the biological plausibility of the signals identified by SHAP.

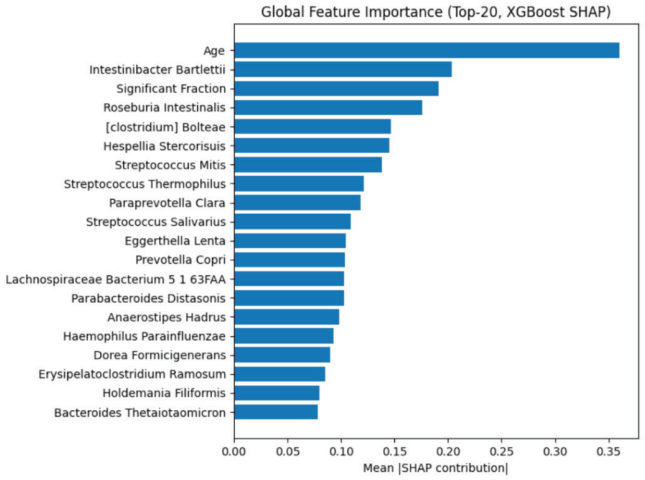

Figure 7: Global feature importance plot based on mean SHAP values from the XGBoost model, ranking top bacterial taxa contributing most to type 2 diabetes classification.

The stability of bacterial features selected using the SelectKBest method was validated in Figure 7, highlighting feature selection performed across 100 bootstraps. The original dataset samples were taken with replacement to create new datasets of equal size but varying compositions. All displayed taxa were selected in ≈0.63-0.71 bootstraps, highlighting the robustness and reproducibility of the bacterial features selected by the model. This demonstrates the biological validity of the features of these features by SelectKBest. Further, the absence of any single biomarker dominating with a near 100% stability highlights the multi-taxa considerations taken for the classification of any sample, justifying the biological grounding of the model. Similarly, the absence of very low-frequency features (<0.4) supports model stability and the effectiveness of feature filtering in eliminating non-significant features that would otherwise highly fluctuate between bootstraps.

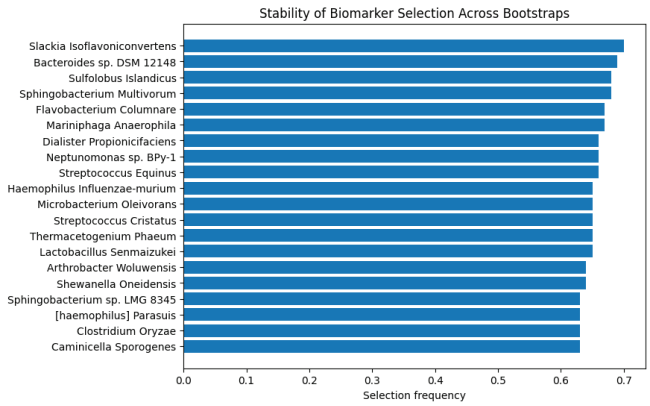

Figure 8: Biomarkers stable at 0.63-0.71 bootstrap frequency confirm robust, reproducible feature selection.

Necessity of Multivariate Modeling over Univariate Analyses:

To illustrate the inter-feature relationships within the microbiome dataset, a correlational network map was created, as shown in Figure 8. Each microbial taxon is represented by a node, with the edges between the nodes depicting the correlations seen. Thicker edges represent stronger correlations, and groups of microbes seen in clusters are termed modules, acting as functional groups that co-occur. The presence of multiple sub-communities indicates microbial co-participation in pathways linked to diabetes status. The high interconnectivity reflects the co-variance of taxa, reinforcing the interdependent structure of the microbiome. Therefore, the presence of a structural and functional network demands the use of multivariate analysis and machine learning to capture network-level dependencies as compared to single-taxon statistical testing.

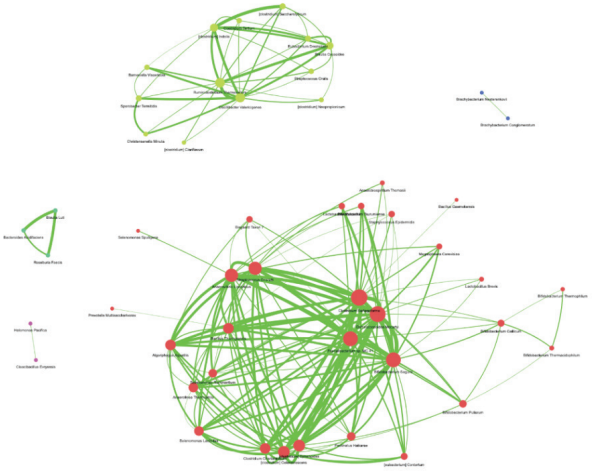

Figure 9: Network reveals interconnected microbial guilds co-varying with diabetes, validating multivariate ML over univariate analysis.

The above two graphs exemplify the need for machine learning to be employed in microbiome data for classification applications, as compared to traditional statistical approaches to such tasks.

In order to biologically validate the basis on which the research was crafted, namely the presence of varying abundances in gut microbiome species between diabetic (red) and healthy (blue) classes, violin plots were constructed as shown in Figure

9. Given the distinct abundance values between the two groups, a clear validation of gut biomarkers for diabetes is present, with either more or fewer enriched species in samples with the disease. Bidirectional shifts in abundance further strengthen the biological validity of the dataset, supporting a multi-feature approach over single-taxon analysis. Further, compared to the healthy violins, the diabetic violins are wider, indicating greater intra-group variance that highlights microbiome instability and dysbiosis, a hallmark of metabolic disorders like T2DM. Finally, significant overlaps in bacterial taxa (*Prevotella copri, Hespellia stercorisuis, Clostridium bolteae*, among others) between the global SHAP values and top SelectKBest features represented in the violin plots (Figure 9) demonstrate the validity of the most important features chosen by the two feature selection methods.

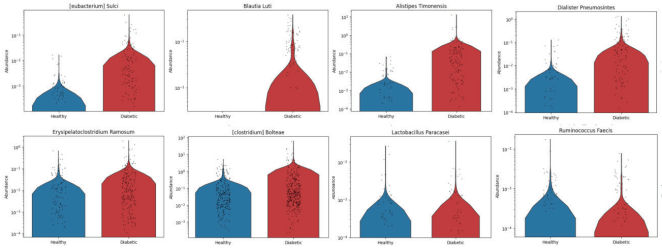

Figure 10: Violin plots of top features from SelectKBest analysis in gut microbiome data for type 2 diabetes prediction, comparing distributions (blue: healthy; red: diabetic) for key taxa. Violin plots confirm distribution shifts in top taxa between groups.

The features represented in the violin plots and the SHAP analyses were compared with those identified gut biomarkers from the literature review. While most were bacterial species that were known, well-researched diabetic signatures, there were certain taxa that had not been documented as biomarkers of T2DM. Further research was conducted on such species, examining potential pathways and intermediary biological components that would correlate the given species with known T2DM processes. For instance, *Lactobacillus paracasei* is known as a probiotic in therapeutic contexts, though it has not been associated with diabetes. However, certain studies prove that in animal models, supplementation of the bacteria improved hyperglycemia, reducing inflammatory markers and also supporting an improved gut-barrier integrity.[31,32]

Alistipes timonesis is another taxon that, while correlated with gut dysbiosis overall, has not been classified as a T2DM marker. Its production of unique metabolites (such as sulfonolipids) ties it with diet-linked gut changes,[33] given that diet is a strong risk factor for diabetes, and high-fat/high-animal-based diets were associated with increased levels of Alistipes,[34] it can be hypothesized that the given species is a potential diabetes biomarker. The identified up-regulation in the diseased state, when further justified by the evident increase of the taxa's abundance in the diabetic samples shown in the violin plots, indicates the plausibility of this gut microbe as a possible diabetic indicator. Another species in the genus Eubacterium contains multiple species that produce butyrate.[35] Given butyrate's known role as a metabolite linked to gut integrity and insulin sensitivity,[36] thereby indicating that the bacteria's presence in the butyrate-producer network could correlate with T2DM risk.

Synbiotic-Mediated Pathway Regulation:

Individual patient compositions led to insights on potential synbiotic formulations that could be administered to upregulate or downregulate bacterial species. Key biomarkers deviating from healthy thresholds were specifically targeted using synbiotics, such as *Lactobacillus rhamnosus* GG and d-tagatose, to upregulate Lactobacillus,[37] as shown in Figure 10. Species were ranked by importance and degree of deviation from the healthy median. For each key feature identified, corresponding synbiotic candidates were recommended. Patient-specific taxa show significant individual variation, thereby supporting personalized synbiotic design.

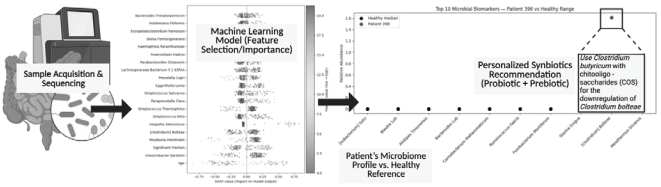

Figure 11: Profiling workflow maps dysregulated taxa to personalized synbiotics, bridging biomarkers to clinical modulation strategies.

Expanding the current repository of experimentally validated synbiotics (Table 4) could enhance intervention design by broadening the range of targetable biomarkers. Further experimental validation of untested synbiotics is essential to ensure clinical reliability. Nonetheless, synbiotics remain a promising route for modulating influential taxa, enabling personalized and targeted treatment strategies.

Table 4: Excerpt of synbiotic intervention data used in post-patient analysis, including modulating influent taxa, synbiotic category, targeted bacterial strains, and clinical strain IDs.

Microbe Name	Substance Name	Substance-Category	Substance-subCategory	Microbe Change	Condition/Disease	Reference_ID
Bifidobacterium longum	Inulin + B. longum	Synbiotic	Prebiotic + Probiotic	Increase	Type 2 Diabetes	PMID: 31456712
Lactobacillus rhamnosus	GOS + L. rhamnosus	Synbiotic	Prebiotic + Probiotic	Increase	Obesity/T2D	PMID: 30011245
Faecalibacterium prausnitzii	FOS + B. adolescentis	Synbiotic	Prebiotic + Probiotic	Increase	Metabolic Syndrome	PMID: 28945121
Akkermansia muciniphila	Polyphenol mix + L. plantarum	Synbiotic	Plant Prebiotic + Probiotic	Increase	Insulin Resistance	PMID: 29123456
Roseburia intestinalis	Resistant starch + L. casei	Synbiotic	Prebiotic + Probiotic	Increase	Gut Dysbiosis	PMID: 32098712

FMT Donor Predictive Scoring:

According to the methodology detailed in Section 2.4, donor-recipient rankings were generated for all the diabetic patients, with the top 20 results summarized in Table 5. Here, FMT is modelled computationally as a hypothesis-generating framework; the donor rankings produced are candidate matches for experimental testing, not clinical prescriptions. The FMT rankings and match scores were calculated for 50 different recipients (diabetic patients), across whom donor match scores exhibited a high range. Top-ranking donors fell between the ≈93.9-94.9 band as shown in the bar plot, highlighting the strong computationally predicted compatibility of a set of donors for certain recipients. The distribution of donor scores highlighted the influence of variations in the magnitude of a certain microbiome community's presence on their suitability to the recipient. While generalized across the 50 recipients, such results highlight the existence of multiple possible donor matches for a given diabetic patient undergoing FMT. Further, variability among donor ranks between different recipients highlights the importance of personalized, recipient-specific

DOI: 10.36838/IJHSR814.36

donor-matching as compared to a universal preference for any given donor.

Table 5: Top donors score 93.9-94.9 for select recipients, showing score drop-off and recipient-specific matching superiority.

Rank	Donor ID	Match Score	Rank	Donor ID	Match Score
1	Donor 210	94.946	11	Donor 251	94.108
2	Donor 293	94.939	12	Donor 25	94.088
3	Donor 306	94.719	13	Donor 295	94.061
4	Donor 41	94.514	14	Donor 12	94.038
5	Donor 291	94.406	15	Donor 253	94.014
6	Donor 323	94.371	16	Donor 11	94.001
7	Donor 150	94.316	17	Donor 278	93.970
8	Donor 245	94.186	18	Donor 222	93.943
9	Donor 129	94.165	19	Donor 307	93.907
10	Donor 30	94.143	20	Donor 185	93.905

By compiling the results of 10 of the 50 recipients for visualization purposes, a heatmap was constructed, depicted in Figure 11. Of greatest significance is the finding that donor-recipient computationally predicted compatibility patterns are heterogeneous, shown in the variations in efficacy of a given donor for different recipients. The presence of certain donors possessing high matching scores across multiple recipients does hint at a potential 'broad spectrum donor' status; however, personalized donor selection is, in almost every case, more effective than the universal selection of any one given donor across recipients. The variation in range and distribution of donor computationally predicted compatibility scores across recipients indicates that some diabetic patients could be easier to match with healthy donors, while others require a larger pool for finding effective matches. Finally, clustering within the heatmap indicates groups of donors sharing microbiome features that confer similar match profiles.

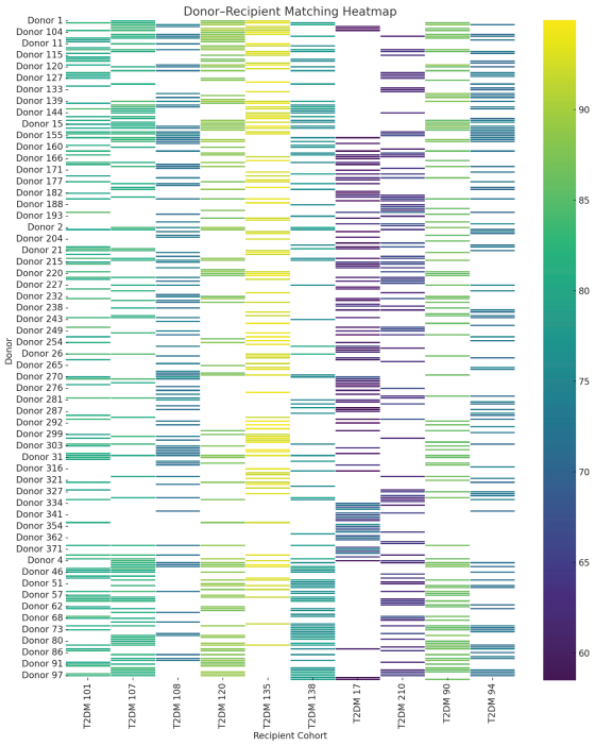

Figure 12: Heatmap illustrates donor-recipient match scores for FMT charts, highlighting trends within recipient sets; wider intervals indicate high variability, and boxplots show donor distribution compared to T2DM patient profiles.

In order to map trends within and across recipients, boxplots were constructed (Figure 12). These charts further extrapolated the idea of inter-recipient variability, wherein certain diabetic patients are easily matched by the set of donor samples taken, while others, possessing lower match scores, are harder to match. However, the presence of higher outliers even for those recipients with lower average scores (such as 'T2DM 17', 'T2DM 210', 'T2DM 108') indicates that even such patients have highly compatible donors, despite being isolated. In comparison, certain recipients (including 'T2DM 135', 'T2DM 120', 'T2DM 90') demonstrate high score ranges with low variance, implying their greater matchability for the given healthy sample group chosen. Recipients with lower inter-quartile ranges suggest more robust donor predictions, while wider intervals indicate high donor-specific performance fluctuations. The construction of such a box plot exemplifies the importance of the continuous nature of donor-recipient match data, as compared to a binary classification that lacks the sophistication of a computed score range.

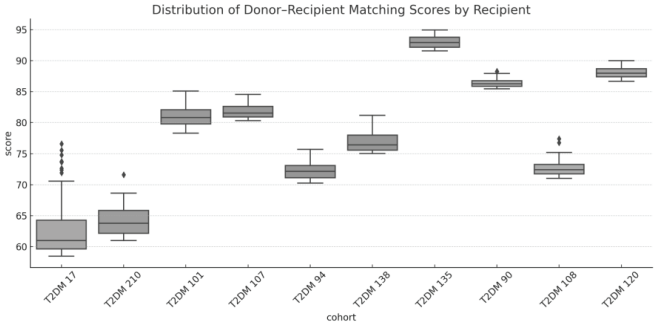

Figure 13: Boxplots reveal broad variability in matchability across recipients, favoring personalized over universal donor selection.

Conclusion

The study employed a machine learning model to classify patients between T2DM and healthy classes based on their gut microbiome readings, aiding improved disease diagnosis and targeted intervention. Such an approach transcends traditional statistical analyses that often use univariate comparisons between groups for predictions, instead employing a multivariate approach better aligned with interdependencies between biological features. When seen in the context of the nature of microbiome data (with a 95.8% sparsity in the training set employed for this study), the achieved 80.00% test accuracy highlights the immense potency of such computational techniques in enabling early detection and intervention. Microbiome-driven disease diagnostics can complement standard glucose and metabolite tests, especially due to their ability to offer a scalable, non-invasive diagnostic that complements precision medicine and personalized therapeutic interventions.

An outcome of the study was the high robustness of ML models in maintaining biological integrity and remaining reliable to the pathophysiological nature of a given disease. The construction of SHAP analyses, correlational network maps, and violin plots confirms that the top predictive taxa used in feature selection methods are highly representative of documented species. The employment of feature analysis further

supported the identification of novel or under-reported taxa that had a highly plausible connection to T2DM processes through various pathways and metabolic intermediaries. Such results highlight the study's informed findings of potential biomarkers that require further research to expand the existing literature on the topic.

Given that only ≈4% of the data used contained non-zero values, the ability for the model to generalize findings towards an 80% testing accuracy, precision, and recall, demonstrates the strong robustness of the created model to noise and class imbalance. In the context of existing studies, large-scale T2D analyses employed XGBoost models, such as the XGBoost-based model in Gut Microbiota and Type 2 Diabetes[38] and the China-specific obesity-related study[39]achieved AUCs of 0.84 and 0.70, respectively, with Bacteroides emerging as a top-ranking feature across both. In contrast, this study achieves comparable or higher AUC (0.85) while maintaining performance across diverse demographic groups. The greater success achieved by our model can largely be attributed to biologically-informed feature selection (GBM and SelectKBest) and regularized designs to minimize overfitting.

The integration of two interventions, FMT and synbiotics, further emphasises the translational value of the model in supporting therapeutic modeling. The pipeline's ability to provide personalized microbiome restoration strategies by pinpointing up/down-regulation highlights the widespread potential of such computational methods. Using precision synbiotic design and donor-recipient matching algorithms, the system developed acts as a closed-loop diagnostic-intervention cycle enabling prediction, validation, and personalisation.

The largest constraint in deploying such models is the dearth of open-source disease-specific microbiome data that ensures standardization of data collection and demographic (metadata) diversity and availability. Future research studies that produce data must integrate host metadata consistently, prevent intra-study sample variation for key criteria (such as the presence of other chronic diseases or illnesses), and expand sample size. Such improvements can enable the use of more sophisticated models to improve inter-study classification validity and prediction accuracy.

Future directions for this research include: a) assimilating datasets with larger sample sizes for improved feature-to-sample ratio, enhancing performance; b) testing model robustness through mimic disease incorporation (obesity, T1DM, other related diseases); c) integrated intermediary stages (prediabetic) between healthy and diabetic for early intervention; d) expanding personalized therapeutic strategies, such as testing efficacy of diet interventions; e) collecting human gut samples from patients belonging to the two cohorts, to conduct a double-blind study on disease status prediction.

By demonstrating the ability of machine learning to classify diabetic patients with high accuracy even in ultra-sparse microbiome environments, staying rooted in biologically-grounded networks and co-occurrences, this work marks a pivotal step toward data-driven, personalized diabetic diagnostics and interventions.

References

1. ElSayed, Nuha A., Rozalina G. McCoy, Grazia Aleppo, Henry Cruz, *et al.* "2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2025." *Diabetes Care*, vol. 48, no. Supplement_1, Dec. 2024, pp. S27–49. https://doi.org/10.2337/dc25-s002.
2. Młynarska, Ewelina, Witold Czarnik, Natasza Dzieża, Weronika Jędraszak, *et al.* "Type 2 Diabetes Mellitus: New Pathogenetic Mechanisms, Treatment and the Most Important Complications." *International Journal of Molecular Sciences*, vol. 26, no. 3, Jan. 2025, p. 1094. https://doi.org/10.3390/ijms26031094.
3. Ogurtsova, Katherine, Leonor Guariguata, Noël C. Barengo, Paz Lopez-Doriga Ruiz, *et al.* "IDF Diabetes Atlas: Global Estimates of Undiagnosed Diabetes in Adults for 2021." *Diabetes Research and Clinical Practice*, vol. 183, Dec. 2021, p. 109118. https://doi.org/10.1016/j.diabres.2021.109118.
4. Tobias, Deirdre K., Jordi Merino, Abrar Ahmad, Catherine Aiken, *et al.* "Second International Consensus Report on Gaps and Opportunities for the Clinical Translation of Precision Diabetes Medicine." *Nature Medicine*, vol. 29, no. 10, Oct. 2023, pp. 2438–57. https://doi.org/10.1038/s41591-023-02502-5.
5. Arias-Marroquín, Ana T., Fabiola M. Del Razo-Olvera, Zaira M. Castañeda, Jesús Salvador Olivares, *et al.* "Personalized Versus Non-personalized Nutritional Recommendations/Interventions for Type 2 Diabetes Mellitus Remission: A Narrative Review." *Diabetes Therapy*, vol. 15, no. 4, Feb. 2024, pp. 749–61. https://doi.org/10.1007/s13300-024-01545-2.
6. Su, Lili, Yanyun Zhang, Wei Chen, Zhiqiang Li, *et al.* "Health Improvements of Type 2 Diabetic Patients Through Diet and Diet Plus Fecal Microbiota Transplantation." *Scientific Reports*, vol. 12, no. 1, Jan. 2022, p. 1152. https://doi.org/10.1038/s41598-022-05127-9.
7. Yu, Yonghua, Yilan Ding, Shuangyuan Wang, Lei Jiang, *et al.* "Gut Microbiota Dysbiosis and Its Impact on Type 2 Diabetes: From Pathogenesis to Therapeutic Strategies." *Metabolites*, vol. 15, no. 6, June 2025, p. 397. https://doi.org/10.3390/metabo15060397.
8. Abavisani, Mohammad, N. Ebadpour, A. Khoshrou, Amirhossein Sahebkar, *et al.* "Deciphering the Gut Microbiome: The Revolution of Artificial Intelligence in Microbiota Analysis and Intervention." *Current Research in Biotechnology*, vol. 7, Jan. 2024, p. 100211. https://doi.org/10.1016/j.crbiot.2024.100211.
9. Sharpton, Thomas J. "An Introduction to the Analysis of Shotgun Metagenomic Data." *Frontiers in Plant Science*, vol. 5, June 2014, p. 209. https://doi.org/10.3389/fpls.2014.00209.
10. Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, Nicola Segata, *et al.* "Shotgun Metagenomics, From Sampling to Analysis." *Nature Biotechnology*, vol. 35, no. 9, Sept. 2017, pp. 833–44. https://doi.org/10.1038/nbt.3935.
11. Costea, Paul I., Georg Zeller, Shinichi Sunagawa, Eric Pelletier, *et al.* "Towards Standards for Human Fecal Sample Processing in Metagenomic Studies." *Nature Biotechnology*, vol. 35, no. 11, Oct. 2017, pp. 1069–76. https://doi.org/10.1038/nbt.3960.[pubmed.ncbi.nlm.nih]
12. Rohland, Nadin, David Reich. "Cost-effective, High-throughput DNA Sequencing Libraries for Multiplexed Target Capture." *Genome Research*, vol. 22, no. 5, Jan. 2012, pp. 939–46. https://doi.org/10.1101/gr.128124.111.[pubmed.ncbi.nlm.nih]
13. Jeon, Sol A., Jihwan Ha, Chang-Bum Hong, Junghyun Park, *et al.* "Comparison of the MGISEQ-2000 and Illumina HiSeq 4000 Sequencing Platforms for RNA Sequencing." *Genomics & Informatics*, vol. 17, no. 3, Sept. 2019, p. e32. https://doi.org/10.5808/gi.2019.17.3.e32.

DOI: 10.36838/IJHSR814.36

14. Langmead, Ben, Steven L. Salzberg. "Fast Gapped-read Alignment With Bowtie 2." *Nature Methods*, vol. 9, no. 4, Mar. 2012, pp. 357–59. https://doi.org/10.1038/nmeth.1923.
15. Segata, Nicola, Levi Waldron, Alessio Ferrer, Danilo Tomaš, *et al.* "Metagenomic Microbial Community Profiling Using Unique Clade-specific Marker Genes." *Nature Methods*, vol. 9, no. 8, June 2012, pp. 811–14. https://doi.org/10.1038/nmeth.2066.
16. Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, *et al.* "MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling." *Nature Methods*, vol. 12, no. 10, Sept. 2015, pp. 902–03. https://doi.org/10.1038/nmeth.3589.
17. Janda, J. Michael, Sharon L. Abbott. "16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls." *Journal of Clinical Microbiology*, vol. 45, no. 9, July 2007, pp. 2761–64. https://doi.org/10.1128/jcm.01228-07.
18. Johnson, Jethro S., Nicholas J. Spakowicz, Cara A. Hong, Andrew J. Petersen, *et al.* "Evaluation of 16S rRNA Gene Sequencing for Species and Strain-level Microbiome Analysis." *Nature Communications*, vol. 10, no. 1, Nov. 2019, p. 5029. https://doi.org/10.1038/s41467-019-13036-1.
19. Tremblay, Julien, Krista Yergeau, Alonso R. Avci, Jean-Pierre Fortin, *et al.* "Primer and Platform Effects on 16S rRNA Tag Sequencing." *Frontiers in Microbiology*, vol. 6, Aug. 2015, p. 771. https://doi.org/10.3389/fmicb.2015.00771.
20. Kozich, James J., Sarah L. Westcott, Niel T. Baxter, Sarah K. Highlander, *et al.* "Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform." *Applied and Environmental Microbiology*, vol. 79, no. 17, June 2013, pp. 5112–20. https://doi.org/10.1128/aem.01043-13.
21. Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, *et al.* "DADA2: High-resolution Sample Inference From Illumina Amplicon Data." *Nature Methods*, vol. 13, no. 7, May 2016, pp. 581–83. https://doi.org/10.1038/nmeth.3869.[pubmed.ncbi.nlm.nih]
22. May, Ali, Sanne Abeln, Wim Crielaard, Jaap Heringa, *et al.* "Unraveling the Outcome of 16S rDNA-based Taxonomy Analysis Through Mock Data and Simulations." *Bioinformatics*, vol. 30, no. 11, Feb. 2014, pp. 1530–38. https://doi.org/10.1093/bioinformatics/btu085.[pubmed.ncbi.nlm.nih]
23. Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, *et al.* "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-based Tools." *Nucleic Acids Research*, Jan. 2013. https://doi.org/10.1093/nar/gks1219.[academic.oup]
24. Akselrud, Caitlin I. Allen. "Random Forest Regression Models in Ecology: Accounting for Messy Biological Data and Producing Predictions With Uncertainty." *Fisheries Research*, vol. 280, Sept. 2024, p. 107161. https://doi.org/10.1016/j.fishres.2024.107161.
25. Jiménez, Álvaro Barbero, José M. Sánchez, Francisco J. Martínez-Estudillo, Alfonso C. Martínez-Estudillo, *et al.* "Finding Optimal Model Parameters by Discrete Grid Search." *Advances in Intelligent and Soft Computing*, 2007, pp. 120–27. https://doi.org/10.1007/978-3-540-74972-1_17.
26. (Proceedings Paper). "Reliable Accuracy Estimates From K-Fold Cross Validation." *IEEE (conference paper; details via IEEE Xplore abstract)*. http://ieeexplore.ieee.org/abstract/document/8698831.
27. Kearney, Sean M., Sean M. Gibbons. "Designing Synbiotics for Improved Human Health." *Microbial Biotechnology*, vol. 11, no. 1, Dec. 2017, pp. 141–44. https://doi.org/10.1111/1751-7915.12885.
28. Huang, Xingxi, Zhiqiang Li, Yuxin Wang, Mengmeng Zhao, *et al.* "Clostridium Butyricum and Chitooligosaccharides in Synbiotic Combination Ameliorate Symptoms in a DSS-Induced Ulcerative Colitis Mouse Model by Modulating Gut Microbiota and Enhancing Intestinal Barrier Function." *Microbiology Spectrum*, vol. 11, no. 2, Mar. 2023, p. e0437022. https://doi.org/10.1128/spectrum.04370-22.
29. Kim, Kyeong Ok, Michael Gluck. "Fecal Microbiota Transplantation: An Update on Clinical Practice." *Clinical Endoscopy*, vol. 52, no. 2, Mar. 2019, pp. 137–43. https://doi.org/10.5946/ce.2019.009.
30. Gupta, Shaan, Ramesh Aggarwal, Rajesh K. Tandon, Yashi K. Mohan, *et al.* "Fecal Microbiota Transplantation: In Perspective." *Therapeutic Advances in Gastroenterology*, vol. 9, no. 2, Oct. 2015, pp. 229–39. https://doi.org/10.1177/1756283x15607414.
31. Toejing, Parichart, Supaporn Phunpee, Waraporn Srisuwannaket, Narissara Phaosri, *et al.* "Influence of Lactobacillus Paracasei HII01 Supplementation on Glycemia and Inflammatory Biomarkers in Type 2 Diabetes: A Randomized Clinical Trial." *Foods*, vol. 10, no. 7, June 2021, p. 1455. https://doi.org/10.3390/foods10071455.
32. Zeng, Zhu, Jing Zhang, Qian Liu, Xiaoqian Wang, *et al.* "Ameliorative Effects of Lactobacillus Paracasei L14 on Oxidative Stress and Gut Microbiota in Type 2 Diabetes Mellitus Rats." *Antioxidants*, vol. 12, no. 8, July 2023, p. 1515. https://doi.org/10.3390/antiox12081515.
33. Parker, Bianca J., Chris J. Wearstler, Emily A. Karlsson, Richard J. S. H. Li, *et al.* "The Genus Alistipes: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health." *Frontiers in Immunology*, vol. 11, June 2020, p. 906. https://doi.org/10.3389/fimmu.2020.00906.
34. Walker, Alesia, Michael R. Lewis, Alessia F. Valdes, Tim D. Spector, *et al.* "Sulfonolipids as Novel Metabolite Markers of Alistipes and Odoribacter Affected by High-fat Diets." *Scientific Reports*, vol. 7, no. 1, Sept. 2017, p. 11047. https://doi.org/10.1038/s41598-017-10369-z.
35. Singh, Vineet, Falguni Patel, Rakesh K. Singh, Analava Mitra, *et al.* "Butyrate Producers, 'The Sentinel of Gut': Their Intestinal Significance With and Beyond Butyrate, and Prospective Use as Microbial Therapeutics." *Frontiers in Microbiology*, vol. 13, Jan. 2023, p. 1103836. https://doi.org/10.3389/fmicb.2022.1103836.
36. Güler, Meryem Saban, Ayşe Nur Tekin, Mehmet Fatih Aslan, Özlem Şimşek, *et al.* "Butyrate: A Potential Mediator of Obesity and Microbiome via Different Mechanisms of Actions." *Food Research International*, vol. 199, Nov. 2024, p. 115420. https://doi.org/10.1016/j.foodres.2024.115420.
37. Lv, Yuanqiang, Jian Chen, Jingjing Li, Qianqian Zhang, *et al.* "Synbiotics Effects of D-tagatose and Lactobacillus Rhamnosus GG on the Inflammation and Oxidative Stress Reaction of Gallus Gallus Based on the Genus of Cecal Bacteria and Their Metabolites." *PLoS ONE*, vol. 20, no. 1, Jan. 2025, p. e0317825. https://doi.org/10.1371/journal.pone.0317825.
38. Jin, Xinqi, *et al.* "Gut Microbiota and Type 2 Diabetes: Genetic Associations, Biological Mechanisms, Drug Repurposing, and Diagnostic Modeling." *International Journal of Molecular Sciences*, vol. 27, no. 2, 21 Jan. 2026, p. 1070, pmc.ncbi.nlm.nih.gov/articles/PMC12842411/, https://doi.org/10.3390/ijms27021070.
39. Ge, Xiaochun, *et al.* "Application of Machine Learning Tools: Potential and Useful Approach for the Prediction of Type 2 Diabetes Mellitus Based on the Gut Microbiome Profile." *Experimental and Therapeutic Medicine*, vol. 23, no. 4, 23 Feb. 2022, https://doi.org/10.3892/etm.2022.11234.

Authors

Tara Pratapa is currently a junior at Indus International School, wanting to pursue computational biology, specifically for genomics and epidemiology.

Ayan Dharod is also a junior, driven to work in the space of precision and personalized medicine for disease prevention, diagnosis, and treatment.

DOI: 10.36838/IJHSR814.36