The Human Microbiome Analysis and Its Effects Associated with Type 2 Diabetes Mellitus

Vaishnavi B. Illindala[1] **& Satyavani Karra**[2]

Received September 4, 2025
Accepted December 9, 2025
Electronic access January 31, 2026

The gut microbiome is integral to metabolic regulation, particularly in modulating glucose homeostasis. In type 2 diabetes, this community becomes dysbiotic-less diverse and imbalanced. This study confirms a strong link between dysbiosis and diabetes, showing decreased levels of the beneficial *Roseburia inulinivorans* and increased levels of *Clostridium hathewayi* and *Clostridium bolteae*. These shifts are associated with metabolic homeostasis and correlate with insulin resistance. Systemic inflammation, particularly marked by Tumor Necrosis Factor-alpha (TNF-alpha), further aggravates the condition. Statistical analyses of metagenomic data confirmed significant differences in bacterial abundance based on diabetes status. Predictive modeling identified age, microbiome composition, and gender as critical factors. An optimized artificial neural network (ANN) model demonstrated good potential for predicting disease outcomes, achieving enhanced accuracy with minimal overfitting. These findings reveal a multifactorial relationship among age, gut microbiota, and inflammation that contributes to the advancement of diabetes. This highlights the promise of microbiome-targeted therapeutic strategies. Furthermore, with continued refinement, machine learning models like ANN could become powerful tools for predicting diabetes and personalizing treatment plans. These findings should be interpreted as exploratory, given dataset constraints and modest model performance, and highlight the need for validation in larger, more detailed cohorts.

Keywords: MetaPhlAn; Clostridium hathewayi; Clostridium bolteae; Roseburia inulinivorans; Tumor necrosis factor-alpha (TNFα); Multivariate analysis of variance (MANOVA); Artificial neural network (ANN)

Introduction

As of 2025, approximately 589 million adults aged 20 to 79 are living with diabetes globally, with type 2 diabetes comprising more than 90% of these cases[1] and contributing to complications such as neuropathy, cardiovascular dysfunction, and gastrointestinal disorders. Recent studies have begun to uncover mechanistic links between gut microbial imbalance and metabolic dysfunction, including insulin resistance and inflammation[2]. Uncontrolled blood glucose levels can impair nerves that regulate digestion, leading to conditions like gastroparesis and diabetic enteropathy, which are associated with food transit and microbial balance.

A schematic overview of the gut microbiome as shown in Figure 1 illustrates its complexity and diversity[3]. Digestion itself begins in the mouth and involves a complex series of biochemical steps that are tightly linked to metabolic health[4]. Microbial populations within the gut-including bacteria, fungi, and viruses-play an integral role in orchestrating host metabolic and immune functions, particularly those linked to glucose regulation and inflammation[3,5]. Understanding this connection could open new pathways for non-invasive diagnostics and therapeutic interventions.

The term "microbiome" refers to the entire genetic material of microbial communities inhabiting specific regions of the body, such as the gut, skin, and oral cavities[5]. These microbes play essential roles in nutrient absorption, energy production, and immune defense. Beneficial gut bacteria help produce short-chain fatty acids and key amino acids, including arginine and glutamine, and vitamins such as folic acid and vitamin K[6]. They also support anti-inflammatory processes and protect against pathogenic overgrowth. The structural composition of a fatty acid chain is illustrated in Figure 2, highlighting the hydrocarbon tail and carboxyl group that define its amphipathic nature. However, environmental factors such as poor diet, stress, and antibiotic use are associated with microbial balance, promoting the proliferation of harmful species[7,8]. This dysbiosis, an imbalance in the composition and function of microbial communities, often characterized by a loss of beneficial microbes, an expansion of potentially harmful ones, or reduced diversity, and metabolic deregulation can impair gut function and may be biomarkers of systemic inflammation and insulin resistance[9]. Stool analysis is a widely used non-invasive method to assess gut health

[1] *High school student*
[2] *Independent Researcher*

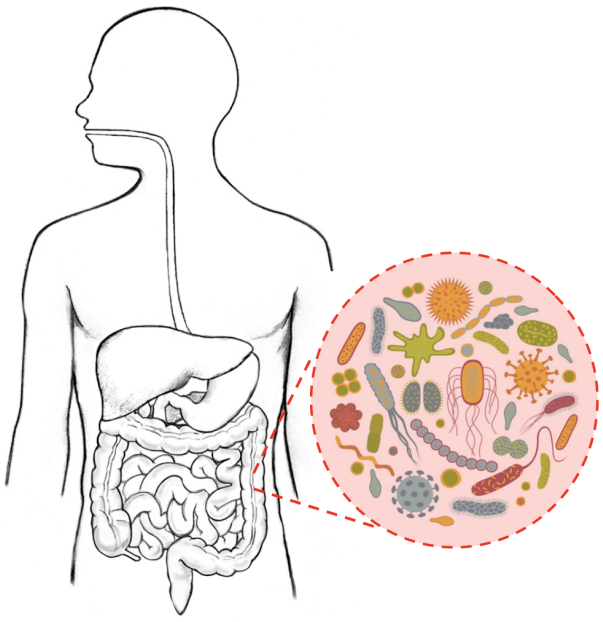

Fig. 1 Gut microbiome schematic (adapted from Han Lab, 2024[3])

and microbial composition[10,11]. Recent studies have identified microbial signatures in type 2 diabetes, highlighting specific taxa and functional pathways that may be biomarker of disease progression[12].

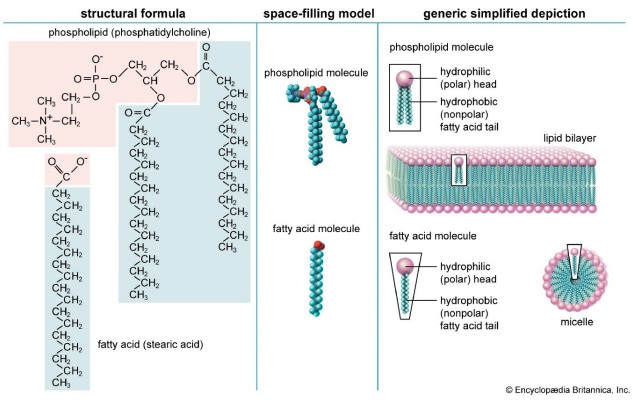

Fig. 2 Fatty acid chain (adapted from Britannica & Rogers, 2019[6])

Emerging studies have linked gut microbiota in the pathophysiology of metabolic conditions, with type 2 diabetes being among the most prominently affected. Disruptions in microbial balance may be biomarkers of complications like gastroparesis, which disproportionately affects women, and diabetic enteropathy, which often presents as constipation. These conditions allow harmful bacteria to multiply rapidly, further compromising gut health[13].

Large-scale initiatives such as the MiBioGen collaboration have begun mapping microbial signatures associated with diabetes, while biotech companies are developing microbiome-based therapies and tools for sequencing and data analysis[14]. Recent reviews have deepened our understanding of how gut microbial dynamics intersect with diabetes pathophysiology, underscoring the promise of microbiome-guided diagnostics and therapies[15]. Additionally, mechanistic insights into microbial contributions to glucose metabolism and insulin sensitivity continue to shape therapeutic strategies for type 2 diabetes[16]. These efforts underscore the growing interest in leveraging microbiome data to understand and potentially treat metabolic diseases.

This study examines how variations in gut microbial composition correlate with type 2 diabetes, using statistical tests and machine learning algorithms to analyze microbial abundance data. Using multivariate linear models and predictive classifiers, the analysis identifies bacterial species associated with diabetic status and evaluates the influence of demographic factors such as sex, age, and Body Mass Index (BMI). The results provide insight into how specific microbes may be biomarkers of metabolic dysfunction and lay the groundwork for future non-invasive diagnostic tools.

The subsequent sections are structured as follows: the Methods section outlines the data acquisition process, preprocessing steps, and exploratory data analysis. The Results section presents findings from statistical tests and machine learning models. The Discussion section interprets these results in the context of existing literature, while the Future Research section outlines plans for Phase 2, including microbial Deoxyribonucleic acid[17] analysis from stool samples[11], collaboration with the MiBioGen initiative[14], and potential partnerships with biotech firms specializing in microbiome therapeutics.

Methods

Data Source and Ethical Approval

The datasets analyzed in this study were originally generated by the Human Microbiome Project Consortium[18] and subsequently compiled and processed into shotgun metagenomic profiles by Pasolli et al.[19]. These processed data are publicly accessible via the Segata Lab GitHub repository (`https://github.com/SegataLab/metaml/tree/master/data`).

For the present study, we performed a secondary analysis by filtering the publicly available dataset to include 489 samples annotated as healthy ($n = 217$), impaired glucose tolerance ($n = 166$), or type 2 diabetes ($n = 106$). This subset was then used for all subsequent analyses focused on metabolic

health outcomes and underwent standardized preprocessing to ensure comparability across cohorts. Specifically, variable names were harmonized, missing values were replaced with equivalent measures (e.g., substituting fasting glucose with FBG, triglycerides with TG, fasting insulin with FINS where available), and redundant columns were removed by Pasolli et al.[19] during dataset compilation. Building on this processed data, we converted all continuous variables to a numeric format, and categorical disease status was encoded as healthy (0), impaired glucose tolerance (1), or type 2 diabetes (2). These steps ensured consistency and reproducibility for downstream statistical and machine learning analyses.

Description of the Data

The dataset used for this study uses machine learning techniques to facilitate meta-analysis across large-scale metagenomic datasets[19]. It contains various medical and microbiome-related variables. The overview of the key features of the dataset is described below:

Medical and demographic variables:

- **Disease:** This is the output variable, likely indicating the presence or type of a disease.
- **Age, gender:** Basic demographic variables.
- **Body Mass Index (BMI):** A measure of body fat based on an individual's weight in relation to their height.
- **Blood pressure:** Includes Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP).

Diabetes-related variables: Includes Fasting Blood Glucose (FBG), Hemoglobin A1c (HbA1c), and insulin levels.

Lipid profile: Includes Triglycerides (TG), Total Cholesterol (TCHO), High-Density Lipoprotein (HDL), and Low-Density Lipoprotein (LDL).

Inflammation Markers: Includes Tumor Necrosis Factor-alpha (TNFα), Interleukin-1 (IL-1family cytokines), and high-sensitivity C-reactive protein (hsCRP), which are commonly used to assess systemic inflammation and disease risk[20].

Microbiome Data: Variables related to specific bacterial taxa, with hierarchical taxonomic information from kingdom to strain or unclassified taxa.

Other Laboratory and Medication variables:

- Statins, insulin, and oral anti-diabetic medication indicate medication usage.
- Various peptides and biomarkers: Adiponectin, Leptin, and Glucagon-Like Peptide-1 (GLP-1).

The intricate relationships linked to type 2 diabetes (T2D) and the human microbiome are examined in this dataset[19], with a particular emphasis on the function of gut bacteria that can form communities that resist pharmaceutical interventions[21] and systemic inflammation on insulin sensitivity. Important questions include determining which gut bacteria affect insulin sensitivity in diabetics, comparing the microbial composition and inflammatory profiles of diabetics and healthy individuals, and identifying bacterial species that may serve as biomarkers for better disease prediction. A comparative analysis of inflammatory markers and microbiome profiles in T2D patients will shed light on the mechanisms behind the connection between gut health and metabolic dysfunction. Furthermore, this study investigates the effects of insulin and oral diabetic medications on an individual's bacterial species, attempting to classify these species based on features of abundance.

Table 1 summarizes the demographic and health characteristics of study participants in our filtered dataset, derived from Pasolli et al.[19] processed data and limited to T2D records.

Table 1 Demographics and health characteristics of study participants (N=489).

Characteristic	Summary / Distribution
Age (years)	Mean = 54.3 SD = 16.1 Range = 13–86
BMI (kg/m^2)	Mean = 24.5 SD = 4.2 Range = 15.6–44.6
Sex distribution	Female: 299 (61.1%) Male: 190 (38.9%)
Health status	Unhealthy: 272 (55.6%) Healthy: 217 (44.4%)
Sex $\times$ Health status	Female Healthy: 133 Female Unhealthy: 166 Male Healthy: 84 Male Unhealthy: 106

Data Collection Process

Human guts are one of the most densely populated habitats on Earth, home to trillions of microorganisms. The dataset utilized for this project was compiled from publicly accessible metagenomic datasets and includes information on the relative abundance of other taxonomic or microbial species within metagenomic samples[19]. Conclusions may be skewed by the inconsistencies and mistakes present in raw data. For example, batch effects occur when distinct sample groups are processed in different ways or at separate times. Technical re-

strictions may also result in insufficient microbiome data[22], since certain microorganisms may not be fully recorded due to constraints in sequencing coverage and analytical sensitivity. Data preparation is used to enhance data quality before analysis to lessen these problems. Preprocessing aids in filling in the gaps in data, minimizing biases and technical noise, and removing features that are not informative.

Standard procedures for data preparation for microbiome sequencing data[22] preparation, illustrated in Figure 3, depict the overall data provenance and preprocessing workflow. While the figure outlines common procedures such as imputation of missing values, batch effect correction, quality filtering, data standardization, and transformation, in this study, we relied on the harmonization and imputation performed by Pasolli et al. (2016) and implemented filtering and encoding as described in Sections Data Source and Ethical Approval and Exploratory Data Analysis. These procedures are essential for ensuring accurate and reproducible downstream analyses. Table 2 summarizes these preprocessing steps, which are essential for ensuring comparability and reproducibility in downstream analyses.

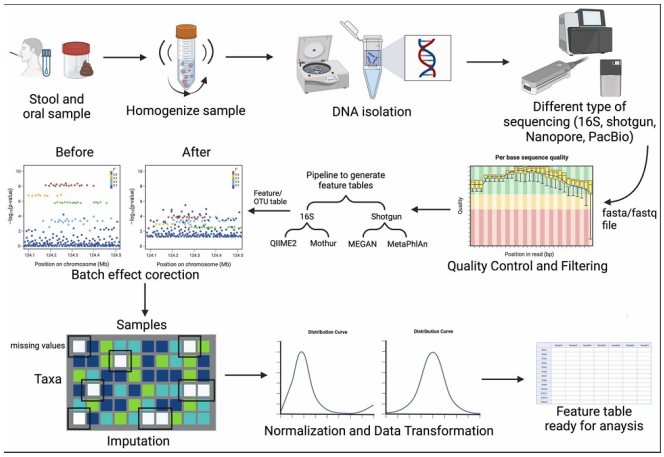

Fig. 3 Data preprocessing workflow for microbiome sequencing (adapted from R. Zhou, S. K. Ng, J. J. Y. Sung, W. W. B. Goh, S. H. Wong. 2023)[22]

MetaPhlAn2[23], a computer tool that allocates metagenomic readings to specific taxonomic groups based on a database of clade-specific marker genes, was used to create the data for this study. The relative abundance is ascertained by mapping the sequencing reads to these flag genes, normalizing by their length and quantity, and averaging the outcomes to obtain the fraction of each species in the whole microbial community. Numerous methods, including read alignment, taxonomic profiling, and standardization, marking unique to a clan were utilized in generating the microbiome data[22]. To further reduce non-biological variation introduced by differences in sequencing runs, cohorts, or laboratories, batch effect correction was performed using the ComBat[24] empirical Bayes framework.

Table 2 Preprocessing of MetaPhlAn2 Feature Tables: Key preprocessing steps applied to microbiome abundance data, with corresponding methods used in this study.

Step & Definition	Method Used
Batch Effect Correction: Removes non-biological variation introduced by differences in sequencing runs, cohorts, or labs.	ComBat[24] (empirical Bayes framework)
Filtering & Imputation: Excludes low-prevalence taxa and replaces zeros/missing values to stabilize downstream transformations.	Prevalence filtering (taxa present in $\geq$10% of samples) + zero replacement (pseudo-count addition)
Normalization: Adjusts abundances to account for sequencing depth and compositional bias, enabling fair comparison across samples.	Relative abundance scaling (MetaPhlAn2[23] output) + Centered log-ratio (CLR) transformation

The Species abundance data provides insight into the associations between human health and the composition of the gut microbiota[14] by utilizing a range of methods that combine more complex microbiome data with traditional clinical parameters. A pattern number of reads for the *Bacteroides caccae* species for each sample, for instance, in the species abundance data is noted in the variable names of the species data. An example is provided below: "k__Bacteria—p__Bacteroidetes—c__Bacteroidia—o__Bacteroidales—f__Bacteroidaceae—g__Bacteroides—s__Bacteroides_caccae". The data described in this pattern is summarized in Table 3.

Table 3 Taxonomic classification patterns of bacterial species based on structured labels

Pattern Name	Description
k_Bacteria	Bacteria Kingdom
p_Bacteroidetes	Bacteroidetes Phylum
c_Bacteroidia	Bacteroidia Class
f_Bacteroidaceae	Bacteroidaceae Family
g_Bacteroides	Bacteroides Genus
s_Bacteroides_caccae	Bacteroides caccae Species

The paper examines how inflammatory markers in the blood, the structure of the gut microbiota, and insulin responsiveness interact in the context of type 2 diabetes. The rela-

tionship between gut flora and inflammation and insulin sensitivity, as well as the differences in the patterns of these two variables in diabetics and healthy people, are important research concerns. Furthermore, the study uses microbial abundance data to classify bacterial species and identify those associated with type 2 diabetes, employing predictive and optimization models through machine learning to determine the impact of insulin and oral antidiabetic drugs on microbial composition.

Exploratory Data Analysis

The gut microbiome includes species such as *Clostridium bolteae* and *Clostridium hathewayi*, both of which have been linked to metabolic conditions such as type 2 diabetes. While these anaerobic, Gram-positive bacteria may be biomarkers of systemic immune activation and resistance to insulin action when overgrown, *Roseburia*-a beneficial butyrate-producing genus-supports gut health and glucose regulation through anti-inflammatory mechanisms. Imbalances in microbial composition are associated with metabolic homeostasis, highlighting the importance of species-level analysis.

In this study, normalized and transformed microbiome data were interpreted using the Python Data Analysis Library, pandas[25], followed by graphical summaries discussed in this section. The dataset[19] comprises microbiome profiles from 489 stool samples analyzed via MetaPhlAn2[23], with 51 features spanning clinical, microbial, laboratory, and medication variables. The primary aim is to compare inflammatory markers and species abundance between healthy and diabetic individuals.

Figure 4 presents a boxplot of log_{10}-transformed read counts for the Gram-positive Clostridium bolteae across three cohorts: individuals without diabetes ("n"), those diagnosed with type 2 diabetes ("t2d"), and participants with impaired glucose tolerance. The median read count is highest in the "t2d" group, intermediate in the healthy cohort, and lowest among those with impaired glucose tolerance. The distribution in all groups exhibits a wide range of variation, suggesting a greater abundance of Clostridium bolteae in individuals with type 2 diabetes.

The boxplot of Gram-negative *Clostridium hathewayi* bacterium in Figure 5 across three groups shows the highest median read count in the "t2d" group, followed by the impaired glucose tolerance group, with the lowest median in the healthy group. Like *Clostridium bolteae*, *Clostridium hathewayi* levels are elevated in individuals with type 2 diabetes. The "t2d" group exhibits the greatest variation in read counts, while both the impaired glucose tolerance and healthy groups show less variability, with the impaired glucose tolerance group displaying higher median levels than the healthy group but lower than those with type 2 diabetes.

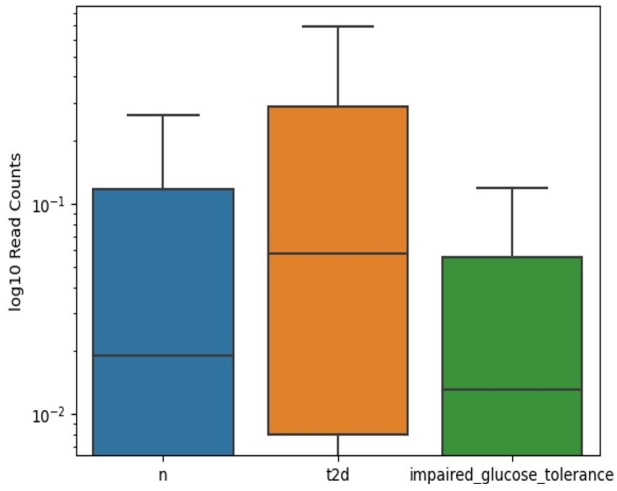

Fig. 4 Boxplot of the Clostridium bolteae species microbiome data

Taxonomic Clarification: Although members of the genus Clostridium are typically Gram positive, Clostridium (Hungatella) hathewayi has been reported to stain Gram negative in clinical isolates[26]. This discrepancy reflects staining behavior rather than phylogenetic classification. In contrast, Roseburia inulinivorans is consistently described as Gram positive rods[27].

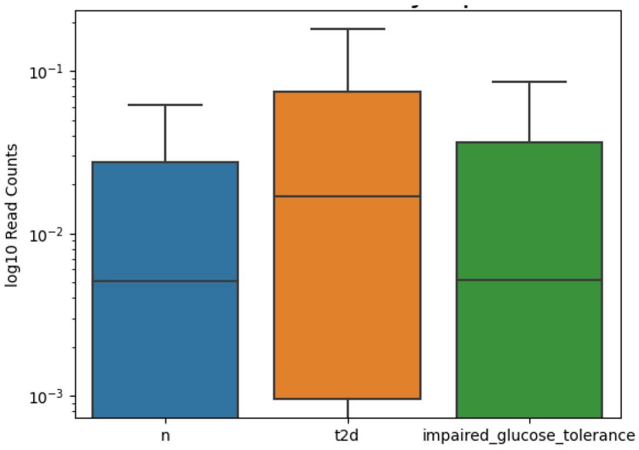

Fig. 5 Boxplot of the Clostridium hathewayi species microbiome data

The boxplot of Gram-positive *Roseburia inulinivorans* bacterium species distribution in Figure 6 across three groups with "t2d" shows the highest median read counts in the healthy group, followed by impaired glucose tolerance, with the lowest in the type 2 diabetes group. This indicates a decline in *Roseburia* abundance from healthy individuals to those with diabetes. The healthy group exhibits the widest range of vari-

abilities, while the "t2d" group has lower median levels and less variation. The impaired glucose tolerance group shows intermediate levels of *Roseburia Inulinivorans*, with moderate variability between the other two groups.

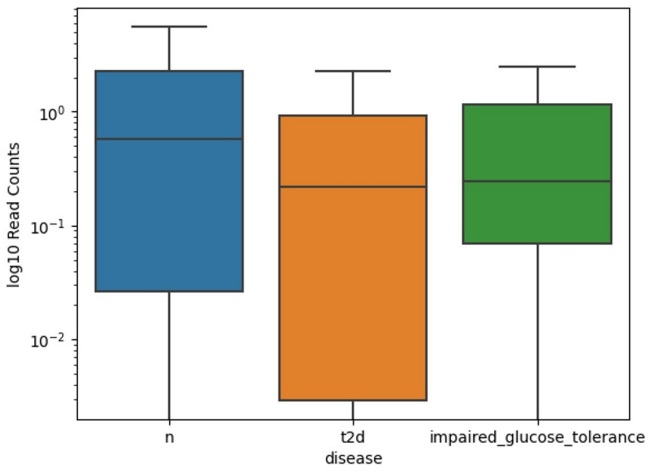

Fig. 6 Boxplot of the Roseburia inulinivorans microbiome data

Results

Guided by principles from Mind on Statistics, this study employed a range of analytical methods to investigate the relationships between microbiome, clinical, and demographic data. Multivariate Analysis of Variance (MANOVA)[28] and Analysis of Variance (ANOVA)[28] were used to assess group-level differences, followed by decision trees[29] and Linear Discriminant Analysis (LDA) for pattern discovery and class separation. Predictive modeling was performed using an Artificial Neural Network (ANN) to capture non-linear relationships and improve classification accuracy. Modern artificial neural networks have evolved to mimic biological intelligence[30].

Machine Learning Models

We employed supervised learning approaches to evaluate classification performance across diagnostic categories (healthy, impaired glucose tolerance, and type 2 diabetes). Decision trees provided transparent, rule-based classification and allowed visualization of feature contributions. LDA offered a linear boundary-based approach with shrinkage regularization to stabilize covariance estimation in moderate-dimensional settings. ANNs enabled nonlinear modeling of complex relationships, with regularization and early stopping applied to enhance generalization. Model performance was assessed using accuracy, precision, recall, and F1-score (defined as the harmonic mean of precision and recall) provide a single measure of a classifier's accuracy that balances false positives and false negatives, with evaluation on held-out test sets to ensure robustness. Table 4 summarizes the full configuration of supervised learning models used in the study, ensuring reproducibility.

Table 4 Machine Model Implementation & Configurations

Model Component	Implementation & Configuration
Decision Tree	DecisionTreeClassifier (scikit-learn v1.2) Criterion: Gini impurity Max depth: 3 Min samples per leaf: 5 Cost complexity pruning: ccp_alpha=0.01 Random seed: 100
LDA	LinearDiscriminantAnalysis (scikit-learn v1.2) Solver: 'lsqr' Shrinkage: 'auto' Missing value handling: Imputation applied Target encoding: LabelEncoder
ANN Architecture	Sequential model (Keras/TensorFlow v2.11) Hidden layers: 64 $\rightarrow$ 32 $\rightarrow$ 16 (ReLU) Output layer: Softmax (3 classes) Regularization: L2 penalty (0.01) Dropout: 0.5 Input features: 12
Optimizer & Training	Adam optimizer Learning rate: 0.001 Loss: Categorical cross-entropy Batch size: 32 Epochs: 50 Early stopping: Patience = 20 (monitoring accuracy) Validation split: 20%

MANOVA (Multivariate Analysis of Variance)

MANOVA evaluates whether the continuous dependent variables (bacterial species in this case) vary as a function of the independent variables (disease and covariates). The disease outcomes ('disease' = impaired_glucose_tolerance, t2d, n) were analyzed for bacterial species ('s_Clostridium_hathewayi', 's_Clostridium_bolteae', 's_Roseburia_inulinivorans') while controlling for covariates

like age, BMI to correct for potential confounding.

Prior to conducting MANOVA, we evaluated the standard assumptions. Univariate normality was assessed using Shapiro–Wilk tests, which indicated significant departures from normality for all bacterial abundances across disease groups ($p < 0.05$). Multivariate normality was tested with Henze–Zirkler's procedure and rejected ($p < 0.001$). Homogeneity of covariance matrices was examined using Box's M test, which revealed heterogeneity across groups ($p < 0.001$). Homogeneity of variances was assessed with Levene's test: equal variances were met for *Clostridium hathewayi* ($p = 0.055$), but violated for *Clostridium bolteae* ($p = 0.025$) and *Roseburia inulinivorans* ($p = 0.015$). Multicollinearity was checked via correlation analysis, with all $|r| < 0.2$, indicating no problematic correlations. Linearity was supported by scatterplot inspection, and outlier detection using Mahalanobis distance flagged 18 multivariate outliers (3.7% of the dataset), which were retained to preserve biological variability.

Given these violations, we relied on Pillai's Trace, which is robust to departures from normality and covariance homogeneity, and we emphasized effect sizes (η^2) rather than p-values alone when interpreting results. The MANOVA results-based on four standard multivariate statistics: Wilks' lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's greatest root[28]-are used to assess whether independent variables (such as sex, age, and BMI) have a statistically significant effect on the combined dependent variables (e.g., microbial species), as modeled in Table 5. The fitted regression equation for the model is:

$$-0.1004 + 0.1376(\text{s\_C.\_hathewayi}) + 0.0395(\text{s\_C.\_bolteae})$$
$$-0.0201(\text{s\_R.\_inulinivorans}) + 0.0171(\text{age}) - 0.0072(\text{bmi})$$

The results indicate that the quantity of the bacterial species is highly influenced by an individual's medical condition, including whether they are healthy, pre-diabetic, or diabetic. While BMI does not seem to have a statistically discernible effect on the number of bacterial species, age does have a small but substantial effect. The linear regression coefficients suggest that various bacterial species may have distinct effects on diabetes status.

The standardized coefficients from the MANCOVA model indicated that *Clostridium hathewayi* ($\beta = 0.138$) and *Clostridium bolteae* ($\beta = 0.040$) were positively associated with disease status, while *Roseburia inulinivorans* ($\beta = -0.021$) showed a negative association. Among covariates, age ($\beta = 0.017$) and BMI ($\beta = 0.007$) contributed only modestly to the model. These values suggest that bacterial taxa contributed more strongly to variation in disease status than demographic covariates.

Further analysis of the effect of different bacterial species on diabetes status (C(disease))-healthy, pre-diabetic, or diabetic was performed to check the effect sizes. Table 6 summary lists the sum of squares, F-values, p-values (PR(>F) denotes the probability of observing a test statistic greater than the calculated F value under the null hypothesis), and effect sizes like eta squared (η^2) and omega squared (ω^2)[31].

Table 5 Multivariate linear model results using MANOVA to assess the impact of sex, age, and BMI on Clostridium hathewayi, Clostridium bolteae, and Roseburia inulinivorans

Predictor	Test Statistic	Value	F Value	Pr > F
Disease Status	Wilks' lambda	0.9678	5.3525	0.0012
	Pillai's trace	0.0322	5.3525	0.0012
	Hotelling–Lawley	0.0322	5.3525	0.0012
	Roy's greatest root	0.0322	5.3525	0.0012
Age	Wilks' lambda	0.9806	3.1810	0.0238
	Pillai's trace	0.0194	3.1810	0.0238
	Hotelling–Lawley	0.0198	3.1810	0.0238
	Roy's greatest root	0.0198	3.1810	0.0238
BMI	Wilks' lambda	0.9971	0.4648	0.7070
	Pillai's trace	0.0029	0.4648	0.7070
	Hotelling–Lawley	0.0029	0.4648	0.7070
	Roy's greatest root	0.0029	0.4648	0.7070

Table 6 ANOVA summary of bacterial classification

Bacteria Classification	F- value	P-Value	Effect size (η)	Omega Squared (ω^2)
s_Clostridium_bolteae	3.8159	0.0227	0.0155	0.0144
s_Clostridium_hathewayi	3.0412	0.0487	0.0124	0.0083
s_Roseburia_inulinivorans	4.7420	0.0091	0.0191	0.0151

There is a statistically significant correlation (p-values all below 0.05) between the three bacterial species and the presence of diabetes. Each species has tiny effect sizes, but s_Roseburia_inulinivorans has the largest correlation with diabetes status, with s_Clostridium_bolteae and s_Clostridium_hathewayi following closely behind. These data imply that the quantity of these bacteria varies according to whether a person is healthy, pre-diabetic, or diabetic, with s__Roseburia_inulinivorans having the most noteworthy difference.

Similar MANOVA tests were conducted for inflammation markers, including TNF-alpha, hsCRP, and IL-1 family cytokines[20]. Among all the markers, only TNF-alpha ($F(2,142) = 4.821, p = 0.0094$) showed a significant effect on disease levels with an eta-squared (η^2) value of 0.0636, suggesting that about 6.36% of the variance in TNF-alpha levels is explained by the disease category (healthy, pre-diabetic, diabetic). This suggests that TNF-alpha levels are likely higher in individuals with diabetes, consistent with the inflammatory

response commonly associated with the condition.

The ANOVA framework was applied to compare group means, following the methodology outlined by Utts and Heckard from Mind on Statistics[28]. ANOVA results of the individual medical and demographic variables and their impact on the disease status in Table 7 reveal that Blood pressure appears to be the most strongly linked to disease, while BMI and HDL have moderate associations, and Table 8 Confidence intervals [95% CI] for five clinical variables comparing impaired glucose tolerance and type 2 diabetes groups to the healthy reference group. Intervals are expressed in bracket notation [lower bound, upper bound]. Blood pressure (SBP, DBP) shows the strongest associations with disease status, BMI and HDL show moderate associations, and LDL shows no significant association (intervals include zero):

Table 7 ANOVA summary of medical and demographic data

Medical and demographic variable	**F-value**	**P-value**	**Effect size (η)**	**Omega Squared (ω^2)**
Body Mass Index	14.42	$8.24e^{-07}$	0.0560	0.0520
Systolic Blood Pressure	11802.37	$2.1e^{-303}$	0.9865	0.9863
Diastolic Blood Pressure	10,148.28	$6.19e^{-293}$	0.9843	0.9841
High-Density Lipoprotein	15.68	$2.56e^{-07}$	0.0622	0.0581
Low-Density Lipoprotein	0.05	0.9515	0.0002	-0.004

Table 8 Confidence intervals [95% CI] for five clinical variables comparing impaired glucose tolerance and type 2 diabetes groups to the healthy reference group.

Medical and Demographic variable	**Impaired glucose tolerance**	**Type 2 diabetes**
Body Mass Index	$[-4.58, -2.04]$	$[-3.36, -0.82]$
Systolic Blood Pressure	$[36.22, 39.60]$	$[43.82, 47.20]$
Diastolic Blood Pressure	$[22.97, 25.20]$	$[25.82, 28.05]$
High-Density Lipoprotein	$[-0.58, -0.23]$	$[-0.68, -0.33]$
Low-Density Lipoprotein	$[-0.24, 0.31]$	$[-0.23, 0.32]$

Confounding Variables: We acknowledge that several important confounders-particularly diabetes medications (metformin, sulfonylureas, insulin), diet, physical activity, statin use, and comorbidities-were not systematically available for all participants. In our main MANOVA analysis, we controlled for age and BMI, which were consistently recorded, but medication and lifestyle data were incomplete. To explore potential medication effects, we conducted a small LDA analysis restricted to individuals with available oral antidiabetic medication data (metformin or sulfonylureas). Due to missing records, this analysis did not yield robust predictive performance. Nevertheless, we report on the LDA framework and coefficients to enable replication and extension by other groups with more complete medication data. In the Discussion section, we emphasize that medication and lifestyle factors may partly drive observed microbiome differences, and that future studies should prioritize medication-naïve cohorts or multivariable models with comprehensive confounder data.

Decision Tree Analysis

Decision tree[29] analysis was conducted using the principles described by Quinlan (1986), which allow for recursive partitioning of data based on predictor variables. The decision tree[29] output gives several insights into how various factors, as age, gender, and microbiome species (particularly s_Clostridium_hathewayi and s_Clostridium_bolteae), impact a person's health status in terms of being healthy, having impaired glucose tolerance (pre-diabetic), or being diabetic. Figure 7 describes the influence of age, microbiomes, and gender on the health of the patient.

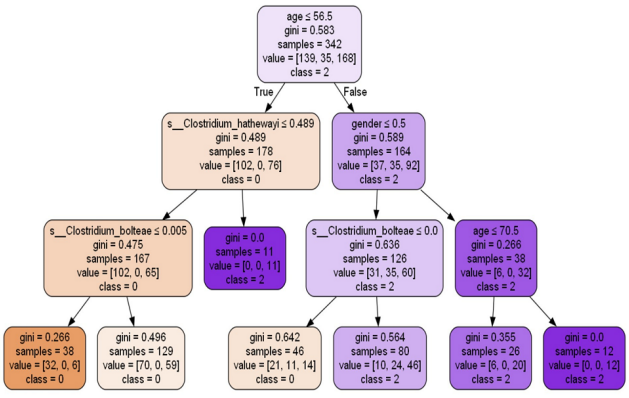

Fig. 7 Decision tree illustrating how age, gender, and microbiome species-particularly s_Clostridium_hathewayi and s_Clostridium_bolteae-influence health status outcomes: healthy, pre-diabetic, or diabetic

Age Influence:

1. The root node starts with age ≤ 56.5 being the primary split. This indicates that age plays a significant role in predicting diabetes, with older individuals (age > 56.5) more likely to be diabetic.

2. As the tree branches further, age $\leq$ 70.5 continues to show a clear impact on the class, suggesting that older age categories tend to be more associated with diabetes (class 2).

   Microbiome Influence:

3. S_Clostridium_hathewayi appears as an important microbial species in the tree. Subjects exhibiting reduced abundance ($\leq$ 0.489) of this species are more likely to be classified as healthy (class 0).

4. S_Clostridium_bolteae shows a similar influence. For instance, those with s_Clostridium_bolteae values $\leq$ 0.005 are also more likely to be healthy (class 0), indicating that the higher presence of these bacterial species might correlate with less favorable health outcomes.

Gender Influence: On the right branch, gender plays a role where females (gender $\leq$ 0.5) appear to be linked to a higher likelihood of being diabetic (class 2), compared to males.

The classification report in Figure 8 shows the model's performance. The overall accuracy of the model is 63.26%, which is moderate. The precision for classifying healthy individuals (class 0) is relatively high (0.70), but the model struggles to correctly classify individuals with impaired glucose tolerance (class 1), as indicated by a precision and recall of 0.0 for this group.

```
Accuracy :  63.26530612244898
Report :                precision    recall  f1-score   support

           0       0.70      0.78      0.74        78
           1       0.00      0.00      0.00        14
           2       0.53      0.58      0.56        55

    accuracy                           0.63       147
   macro avg       0.41      0.45      0.43       147
weighted avg       0.57      0.63      0.60       147
```

Fig. 8 Classification report summarizing the performance of the decision tree model, including precision, recall, and F1-score across health status categories.

ROC and AUC curves: To evaluate the performance of the decision tree classifier across the three disease categories (healthy, impaired glucose tolerance, and type 2 diabetes), we computed one-vs-rest ROC curves and corresponding AUC values displayed in Figure 9. The model achieved an AUC of 0.713 for healthy individuals (Class 0), 0.842 for impaired glucose tolerance (Class 1), and 0.644 for type 2 diabetes (Class 2). These curves provide a visual summary of the classifier's ability to distinguish each class from the others. While the AUC for Class 1 appears high, this reflects the model's confidence rather than its accuracy, as the precision and recall for this class were both zero. These results highlight the importance of interpreting ROC/AUC alongside confusion matrices and classification metrics.

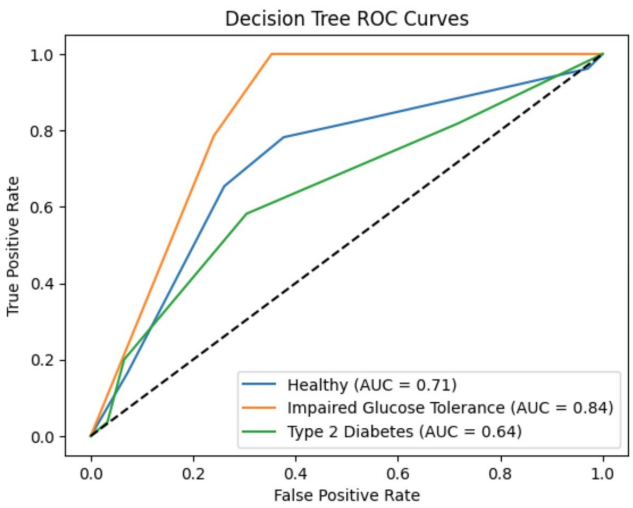

Fig. 9 ROC Curves for Decision Tree Classifier.

Confusion matrix: Table 9 presents the confusion matrix for the decision tree classifier using the Gini index. The model correctly classified 61 of 78 healthy individuals (Class 0), but misclassified all 14 impaired glucose tolerance cases (Class 1), assigning them to other categories. For type 2 diabetes (Class 2), 32 of 55 cases were correctly identified. These results are aligned with the classification report: the model performs well for healthy individuals (precision = 0.70, recall = 0.78), moderately for T2D (precision = 0.53, recall = 0.58), and poorly for IGT (precision = 0.00, recall = 0.00). A baseline model that always predicted T2D achieved only 37% accuracy, confirming that the decision tree provides meaningful improvement despite limitations in identifying intermediate disease states.

Table 9 Confusion matrix for decision tree classifier (Gini index, 3-class diseases outcome).

True Class ↓ / Pred →	**Class 0 (Healthy)**	**Class 1 (IGT)**	**Class 2 (T2D)**
Class 0 (Healthy)	61	0	17
Class 1 (IGT)	3	0	11
Class 2 (T2D)	23	0	32

Linear Discriminant Analysis for Studying the Impact of Medications

Recent findings suggest that the microbial composition in the gut can influence how diabetes medications, such as metformin, are metabolized and how effectively they function. For example, certain bacterial species have been associated

with improved responses to metformin. Linear Discriminant Analysis (LDA) is a statistical method used for dimensionality reduction and classification. In the context of classification, LDA[32] seeks to model the difference between groups by assuming that the data from each class are drawn from multivariate normal distributions that differ in their means but share an identical covariance structure. It operates by constructing a weighted sum of input features that maximizes class separation, making it well-suited for datasets with clear group distinctions and limited sample sizes. The test was conducted to analyze the impact of two drugs: Metformin and sulfonylurea. The summary of the results is shown in Figure 10.

```
Classification Report:
              precision    recall  f1-score   support

         met       0.75      0.75      0.75         4
       sulph       0.00      0.00      0.00         1

    accuracy                           0.60         5
   macro avg       0.38      0.38      0.38         5
weighted avg       0.60      0.60      0.60         5

LDA Coefficients (Impact of bacterial species on medication type):
s__Clostridium_bolteae: 12.191591591612315
s__Clostridium_hathewayi: -0.5074119299846949
s__Roseburia_inulinivorans: -2.4956634436932834
```

Fig. 10 Linear Discriminant Analysis (LDA) summary illustrating how gut microbiome features may be biomarkers of to class separation in diabetes status prediction, including responses to metformin.

Precision, recall, and F1-score: For ”met” (metformin), precision, recall, and F1-score[33] are all 0.75, suggesting that the model correctly identifies 75% of cases where the medication is metformin and does so with relatively high precision. However, for ”sulph” (sulfonylurea), the model fails to correctly predict any case (precision, recall, and F1-score[33] are all 0.00). The result could stem from insufficient sample representation, as the dataset includes just one instance of the ”sulph” category (support = 1). This single data point doesn't provide enough information for the model to learn and predict effectively.

Accuracy: The overall accuracy is 0.60, meaning the model correctly predicted 60% of cases. Given the small sample size (5 instances total), this accuracy figure is not very reliable, and the model is likely to be overfitting to the limited data.

Interpretation of bacterial species coefficients: S_Clostridium_bolteae: The large positive coefficient (12.19) suggests that higher levels of this bacterial species are strongly associated with the likelihood of the medication being ”met”. This implies that s_Clostridium_bolteae may play a significant role in the discrimination between metformin and sulfonylurea users. S_Clostridium_hathewayi: The negative coefficient (-0.50) suggests a slight inverse relationship with ”met”. As levels of s_Clostridium_hathewayi increase, the model tends to predict ”sulph” rather than ”met”, though the effect is small compared to s_Clostridium_bolteae. S_Roseburia_inulinivorans: Similarly, a negative coefficient (-2.49) indicates that higher levels of this bacterial species are associated with ”sulph” rather than ”met”, though the effect is still smaller compared to s_Clostridium_bolteae.

The tiny and unbalanced sample size limits the model's practical application, especially for sulfonylurea. Although there may be some biological validity to the positive correlation between s_Clostridium_bolteae and metformin, a meaningful interpretation is impeded by the absence of information regarding sulfonylureas. The dataset[19] should be enlarged, particularly sulfonylurea users, to provide more dependable insights and better capture the associations between bacterial species and diabetes drug types.

Predicting the Outcome of the Disease Using Artificial Neural Network Models

Through the identification of intricate patterns in large datasets[19], Artificial Neural Networks (ANNs)-potent machine learning models-are increasingly utilized in clinical research to detect disease-related outcomes. Modern ANN[30] architecture has evolved significantly in recent years. ANNs[30] comprise structured layers of interconnected nodes-commonly referred to as neurons-that simulate the architecture of the human brain. These networks collaboratively transform and analyze input data, including microbial, clinical, and demographic variables. Compared to traditional statistical methods, ANNs[30] are better equipped to model non-linear relationships between features like genetic markers, inflammatory proteins, and microbiome composition, enabling researchers to predict outcomes such as diabetes, cancer, or cardiovascular conditions. By identifying underlying structures in data and extending those insights to unfamiliar inputs, these models offer significant value in personalized diagnostics and predictive healthcare. Figures 11 and 12 display the initial model loss and accuracy trends for the ANN[30].

The model was developed with an input layer containing 12 features derived from clinical and microbiome data. The neural network architecture began with a first hidden layer containing 64 neurons, each utilizing the Rectified Linear Unit (ReLU) activation function introduces non-linearity by outputting zero for negative inputs and the input value for positive inputs, enabling efficient training of deep networks. Two subsequent hidden layers-comprising 32 and 16 neurons, respectively - were positioned between the input and output layers. The output layer included 3 neurons with a SoftMax activation function (converts raw output scores into normalized probabilities, facilitating multi-class classification), corresponding to three classification categories: Diabetic, Impaired Glucose Tolerance, and Healthy individuals. Based on the model loss

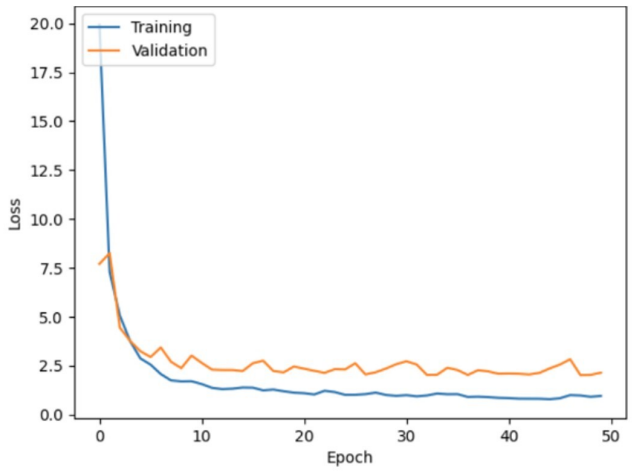

Fig. 11 Model loss trend plot for the ANN before hyperparameter tuning

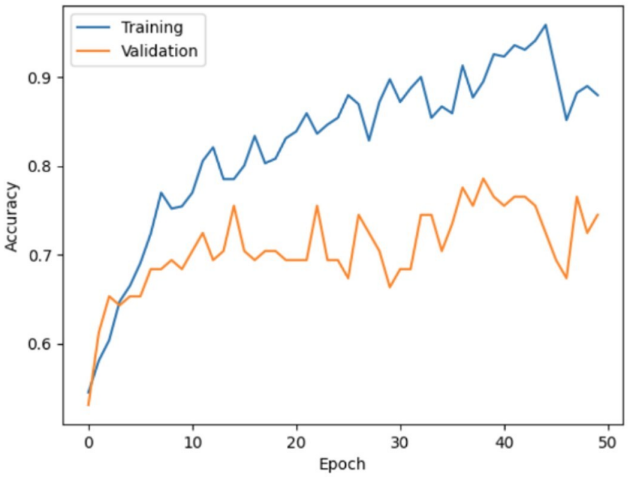

Fig. 12 Model accuracy trend plot for the ANN before hyperparameter tuning

and accuracy graphs, we can draw the following conclusions:

Training vs. Validation Performance:

- A consistent decline in training loss accompanied by rising training accuracy indicates effective model learning from the input data.
- Validation loss does not follow the same trend. It fluctuates significantly after the initial epoch and does not stabilize, implying that the model is not generalizing well to the validation data. This is supported by the validation accuracy, which also fluctuates and does not show consistent improvement.

Overfitting:

- The training accuracy reaches around 95%, while the validation accuracy remains much lower, fluctuating around 70-75%. This large gap between validation and training accuracy indicates that the model has become overly tailored to the training data.
- Overfitting arises when the model memorizes the training data, including noise or non-generalizable features, leading to diminished performance on new inputs.

Possible Remedies: Early stopping may mitigate overfitting by halting training once the validation loss ceases to improve, preventing the model from over-adapting to the training set. Regularization techniques like dropout and L1/L2 regularization[34] could help reduce overfitting by making the model less sensitive to the training data. Increasing the dataset size or applying data augmentation can expose the model to more diverse patterns, improving its ability to generalize. Adjusting the learning rate may help the model converge more effectively, especially if it struggles to minimize validation loss. Model tuning for this case study incorporated callbacks for early stopping, L2 regularization, and a dropout layer added after the second hidden layer. After applying these modifications, the model was re-run, and the resulting performance plots are presented in Figures 13 and 14.

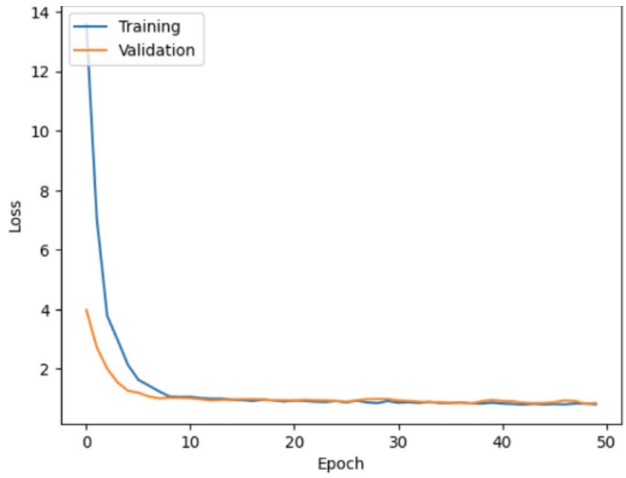

Fig. 13 Model loss trend plot for the ANN after hyperparameter tuning

In these updated graphs, it seems that the model's performance has improved significantly compared to the previous version: Both the training and validation loss have dropped to near zero, with a smooth convergence around 5-10 epochs. These results suggest that the model is effectively learning from the data, with reduced risk of overfitting compared to earlier iterations. Training accuracy varies between 60% and 70%, and validation accuracy closely mirrors this trend. The

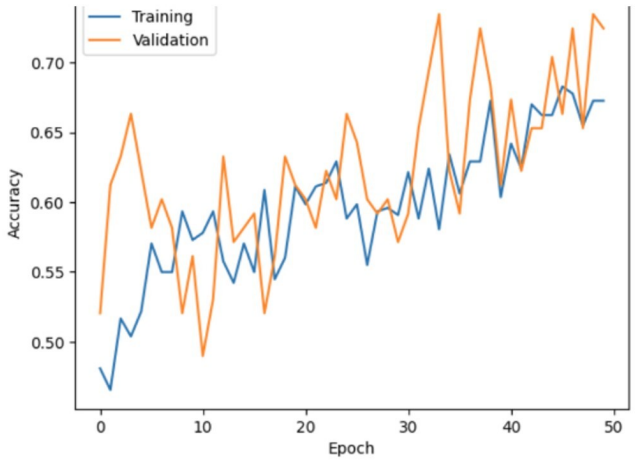

Fig. 14 Model accuracy trend plot for the ANN after hyperparameter tuning

minimal gap between the two performance curves indicates strong generalization to unseen data. Compared to the previous run, the validation accuracy is now much more stable and closer to the training accuracy, indicating that overfitting is significantly reduced, and the model has become more robust. The model accuracy is now better for both the validation and training sets, and the validation accuracy is very close to the training, indicating that the tuning (likely through hyperparameter adjustments or regularization) resulted in a much better generalized performing model. The loss curves show that the model converges quickly, and the reduced overfitting indicates effective tuning.

Discussion

Based on the various statistical tests, there exists a significant association between the gut microbiome and diabetes status. Individuals with type 2 diabetes exhibit higher levels of dysbiosis, marked by an imbalance in bacterial species such as *Clostridium hathewayi* and *Clostridium bolteae*, which may be biomarkers of increased inflammation and worsened glucose control. The reduction in *Roseburia inulinivorans* from healthy individuals to individuals exhibiting reduced glucose tolerance and diabetes highlights its protective role[35] in metabolic health. Additionally, inflammatory markers like TNF-alpha show a strong correlation with diabetes, while IL-1 family cytokines (interleukin) though not statistically significant, suggests a possible role in the disease's progression.

The MANOVA results confirm that the abundance of specific bacterial species varies significantly with diabetes status, particularly *Roseburia inulinivorans*. Moreover, decision tree analysis[29] underscores the importance of age, microbiome composition, and gender in predicting diabetes, with older age and higher levels of *Clostridium species* being linked to poorer health outcomes. The model's accuracy is moderate (63.26%), and while it performs well in classifying healthy individuals, it struggles with those exhibiting impaired glucose tolerance. These insights suggest that age, gut microbiome composition, and inflammation are crucial factors in diabetes development and progression.

The ReLU activation function-based revised training results of the ANN[30] model demonstrate a notable enhancement in the model's performance. Within 5–10 epochs, the training and validation losses converge gradually to almost zero, suggesting effective learning and less overfitting. The training and validation accuracy graphs show a tight alignment, both varying between 60% and 70%, according to the accuracy graphs. This small gap indicates good generalization to unknown data, which is a distinct improvement over the prior version. Furthermore, after regularization or hyperparameter adjustment, the validation accuracy is consistent and closely matched, indicating a stronger model and verifying the improved performance and dependability of the model.

Because this study is cross-sectional and exploratory, the observed associations between microbiome composition and metabolic outcomes cannot establish causality. Only longitudinal or interventional studies can determine whether microbiome changes may be biomarkers of the development of T2D or result from it. Our findings should therefore be interpreted as hypothesis-generating, requiring replication and confirmation in independent cohorts.

Future Research

The role of the gut microbiota in both the development and management of type 2 diabetes (T2D) has gained substantial attention in recent years[14]. Chronic inflammation and insulin resistance, two important variables in the development of type 2 diabetes, have been connected to dysbiosis, an imbalance in the gut microbiota. According to studies, depending on their variety and abundance, specific bacterial strains in the gut can either improve or correlate with the outcomes of diabetes by producing metabolites such as short-chain fatty acids that can affect immunological responses and glucose metabolism. For example, certain probiotics have demonstrated potential in assisting diabetic patients in managing their blood glucose levels.

A potential link between dysbiosis in the gut microbiota and the development of type 2 diabetes has also been revealed by a recent study that used Mendelian randomization techniques[36]. Researchers verified links between specific microbial families and the incidence of type 2 diabetes by examining genetic markers in a variety of groups, highlighting the gut microbiota as a changeable element that may be targeted for diabetes

prevention and management. In the future, the area is heading towards personalized[37] medicine techniques that could allow for the customization of microbiome-based treatments for each patient according to their gut microbiota patterns.

Behavioral traits such as aggression may be modulated by the gut-brain axis, which has been shown to play a critical role in neurobehavioral regulation[35]. To advance understanding of microbiota-host dynamics and uncover novel therapeutic targets within the gut ecosystem, future studies may investigate the integration of fecal microbiota transplantation, probiotics, and prebiotics into diabetes management approaches. With continued research assisting in a better understanding and utilization of the microbiome's involvement in metabolic health, these developments point to a bright future for microbiome-based therapies aimed at regulating type 2 diabetes.

Building on the current findings, the second phase of this study will focus on exploring non-invasive approaches to assess gut microbial shifts linked to type 2 diabetes. To examine the relationship between microbial composition and glucose metabolism, microbial DNA analysis from stool samples will serve as a foundational technique[11]. This phase will also aim to engage with large-scale initiatives such as the MiBioGen collaboration, which aggregates microbiome data and has investigated links between gut microbial profiles and type 2 diabetes[14]. In addition, partnerships with biotech companies specializing in microbiome therapeutics may provide access to sequencing instruments, data processing software, and other resources that could significantly support the expansion of this research.

Limitations

This study has several limitations that should be considered when interpreting the findings. The dataset was derived from publicly available metagenomic resources, which provided breadth but limited demographic detail (age, sex, BMI only), restricting generalizability. Subgroup analyses were underpowered, particularly for medication effects, with only 22 participants reporting treatment data. The cross-sectional design prevents causal inference, so associations should be interpreted as correlational and hypothesis-generating. Statistical assumption checks revealed departures from normality and covariance homogeneity in MANOVA; we therefore relied on Pillai's Trace for robustness and emphasized effect sizes rather than p-values alone. Machine learning performance was modest, with the decision tree failing to classify impaired glucose tolerance cases and the ANN achieving only moderate accuracy, underscoring challenges with imbalanced microbiome datasets. Finally, preprocessing decisions such as harmonizing variable names and imputing missing values may introduce bias, though these steps were necessary for consistency. Taken together, these limitations highlight the exploratory nature of the study and the need for larger, more balanced cohorts with detailed demographic and clinical data to validate and extend these findings.

Data Availability

The datasets analyzed in this study are publicly available from the Segata Lab GitHub repository (`https://github.com/SegataLab/metaml/tree/master/data`; Pasolli[19] et al. The original data were generated by the Human Microbiome Project Consortium[18]. Analysis code is available from the corresponding author upon reasonable request.

References

1 B. B. Duncan, D. J. Magliano and E. J. Boyko, *Nephrology Dialysis Transplantation*, 2025.

2 A. Barlow and V. Mathur, *Journal of the Endocrine Society*, 2023, **7**, 1–12.

3 H. Lab, *Gut microbiome schematic*, `https://thehanlab.com/publications`, 2024.

4 X. A. Singh, *Journal of Hepatology and Gastrointestinal Disorders*, 2014.

5 X. R. Joos, K. Boucher, A. Lavelle, M. Arumugam, M. J. Blaser, M. J. Claesson *et al.*, *Nature Reviews Microbiology*, 2024, **22**, 707–720.

6 K. Rogers, *Fatty acid: Definition, structure, functions, properties, & examples*, `https://www.britannica.com/science/fatty-acid`, 2019.

7 A. B. Shreiner, J. Y. Kao and V. B. Young, *Current Opinion in Gastroenterology*, 2015, **31**, 69–75.

8 A. M. Valdes, J. Walter, E. Segal and T. D. Spector, *British Medical Journal (BMJ)*, 2018, **361**, k2179.

9 L. Zhao, F. Zhang, X. Ding and Y. Wang, *Frontiers in Endocrinology*, 2022, **13**, 1001234.

10 X. Q. Tang, G. Jin, G. Wang, T. Liu, X. Liu, B. Wang and H. Cao, *Frontiers in Cellular and Infection Microbiology*, 2020, **10**, 151.

11 J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang *et al.*, *Nature*, 2012, **490**, 55–60.

12 M. Gurung, Z. Li, H. You, R. Rodrigues and Q. Cao, *Nature Medicine*, 2020, **26**, 1576–1583.

13 M. Camilleri, *Gastroenterology*, 2017, **152**, 1679–1693.

14 O. Okobi, *Medical Research Archives*, 2024, **12**, 5454.

15 M. D'Alessio, J. M. Gonzalez and T. R. Patel, *Diabetes Care*, 2024, **47**, 1489–1499.

16 M. N. Sikalidis and J. Maykish, *Biomedicines*, 2020, **8**, 8–22.

17 J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh *et al.*, *Nature*, 2010, **464**, 59–65.

18 Human Microbiome Project Consortium, *Nature*, 2012, **486**, 207–214.

19 E. Pasolli, D. T. Truong, F. Malik, L. Waldron and N. Segata, *PLOS Computational Biology*, 2016, **12**, e1004977.

20 C. Wunderle, E. Martin, A. Wittig, P. Tribolet, T. A. Lutz, C. Köster-Hegmann *et al.*, *Journal of Inflammation*, 2025, **22**, 16.

21 H. M. Seidler, A. J. Salinas, J. R. Marchesi and P. D. Cotter, *Nature Reviews Microbiology*, 2022, **20**, 353–368.

22 R. Zhou, S. K. Ng, J. J. Y. Sung, W. W. B. Goh and S. H. Wong, *Computational and Structural Biotechnology Journal*, 2023, **21**, 4804–4815.

23 D. T. Truong, E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli *et al.*, *Nature Methods*, 2015, **12**, 902–903.

24 Y. Zhang, G. Parmigiani and W. E. Johnson, *NAR Genomics and Bioinformatics*, 2020, **2**, lqaa078.

25 The pandas development team, *pandas-dev/pandas: pandas [software]*, Zenodo, 2020.
26 P. C. Y. Woo, S. K. P. Lau, J. L. L. Teng, H. Tse, K. H. Chan and K. Y. Yuen, *Journal of Clinical Microbiology*, 2004, **42**, 530–534.
27 S. H. Duncan, G. Holtrop, A. G. Calder, C. S. Stewart and H. J. Flint, *International Journal of Systematic and Evolutionary Microbiology*, 2006, **56**, 2437–2441.
28 J. M. Utts and R. F. Heckard, *Mind on Statistics*, Brooks/Cole, 2nd edn, 2006.
29 J. R. Quinlan, *Machine Learning*, 1986, **1**, 81–106.
30 A. Bahmer, D. Gupta and F. Effenberger, *Neural Computation*, 2023, **35**, 765–780.
31 A. P. Field, *Discovering Statistics Using IBM SPSS Statistics*, Sage Publications, 5th edn, 2017.
32 L. Qu and Y. Pei, *Processes*, 2024, **12**, 1382.
33 D. J. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, Springer, 2nd edn, 2021.
34 O. Demir-Kavuk, M. Kamada, T. Akutsu and E.-W. Knapp, *BMC Bioinformatics*, 2011, **12**, 412.
35 J. F. Cryan, K. J. O'Riordan, C. S. Cowan, T. F. Sandhu, G. Bastiaanssen, M. Boehme *et al.*, *Physiological Reviews*, 2019, **99**, 1877–2013.
36 K. Sun, G. Yan, H. Wu and X. Huang, *Frontiers in Public Health*, 2023, **11**, 1255059.
37 J. Wang, Y. Li, J. Qin, J. Li, Y. Zhang, J. Yu *et al.*, *Nature Medicine*, 2021, **27**, 155–164.