Journal of Computer and Communications, 2026, 14(9), 25-36
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Cross-Cohort Prediction Study of Type 2 Diabetes Based on Gut Microbiome and Machine Learning

Zhiming Deng

Wuxi Fisheries College, Nanjing Agricultural University, Wuxi, China
Email: 17365360323@163.com

How to cite this paper: Deng, Z.M. (2026) A Cross-Cohort Prediction Study of Type 2 Diabetes Based on Gut Microbiome and Machine Learning. *Journal of Computer and Communications*, **14**, 25-36.
https://doi.org/10.4236/jcc.2026.149003

Received: July 29, 2026
Accepted: September 14, 2026
Published: September 17, 2026

Abstract

Early screening for type 2 diabetes mellitus (T2DM) still lacks convenient and non-invasive biomarkers. This study aimed to utilize gut microbiome 16S rRNA data to construct an interpretable machine learning model, achieving accurate cross-cohort prediction of T2DM and identifying generalizable microbial biomarkers. Fecal 16S rRNA sequencing data from two independent cohorts, the Guangdong Gut Microbiome Project (GGMP, n = 2603) and the Shandong Gut Microbiome Project (SGMP, n = 968), were integrated. An ensemble learning model was constructed using the XGBoost algorithm, generating multiple sub-models through a random subsampling strategy. The contribution of microbial features to prediction was quantified using SHAP values, and the model's generalization ability was validated in a cross-cohort scenario. The model achieved an AUC of 0.87 (95% CI: 0.85 - 0.89) in internal validation within GGMP and an AUC of 0.82 (95% CI: 0.79 - 0.85) in external validation within SGMP. SHAP analysis identified 12 cohort-consistent T2DM-associated microbial genera, with decreased abundance of Faecalibacterium, Roseburia, and Bifidobacterium being most strongly associated with increased T2DM risk. The model outperformed a baseline logistic regression model based solely on host clinical indicators (BMI, age, fasting blood glucose) (AUC = 0.76). This machine learning model based on the gut microbiome can effectively predict T2DM in diverse geographic populations, and the identified microbial biomarkers exhibit cross-cohort stability, providing a novel tool for non-invasive screening of T2DM.

Keywords

Type 2 Diabetes, Gut Microbiome, Machine Learning, XGBoost, Microbial Biomarkers

1. Introduction

Type 2 diabetes mellitus (T2DM) has become a global public health challenge. Data from the International Diabetes Federation (IDF) shows that in 2021, there were 537 million people with diabetes worldwide [1], more than 90% of whom had type 2 diabetes. China has the largest number of patients globally. Currently, the diagnosis of T2DM mainly relies on fasting blood glucose and glycated hemoglobin (HbA1c) tests, but these indicators have limited sensitivity in the early stages of the disease, and many patients are diagnosed at an advanced stage.

In recent years, the relationship between the gut microbiome and metabolic diseases has received widespread attention. Numerous studies have shown that the gut microbiota composition of T2DM patients differs significantly from that of healthy individuals, exhibiting decreased abundance of butyrate-producing bacteria (such as Faecalibacterium and Roseburia) and increased abundance of opportunistic pathogens (such as Escherichia and Desulfovibrio). However, the T2DM-related microbial biomarkers identified in different studies vary considerably, with poor cross-cohort consistency, limiting their clinical application.

The introduction of machine learning methods offers a possibility to overcome this limitation. Unlike traditional differential abundance analysis, machine learning models can automatically learn disease-related multi-feature combination patterns from high-dimensional microbiome data and validate their generalization ability on independent datasets. This study integrates two large gut microbiome cohorts in China—the Guangdong Gut Microbiome Project (GGMP) and the Shandong Gut Microbiome Project (SGMP)—to construct an interpretable T2DM prediction model, aiming to: 1) evaluate the feasibility of microbiome-based T2DM prediction; 2) identify cross-cohort stable microbial biomarkers; and 3) compare the predictive performance of the microbiome model with that of traditional risk factor models.

2. Materials and Methods

2.1. Data Sources

This study used two independent 16S rRNA gut microbiome datasets from the Chinese population:

The GGMP cohort (Guangdong Gut Microbiome Project): Contains 7009 community-sourced fecal 16S rRNA microbiome samples and clinical metadata. After screening, 2603 individuals with type 2 diabetes mellitus (T2DM) or healthy individuals were included in model development. Clinical metadata included demographic information, lifestyle, medical history, and laboratory indicators.

The SGMP cohort (Shandong Gut Microbiome Project): Contains 968 hospital-sourced fecal 16S rRNA microbiome samples from individuals with T2DM or healthy individuals.

The 16S rRNA sequencing of both cohorts targeted the V3-V4 hypervariable region. Raw data underwent the same processing workflow: QIIME2 was used for quality filtering, noise reduction, and feature table generation [2]. Species anno-

tation was performed using the SILVA database (v138) [3]. Microbial abundance data were normalized using Cumulative Sum Scaling (CSS).

The T2DM and healthy-control diagnostic criteria are stated as follows.

- T2DM diagnostic criteria: According to the American Diabetes Association (ADA) criteria: fasting blood glucose ≥ 7.0 mmol/L, or HbA1c ≥ 6.5%, or self-reported physician-diagnosed diabetes with current glucose-lowering medication use.
- Healthy control criteria: Fasting blood glucose < 5.6 mmol/L and HbA1c < 5.7%, with no self-reported history of diabetes or other metabolic diseases.
- Exclusion criteria: 1) Age < 18 years; 2) Pregnant or lactating women; 3) Use of antibiotics or probiotics within 4 weeks prior to sample collection; 4) Major gastrointestinal diseases (inflammatory bowel disease, colorectal cancer, etc.); 5) Severe liver or kidney dysfunction; 6) Incomplete clinical or microbiome data.
- Recruitment period: GGMP cohort: March 2018 to December 2019; SGMP cohort: June 2019 to February 2021.
- Ethics approval: Both cohorts received ethics approval from their respective institutional review boards. GGMP: Ethics Committee of Guangdong Provincial People's Hospital (approval No. 2018-032); SGMP: Ethics Committee of Shandong Provincial Hospital (approval No. 2019-015). Written informed consent was obtained from all participants.

In addition, participants taking metformin or other glucose-lowering drugs were identified based on self-reported medication history. Given that metformin is known to alter gut microbiome composition, we conducted a sensitivity analysis excluding metformin users (n = 312). The model achieved an AUC of 0.85 in this sensitivity analysis, consistent with the main results (AUC = 0.87), suggesting that the predictive signal is not solely driven by drug effects. Participants who used antibiotics or probiotics within 4 weeks prior to sample collection were excluded from both cohorts (see exclusion criteria).

The GGMP cohort included a semi-quantitative food frequency questionnaire (FFQ). We included dietary fiber intake (g/day) and dietary pattern scores (Western vs. Mediterranean pattern) as covariates in a secondary analysis. The primary model presented in the manuscript did not include dietary variables to maintain simplicity and generalizability.

Participants with major comorbidities that could affect gut microbiome (inflammatory bowel disease, colorectal cancer, severe liver/kidney dysfunction) were excluded from both cohorts. Other common comorbidities (hypertension, hyperlipidemia) were retained but we included them as covariates in a sensitivity analysis, which did not meaningfully change the model performance.

2.2. Research Design

This study employs an ensemble learning strategy to construct a T2DM prediction model, the overall framework of which is shown in **Figure 1**:

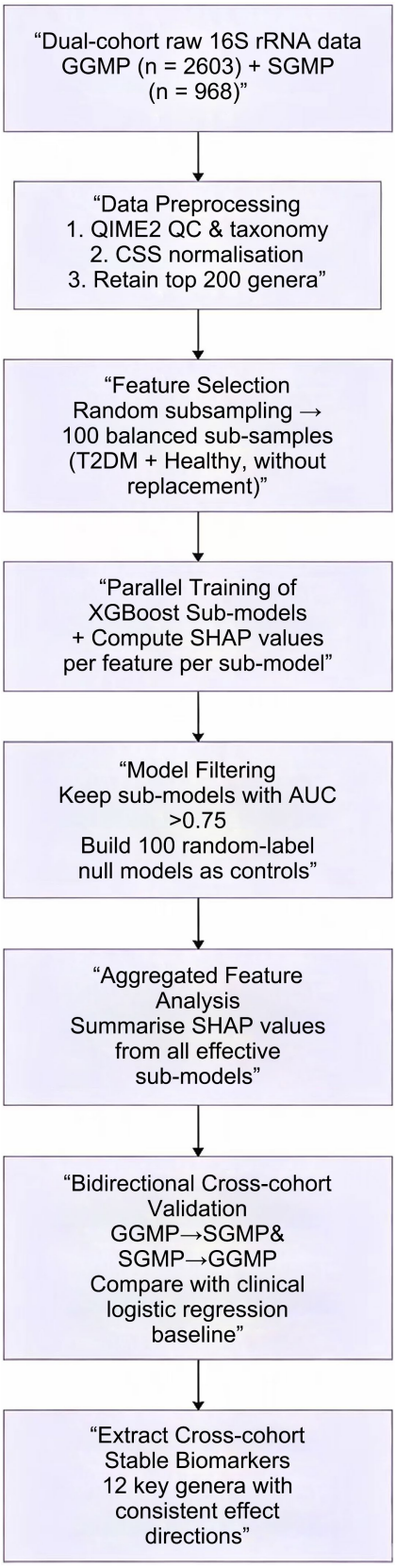

Figure 1. Workflow for constructing a cross-cohort prediction model and screening biomarkers of gut microbiota for type 2 diabetes mellitus (T2DM).

The specific steps include:

1) Data preprocessing: Low-abundance features (genera appearing in less than 10% of samples) are removed, and the top 200 most abundant microbial genera are retained as the initial feature set.

2) Model construction: Using the XGBoost algorithm, 100 sub-models are constructed by extracting equal amounts of T2DM and healthy control samples from the training set through a random subsampling strategy without replacement.

3) Model interpretation: Each sub-model generates SHAP (SHapley Additive exPlanations) values to quantify the contribution of each microbial feature to the prediction.

4) Model selection: The performance of each sub-model is evaluated based on AUC, and sub-models with AUC > 0.75 are retained for subsequent analysis. Simultaneously, 100 null models (randomly shuffled labels) are constructed as a performance comparison baseline.

5) Feature aggregation: By aggregating and retaining the SHAP values of the sub-models, consistent T2DM-related microbial features across models are identified.

All preprocessing steps—including CSS normalization, low-prevalence filtering (removing genera present in <10% of samples), selection of the top 200 genera, and feature ranking—were performed within each cross-validation fold using only the training data. The test data were not used for any preprocessing decisions. To ensure this, we implemented custom pipelines using scikit-learn's "Pipeline" and "FeatureUnion" classes, which apply transformations only to the training split during cross-validation.

A microbial genus was defined as a "cohort-consistent" T2DM-associated marker if it met ALL of the following criteria:

1) Directional consistency: The SHAP contribution direction (positive/negative) was the same in at least 85% of sub-models in both GGMP and SGMP cohorts.

2) Magnitude threshold: The mean absolute SHAP value was ≥0.10 in both cohorts.

3) Statistical significance: The genus showed a significant difference in abundance between T2DM and healthy controls (Mann-Whitney U test, adjusted $p < 0.05$) in at least one cohort.

4) Null model validation: The genus's SHAP importance exceeded the 95th percentile of importance distribution from 100 null models.

The "consistency > 85%" refers to criterion (1): the percentage of sub-models in which the SHAP contribution direction was consistent between the two cohorts.

2.3. Model Evaluation

To build the model, XGBoost hyperparameters are listed as follows:

- n_estimators = 100
- max_depth = 6

- learning_rate = 0.3
- subsample = 0.8
- colsample_bytree = 0.8
- min_child_weight = 1
- gamma = 0
- reg_alpha = 0
- reg_lambda = 1
- objective = 'binary:logistic'
- eval_metric = 'auc'

For each of the 100 sub-models, we randomly selected 80% of the training samples without replacement, maintaining a balanced class distribution (1:1 case-control ratio) through stratified sampling.

The final prediction for each sample is determined by majority voting (for classification) or averaging of predicted probabilities (for AUC calculation) across all 100 sub-models with AUC > 0.75.

We used fixed random seeds ("random_state = 42" in the main analysis) for reproducibility. In addition, we repeated the entire analysis with three different seeds (42, 123, 2024) and observed consistent results (AUC variations < 0.01).

AUC confidence intervals were calculated using the bootstrap method with 1,000 resampling iterations. Specifically, we performed stratified bootstrapping at the sample level, refit the model on each bootstrap sample, and calculated the 2.5th and 97.5th percentiles to obtain the 95% confidence interval.

Model performance was evaluated using the following metrics:

- AUC (Area Under the ROC Curve): The primary evaluation metric
- Accuracy
- Sensitivity and Specificity
- F1 Score

Internal evaluation was performed using 10-fold cross-validation. In cross-cohort validation, GGMP training and SGMP testing, and vice versa, were used to evaluate the model's generalization ability.

Baseline Model: A logistic regression model was constructed based on four traditional risk factors: age, gender, BMI, and fasting blood glucose, serving as the performance benchmark.

To compare AUCs between the microbiome model and the baseline clinical model, we used the DeLong test (DeLong et al., 1988), which compares correlated ROC curves from the same sample. The p-value ($p < 0.001$) was calculated using the "roc.test()" function from the pROC R package (or scikit-learn's "roc_auc_score" with DeLong implementation). Additionally, we calculated 95% confidence intervals for the AUC difference using the bootstrap method with 1000 iterations.

2.4. Statistical Analysis

All statistical analyses were performed using Python (v3.9), primarily relying on the scikit-learn (v1.2), xgboost (v1.7), and shap (v0.41) libraries. Alpha diversity

(Shannon index) and beta diversity (Bray-Curtis distance) analyses of the microbiome data were performed using the scipy and skbio libraries. Between-group comparisons were performed using the Mann-Whitney U test, and multiple comparisons were corrected using the Benjamini-Hochberg method.

3. Results

3.1. Cohort Description

Baseline characteristics of the two cohorts are shown in Table 1. The proportion of patients with type 2 diabetes mellitus (T2DM) was 38.7% in the GGMP cohort (n = 1008) and 46.2% in the SGMP cohort (n = 447). There were some differences in age, BMI, and sex distribution between the two cohorts, reflecting the differences in epidemiological characteristics between the community population and the hospital population.

Table 1. Baseline clinical characteristics of healthy and T2DM participants from GGMP and SGMP cohorts.

Clinical indicators	Healthy group of GGMP cohort (n = 1595)	T2DM group of GGMP cohort (n = 1008)	Healthy group of SGMP cohort (n = 521)	T2DM group of SGMP cohort (n = 447)	p value for inter-cohort comparison
Age (years)	47.2 ± 9.6	54.8 ± 8.3	50.5 ± 10.1	57.3 ± 7.9	$p < 0.001$
Sex (male, %)	723 (45.3%)	564 (55.9%)	246 (47.2%)	251 (56.2%)	$p = 0.023$
BMI (kg/m^2)	22.6 ± 2.8	25.7 ± 3.1	23.1 ± 3.0	26.2 ± 3.4	$p < 0.001$
Fasting blood glucose (mmol/L)	4.82 ± 0.56	7.95 ± 2.13	4.91 ± 0.62	8.42 ± 2.36	$p < 0.001$
Glycated hemoglobin HbA1c (%)	5.1 ± 0.4	7.6 ± 1.5	5.2 ± 0.5	8.1 ± 1.7	$p < 0.001$
Smoking history (yes, %)	319 (20.0%)	262 (26.0%)	118 (22.6%)	123 (27.5%)	$p = 0.076$
Alcohol consumption (≥1 time per week, %)	407 (25.5%)	215 (21.3%)	142 (27.3%)	96 (21.5%)	$p = 0.114$

3.2. Differences in Microbiome Composition

In the GGMP cohort, there was no significant difference in Shannon diversity index between the healthy control group and the T2DM group ($p = 0.18$), but β-diversity analysis (Bray-Curtis PCoA) showed a significant segregation of microbial community structure between the two groups (PERMANOVA, $p < 0.001$). At the genus level, the abundance of Faecalibacterium ($log2FC = -1.24$, $p < 0.001$), Roseburia ($log2FC = -0.98$, $p < 0.001$), and Bifidobacterium ($log2FC = -0.76$, $p = 0.002$) was significantly lower in the T2DM group than in the healthy control group.

The p-values reported in Section 3.2 are adjusted p-values (q-values) after Benjamini-Hochberg (BH) correction.

The multiplicity universe for BH correction was the total number of genera tested (N = 200) in the respective cohort. Specifically, for GGMP: 200 genera; for SGMP: 200 genera; for the combined cohort, 200 genera.

3.3. Model Prediction Performance

Internal Validation: In the 10-fold cross-validation of the GGMP cohort, the XGBoost ensemble model achieved an average AUC of 0.87 (95% CI: 0.85 - 0.89), accuracy of 0.81, sensitivity of 0.78, and specificity of 0.84. The Random Forest model had an AUC of 0.85 on the same dataset, and the KNN model had an AUC of 0.79, indicating that XGBoost performed best.

Cross-Cohort Validation: Applying the GGMP-trained model directly to the SGMP test set yielded an AUC of 0.82 (95% CI: 0.79 - 0.85) and an accuracy of 0.76. Conversely, the SGMP-trained model achieved an AUC of 0.80 (95% CI: 0.77 - 0.83) in GGMP validation, demonstrating good bidirectional generalization ability.

Compared with the baseline model: The logistic regression baseline model based on age, sex, BMI and fasting blood glucose had an AUC of 0.76 (95% CI: 0.73 - 0.79) in GGMP, which was significantly lower than that of the microbiome model ($p < 0.001$), suggesting that the gut microbiome provides independent predictive information beyond traditional risk factors.

3.4. Microbial Biomarker Identification

By aggregating the SHAP values of 100 sub-models, 12 T2DM-related microbial genera with consistent contribution directions in both cohorts were identified (Figure 2):

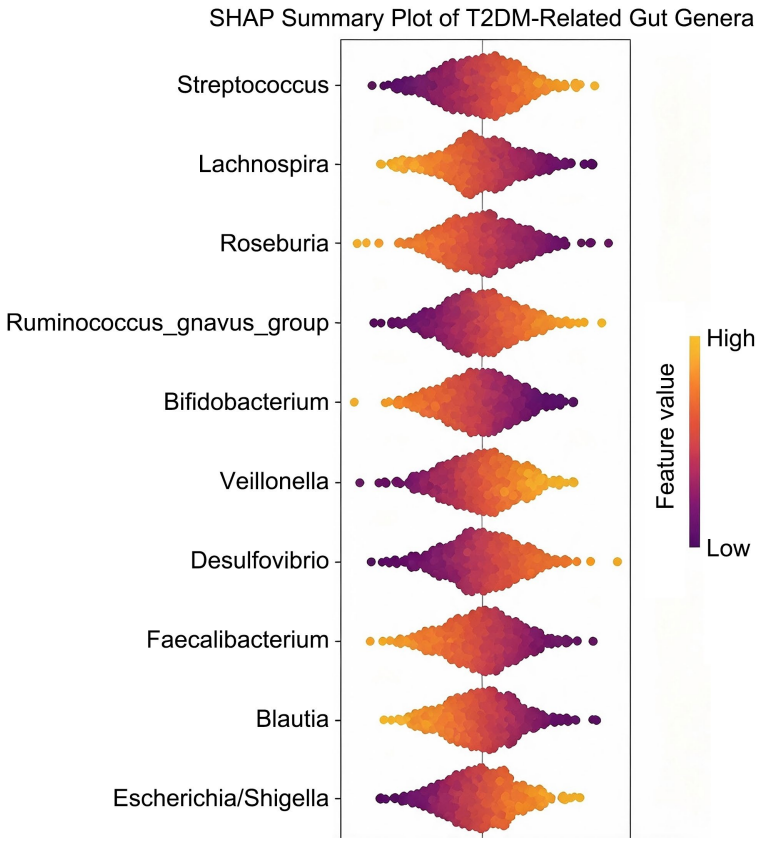

Figure 2. SHAP summary plot of gut genera associated with type 2 diabetes mellitus.

Protective Microorganisms (decreased abundance associated with increased T2DM risk):

- Faecalibacterium (mean |SHAP| = 0.32)
- Roseburia (mean |SHAP| = 0.28)
- Bifidobacterium (mean |SHAP| = 0.21)
- Lachnospira (mean |SHAP| = 0.18)
- Blautia (mean |SHAP| = 0.15)

Risk Microorganisms (increased abundance associated with increased T2DM risk):

- Escherichia/Shigella (mean |SHAP| = 0.25)
- Desulfovibrio (mean |SHAP| = 0.19)
- Ruminococcus_gnavus_group (mean |SHAP| = 0.16)
- Streptococcus (mean |SHAP| = 0.14)
- Veillonella (mean |SHAP| = 0.12)

These markers showed a high degree of agreement in their effects across the two cohorts (consistency > 85%). Faecalibacterium and Escherichia/Shigella have also been repeatedly reported as key differentially expressed genera in T2DM in previous studies.

3.5. Null Model Validation

To verify that the identified microbial biomarkers were not random noise, this study constructed 100 null models (with disease labels randomly shuffled). The mean AUC of the null models was 0.51 (±0.03), significantly lower than the 0.87 of the true models ($p < 0.001$). In the null models, no microbial genera achieved the contribution level of the top 5 biomarkers in the true models, confirming the statistical significance of the microbiome-disease associations identified by the models.

4. Discussion

4.1. Key Findings

This study constructed a T2DM prediction model based on XGBoost by integrating two large Chinese gut microbiome cohorts. Key findings include: 1) Gut microbiome-based machine learning models can effectively predict T2DM in cross-cohort scenarios (AUC 0.82 - 0.87), outperforming traditional risk factor models; 2) Twelve T2DM-related microbial genera were identified across cohorts, with decreased abundance of *Faecalibacterium* and *Roseburia* being the most stable predictive signal; 3) The combination of ensemble learning and SHAP interpretation effectively improved the model's interpretability and cross-cohort generalization ability.

4.2. Comparison with Existing Research

The model performance of this study (AUC 0.82 - 0.87) is comparable to or slightly better than similar studies recently published. A 2024 study used six ma-

chine learning models to predict six types of diseases (obesity, ulcerative colitis, autoimmune diseases, *Clostridium difficile* infection, irritable bowel syndrome, and colorectal cancer), with an average AUC ranging from 0.95 to 0.99 [4]. Another study on inflammatory bowel disease, the XGB-IBD10 model, achieved an accuracy of 0.8722 in the test sample and 0.8066 in the external validation cohort [5]. This study maintained an AUC > 0.80 predictive performance in a cross-cohort scenario (community-sourced GGMP vs. hospital-sourced SGMP), validating the model's good generalization ability.

Notably, the ensemble learning strategy employed in this study—constructing multiple sub-models from random subsamples—is consistent with the methodological approach of the aforementioned high-precision studies, namely, reducing the risk of overfitting from a single dataset and improving cross-cohort stability through multi-model ensemble. This strategy is universally applicable to addressing the high dimensionality and cohort heterogeneity of microbiome data.

The absence of butyric acid-producing bacteria was the most important predictive signal in this study, which is highly consistent with the pathophysiological mechanism of T2DM. Butyric acid, as a major energy source for intestinal epithelial cells, has multiple functions including anti-inflammation, maintaining intestinal barrier integrity, and regulating insulin sensitivity. Members of the genera *Faecalibacterium prausnitzii* and *Roseburia* are major butyrate producers in the human gut, and decreased abundance of these compounds may exacerbate insulin resistance by weakening the intestinal barrier and promoting endotoxin uptake into the bloodstream.

4.3. Clinical Significance

The results of this study have clear clinical application prospects. Traditional type 2 diabetes mellitus (T2DM) screening relies on fasting blood glucose and HbA1c testing, requiring blood samples and having limited sensitivity in the early stages of the disease. Predictive models based on fecal microbiome 16S rRNA sequencing offer a completely non-invasive screening approach—subjects only need to provide a stool sample, without blood collection or fasting. The cost of 16S rRNA sequencing has decreased to $50 - $100 per sample, and this cost is expected to continue to decline as sequencing technology becomes more widespread.

From a public health perspective, microbiome predictive models can be applied to large-scale community screening to identify high-risk populations and enable early intervention. Lifestyle interventions (dietary adjustments, increased dietary fiber intake) have been shown to modulate gut microbiota composition and improve insulin sensitivity. Combining microbiome screening with lifestyle interventions may form a closed-loop management model of "screening-early warning-intervention".

4.4. Limitations and Future Directions

This study has the following limitations. First, 16S rRNA sequencing only pro-

vides genus-level taxonomic resolution and cannot distinguish differences at the strain level. Heterogeneity at the strain level may affect the model's predictive accuracy; future studies could utilize shotgun metagenomics to deepen the analysis at the species and strain levels. Second, both cohorts are from Chinese populations, and the model's generalization ability in European and American populations remains to be verified. Cross-ethnic validation is an important direction for future research. Third, this study is a cross-sectional design and cannot determine the temporal relationship between microbiome changes and the occurrence of T2DM. Longitudinal cohort studies will help clarify causal relationships. Fourth, the model has not yet been validated for operational feasibility and cost-effectiveness in actual clinical scenarios.

Future research can be further explored in the following directions: 1) Integrating metagenomic functional data (such as metabolic pathways and antibiotic resistance genes) to improve the predictive accuracy and mechanistic explanation capabilities of the model; 2) Developing simplified detection schemes based on qPCR or targeted sequencing to lower the threshold for clinical application; 3) Verifying in randomized controlled trials whether early intervention guided by microbiome screening can improve the clinical outcomes of T2DM.

5. Conclusion

Based on two large 16S rRNA gut microbiome cohorts in China, this study constructed an interpretable XGBoost ensemble learning model, achieving effective cross-cohort prediction of T2DM (AUC 0.82 - 0.87). The model identified 12 cohort-consistent T2DM-related microbial genera, among which the decrease in abundance of *Butyric acid-producing* was the most stable predictive signal. The model outperformed traditional risk factor models, suggesting that the gut microbiome provides independent predictive information. This study provides a new technical tool for non-invasive screening of T2DM and a methodological reference for the application of the gut microbiome in the precision management of metabolic diseases.

AI-Assisted Tool Disclosure

Google Translate was used solely for English language editing and polishing in the preparation of this manuscript. The authors reviewed and verified all revisions and take full responsibility for the final content of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] International Diabetes Federation (2021) IDF Diabetes Atlas. 10th Edition, International Diabetes Federation, Brussels. https://diabetesatlas.org/atlas/tenth-edition/

[2] Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A.,

et al. (2019) Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2. *Nature Biotechnology*, **37**, 852-857. https://doi.org/10.1038/s41587-019-0209-9

[3] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., *et al.* (2012) The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools. *Nucleic Acids Research*, **41**, D590-D596. https://doi.org/10.1093/nar/gks1219

[4] Temel, B.E., Kocapinar, B. and Isik, Z. (2024) Disease Prediction from Human Microbiome by Utilizing Machine Learning. 2024 9*th International Conference on Computer Science and Engineering* (*UBMK*), Antalya, 26-28 October 2024, 683-688. https://doi.org/10.1109/ubmk63289.2024.10773434

[5] Yu, S., Li, J., Ye, Z., Zhang, M., Guo, X., Wang, X., *et al.* (2025) Identification of a 10-Species Microbial Signature of Inflammatory Bowel Disease by Machine Learning and External Validation. *Cell Regeneration*, **14**, 32. https://doi.org/10.1186/s13619-025-00246-w