Microbiome-Based Classification of Type 2 Diabetes Using Public Gut Metagenomic Data and Stacked Machine Learning Models

Margarido Fonseca. R1, Second B. Author2, and Third C. Author, Jr.3, Member, IEEE

1Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal.

Corresponding author: Margarido Fonseca. R (e-mail: a2019135871@isec.pt).

ABSTRACT Accurately classifying gut microbiome profiles offers a promising route for non-invasive detection of Type 2 Diabetes (T2D). This study explores whether a machine learning pipeline can reliably distinguish between healthy and diabetic individuals using species-level microbiome data. Using 145 curated samples from the Data repository for human gut microbiota (GMRepo) PRJEB1786 dataset (92 healthy, 53 diabetic), I constructed a classification pipeline that integrates mutual information-based feature selection, SMOTE resampling, and standardized abundance features. A stacking classifier combining optimized Support Vector Machine, Random Forest, and Logistic Regression models achieved 92% test accuracy. However, when the model was applied to an extended dataset incorporating additional samples from project PRJNA422434, accuracy dropped to 62%, indicating poor generalizability across cohorts. These findings suggest that while high classification performance is achievable within a controlled dataset, broader deployment will require improved cross-cohort normalization and robust model validation strategies.

INDEX TERMS Gut Microbiome, Machine Learning, Type 2 Diabetes (T2D).

1. INTRODUCTION

Type 2 Diabetes (T2D) is a chronic metabolic disorder that continues to rise at an alarming rate. As of 2024, an estimated 589 million adults (aged 20-79) are living with diabetes globally, with projections to reach 853 million by 2050 [1]. Additionally, over **252 million people** are believed to be living with **undiagnosed diabetes**, underscoring the urgent need for improved detection strategies [2]. T2D significantly increases the risk of cardiovascular disease, kidney failure, and neuropathy, making early diagnosis critical to reduce complications and long-term healthcare burden. Traditionally attributed to genetic, dietary, and lifestyle factors, growing evidence suggests that the gut microbiome—an ecosystem of trillions of microorganisms residing in the gastrointestinal tract—plays a significant role in metabolic regulation and disease progression [3]. Disruptions in microbial composition, or dysbiosis, have been linked to insulin resistance, chronic inflammation, and altered glucose metabolism.

With advances in metagenomic sequencing, researchers now have access to species-level microbiome data, which presents a new opportunity for non-invasive diagnostic modeling. However, extracting clinically relevant patterns from this data remains challenging due to its high dimensionality, sparsity, and inter-individual variability. Machine learning offers a promising approach to tackle these challenges, enabling pattern discovery and disease classification at scale.

This study investigates the use of machine learning pipelines, including ensemble models, to classify individuals as healthy or diabetic based on their gut microbiome profiles. The primary aim was to optimize predictive performance using publicly available metagenomic data. A species-level dataset (PRJEB1786) from GMRepo was used as the primary source, with preprocessing steps including abundance normalization, metadata integration, and univariate feature selection based on mutual information scores.

The resulting classification pipeline—built around a stacking ensemble of Random Forest, Support Vector Machine, and Logistic Regression models—achieved 92% test accuracy on the internal dataset. However, when samples from a second cohort (PRJNA422434) were introduced, model accuracy dropped to 62%, revealing limitations in cross-cohort generalization. These findings highlight both the promise and current constraints of microbiome-based disease prediction, and motivate future work focused on model robustness, cohort harmonization, and biological validation.

PROBLEM DESCRIPTION

Despite mounting evidence linking gut microbiome composition to Type 2 Diabetes (T2D), leveraging this information for clinical prediction remains a complex task. Microbiome data is inherently high-dimensional, sparse, and biologically noisy, making it difficult to extract stable and generalizable patterns using conventional statistical methods. Additionally, microbial populations vary significantly across individuals due to factors such as geography, diet, age, and sequencing protocol—further complicating classification efforts.

While previous studies have identified associations between specific microbial taxa and metabolic disorders, most do not provide access to the underlying datasets, limiting reproducibility and independent evaluation. Among the few datasets that are publicly available, many lack critical metadata—such as host age, geography, or health context—that could improve model performance or interpretability. Additionally, the classification accuracy or predictive validity of published models is often unclear or unreported, making it difficult to assess their real-world applicability or compare approaches. This lack of transparency presents a significant obstacle for advancing microbiome-based diagnostics.

The central problem this work addresses is whether machine learning pipelines can reliably classify individuals as diabetic or healthy based solely on microbiome species abundance, potentially supplemented by metadata. The goal was to not only maximize predictive accuracy but also assess the extent to which such models can generalize to unseen cohorts. In doing so, this study aims to bridge the gap between microbiome association studies and practical, model-driven diagnostics.

RELATED WORK

The human gut microbiome has been widely implicated in metabolic health and disease. Numerous studies have shown that microbial dysbiosis—disruptions in gut microbial composition—can contribute to insulin resistance, systemic inflammation, and abnormal glucose metabolism, all of which are hallmarks of Type 2 Diabetes (T2D).

Dash et al. applied machine learning to identify microbial taxa and metabolic pathways associated with Type 2 Diabetes (T2D) using nanopore sequencing. While they reported distinct differences in microbial composition between diabetic and healthy individuals, the study did not include standard performance metrics (e.g., accuracy, F1-score), and the underlying datasets were not publicly shared, limiting the possibility of external validation.

Młynarska et al. provide a descriptive review of the role of gut microbiota in the development of Type 2 Diabetes, focusing on mechanisms such as the production of short-chain fatty acids (SCFAs)—molecules produced by gut bacteria when they break down dietary fiber, which help regulate metabolism and inflammation—along with immune system interactions and the integrity of the gut barrier. While informative, their review does not engage with computational tools or predictive modeling techniques such as machine learning. While informative, their review does not engage with computational tools or predictive modeling techniques such as machine learning. In contrast, microbiome-focused ML research has explored classification using models like Random Forests, support vector machines (SVMs), and neural networks. However, these efforts often prioritize biomarker discovery over robust, generalizable classification pipelines, limiting real-world applicability due to a lack of validation across external datasets.

The GMRepo database provides a curated, publicly accessible resource for human gut metagenomic data with phenotype labels, making it an ideal candidate for reproducible ML applications. However, its use in published ML-based disease classification studies remains limited. Most studies leveraging GMRepo focus on compositional analysis or biomarker enrichment rather than supervised learning tasks with quantifiable evaluation metrics.

Of relevance is project PRJEB1786 [6], the dataset used in this study. According to its description, the original researchers developed a mathematical model to classify T2D status based on microbial composition among European women. While they claim the model achieved “high accuracy,” no specific performance metrics are reported, and neither the code nor a reproducible pipeline is available. Furthermore, the project mentions that their model performed poorly when applied to a Chinese cohort, noting that geographic differences in microbiomes affect prediction. However, no dataset ID or quantitative performance details are provided.

In this study, the stacking classifier was trained on PRJEB1786 and then tested on a second GMRepo dataset, PRJNA422434, which also originates from a Chinese population. Although it cannot be confirmed whether this is the same cohort referenced in the original study, the observed drop in accuracy—from 92% to 62%—mirrors their claim of poor generalization. Unlike the original work, this study explicitly reports performance metrics, documents the full classification pipeline, and tests generalization using a second publicly available cohort.

In doing so, this work addresses key gaps in the literature by:

* Using fully open and curated microbiome data
* Building a reproducible, performance-optimized ML pipeline
* Evaluating internal classification accuracy
* Testing cross-cohort generalization with quantitative results

WORK CONDUCTED

This section outlines the full methodology used to build a high-performing classification model for predicting diabetes status based on gut microbiome composition. The process included dataset acquisition and correction, species-level abundance normalization, metadata integration, feature selection, oversampling for class balance, model training, and evaluation. Special emphasis was placed on optimizing predictive performance through ensemble learning and assessing generalizability to external cohorts.

1. DATA ACQUISITION

Data for this study was obtained from theGMRepodatabase **(**[**https://gmrepo.humangut.info/home**](https://gmrepo.humangut.info/home)**).** The sample search interface was used to filter for entries where the phenotype was set to **"Diabetes Mellitus, Type 2"**. Although the interface allows for further filtering by host metadata such as age and BMI, these filters were removed to maximize sample inclusion. The correct approach involved deleting any automatically added **"AND/OR"** groups that narrow the query logic.

This query returned **424 sequencing runs**, corresponding to **two different projects**. For this study, data from project **PRJEB1786** was selected. Within the project page, the relevant entries were accessed by scrolling to the “Samples” section, which contains curated data. Downloads were made in the species-level format, as genus-level annotations were available but deemed less informative for classification tasks.

1. EXPLORATORY DATA ANALYSIS

After data acquisition, each sample was downloaded as an individual tab-delimited file containing species-level relative abundance values, along with taxonomic identifiers and minimal metadata. An excerpt from one such file is shown in Figure 1..

A screenshot of a computer

AI-generated content may be incorrect.

(legenda ou anexo de figuras)

The files **did not include structured metadata** such as phenotype, age, sex, BMI, or geographic location. While this information is available on the GMRepo website, it must be **manually retrieved and persisted** if needed for downstream analysis. In the early stages of the study, the focus was placed exclusively on the species-level data.

To enable group-level inspection, the sample files were first divided by phenotype and then **merged into two aggregate tables**: one for all healthy samples and another for all diabetic samples. This bundling facilitated an initial comparative analysis of microbial composition between the two groups.

A colorful pie chart with text

AI-generated content may be incorrect.

(healthy, talvez colocar também a imagem das espécies mais comuns? Estas 2 imagens em anexo?)

A colorful pie chart with numbers

AI-generated content may be incorrect.

(diabetic, talvez colocar também a imagem das espécies mais comuns? Estas 2 imagens em anexo?)

Species abundance and prevalence were explored by calculating the **total relative abundance per species** across all samples in each group, as well as the **number of samples** in which each species appeared. This distinction revealed that certain species—such as *Faecalibacterium prausnitzii* and *Eubacterium rectale*—were both **highly abundant** and **widely prevalent** in both groups. However, subtle shifts in composition were observed between healthy and diabetic samples.

In addition, many entries were labeled as **"Unknown" species** or assigned invalid identifiers such as **“-1”**. These entries were present in both groups and, although abundant, lacked biological interpretability. For this reason, they were excluded from all subsequent stages of analysis. The impact of this filtering step is illustrated in **Figure 2**, which shows the difference in species distribution before and after removal.

A colorful pie chart with numbers

AI-generated content may be incorrect.

(healthy)

A colorful pie chart with text

AI-generated content may be incorrect.

(diabetic)

The insights gathered during this exploratory phase helped guide downstream decisions regarding feature selection, normalization, and model design.

1. FEATURE ENGINEERING AND DATASET STRUCTURING

Following initial data exploration, the cleaned healthy and diabetic sample tables were merged into a single dataset. Each row represented a unique sample, and each column corresponded to the **relative abundance of a single microbial species**, identified by its ncbi\_taxon\_id. An additional binary target column labeled samples as **1 for healthy** and **0 for diabetic** to enable supervised classification.

A screen shot of a white grid

AI-generated content may be incorrect.

To incorporate contextual information, **hostage** and **country of origin** were manually retrieved from the GMRepo website and persisted for each sample. All participants in this dataset were **female**, and **BMI data was not available**. The collected metadata was preprocessed and encoded appropriately: **country** was one-hot encoded, while **age** was treated as a continuous numerical feature and standardized along with microbial abundance values. These metadata features were appended to the final feature matrix, providing the model with potential covariates that could capture geographic or demographic variation in microbial composition.

The resulting dataset contained **372 features** in total, including microbial species and encoded metadata. It was both **high-dimensional** and **sparse**, as many microbial species were only present in a subset of samples. To reduce noise and highlight the most informative features, **univariate feature selection** was applied using SelectKBest with **mutual information** as the scoring function. The top **200 features** most predictive of the health label were retained. These included both species-level features and metadata columns. **Recursive Feature Elimination (RFE)** was also tested as an alternative method and, coincidentally, selected the same number of features. However, it did not lead to improved model performance and was therefore not included in the final pipeline.

All numerical features—microbial abundance values and age—were then standardized using StandardScaler to ensure uniform scaling. The dataset was subsequently split into training and test subsets using an **80/20 stratified split**, preserving the original ratio of healthy to diabetic samples.

Given the moderate class imbalance (92 healthy vs. 53 diabetic), **SMOTE (Synthetic Minority Over-sampling Technique)** was applied to the training set. This method generated synthetic samples for the minority class (diabetic) to balance the class distribution, reducing the risk of biased learning and improving model performance.

At this stage, the dataset was fully structured, feature-selected, scaled, and balanced—ready for model development and training.

1. MODEL DEVELOPMENT AND OPTIMIZATION

With the final dataset prepared, several machine learning models were developed to classify individuals as healthy or diabetic based on their gut microbiome profiles. The objective was to maximize predictive accuracy through robust feature representation, balanced training data, and optimized model configuration.

Three supervised classifiers were selected as base learners due to their complementary strengths and interpretability: Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR). Each model underwent hyperparameter tuning using GridSearchCV, with F1-score as the evaluation metric and 3-fold cross-validation to ensure robustness across training splits.

For the Random Forest classifier, the grid search explored parameters such as the **number of estimators**, **maximum tree depth**, **minimum samples per split**, and **minimum samples per leaf**. The Support Vector Machine was tuned by varying the **regularization parameter C**, **kernel type** (linear and radial basis function), and **gamma**. In the case of Logistic Regression, tuning focused on the **regularization strength** and the **solver** (using lbfgs) to ensure stable convergence on the standardized feature space.

Once the best configurations were selected, these models were integrated into a **stacking ensemble classifier**, using Logistic Regression as the **meta-learner**. The stacking approach was chosen to leverage the diversity of the base learners—combining the decision boundaries of SVM, the ensemble power of Random Forest, and the linear separability of Logistic Regression.

The final model was constructed using StackingClassifier with passthrough=True, allowing the meta-model to access both the raw input features and the predictions of the base models. The ensemble was trained on the **SMOTE-balanced** training set to mitigate class imbalance.

This configuration achieved the best overall performance during experimentation, outperforming individual models and simpler ensemble methods such as majority voting. The optimized stacking pipeline was then evaluated on the held-out test set, as described in the next section.

1. PERFORMANCE EVALUATION

The final stacking classifier was evaluated on the held-out test set using a variety of standard metrics: **accuracy**, **precision**, **recall**, and **F1-score**. The model achieved an accuracy of **92%**, marking a substantial improvement over the individual base models, which typically ranged from **61% to 75%**. The ensemble's strong performance was attributed to the complementary strengths of its components, as well as the careful feature selection, metadata integration, and class balancing steps applied earlier in the pipeline.

The performance metrics for each individual model and the final ensemble are summarized in **Table X**.

(Colocar tabela com valores da acc e todas as “iterações” que fiz até chegar a 92% de accuracy)

In addition to internal evaluation, the model's generalizability was tested by introducing additional microbiome samples from a separate GMRepo project, **PRJNA422434**. These samples were processed and formatted in the same way as the original dataset to ensure compatibility. However, when the trained classifier was applied to this extended dataset, **performance dropped to 62% accuracy**. This result highlights a key limitation: while the model performs well within its original domain, it does not generalize reliably to unseen cohorts.

The drop in performance is likely due to **cohort-specific effects**, such as variations in sample preparation, sequencing methods, or population-level microbiome differences. This emphasizes the need for **cross-cohort normalization strategies**, **domain adaptation**, or **larger and more diverse training datasets** in future work.

Despite this limitation, the high internal accuracy demonstrates that gut microbiome composition, when properly processed and modeled, holds substantial promise for classifying Type 2 Diabetes status in controlled settings.

FUTURE WORK

The While the stacking classifier developed in this study achieved high predictive accuracy on the internal dataset, several important directions remain for improving model robustness, generalizability, and biological relevance.

First, future work should focus on enhancing **cross-cohort generalization**. The significant performance drop observed when applying the model to an external dataset (PRJNA422434) suggests that microbiome-based classifiers are sensitive to cohort-specific factors such as geographic diversity, sequencing methods, or environmental exposures. Approaches such as **domain adaptation**, **batch effect correction**, and **multi-cohort training** could help mitigate this issue and improve transferability across datasets.

Second, the integration of additional **metadata** could improve model performance and interpretability. While this study incorporated host age and country, other factors such as diet, medication, lifestyle, and medical history could further explain variation in microbial composition and enhance predictive power. However, these variables are often inconsistently reported or entirely missing from public datasets, underscoring the need for better-curated and metadata-rich microbiome repositories.

Third, the current study focused on **species-level taxonomic profiles**, but functional information—such as gene content, metabolic pathway activity, or strain-level variation—could provide more direct insights into host-microbiome interactions. Incorporating such **functional metagenomics data** may improve biological interpretability and enhance classification accuracy.

Additionally, while the current model used traditional machine learning techniques, future work could explore **deep learning** or **generative approaches**, such as **variational autoencoders (VAEs)** or **generative adversarial networks (GANs)**, for learning compact representations or augmenting limited datasets.

Finally, a natural next step involves applying the model to **larger, more diverse cohorts**, possibly in clinical contexts, to assess real-world utility. Validation against longitudinal datasets could also help evaluate the model's ability to detect early microbiome shifts that precede clinical onset of Type 2 Diabetes.

CONCLUSION

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

REFERENCES

[1] “IDF Diabetes Atlas | Global Diabetes Data & Statistics.” Accessed: Apr. 26, 2025. [Online]. Available: https://diabetesatlas.org/

[2] “million”.

[3] J. Wang *et al.*, “A metagenome-wide association study of gut microbiota in type 2 diabetes,” *Nature 2012 490:7418*, vol. 490, no. 7418, pp. 55–60, Sep. 2012, doi: 10.1038/nature11450.

\\192.168.0.8\Design2\Indesign Projects\005 Series\03 OA Word templates\Work\Access-Template\Images\Fig1.tifFIRST A. AUTHOR (M’76–SM’81–F’87) and all authors may include biographies. Biographies are often not included in conference-related papers. This author became a Member (M) of IEEE in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author’s educational background is listed. The degrees should be listed with type of degree in what field, which institution, city, state, and country, and year the degree was earned. The author’s major field of study should be lower-cased.

The second paragraph uses the pronoun of the person (he or she) and not the author’s last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (publisher name, year) similar to a reference. Current and previous research interests end the paragraph.

The third paragraph begins with the author’s title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter). List any memberships in professional societies other than the IEEE. Finally, list any awards and work for IEEE committees and publications. If a photograph is provided, it should be of good quality, and professional-looking. Following are two examples of an author’s biography.

**\\192.168.0.8\Design2\Indesign Projects\005 Series\03 OA Word templates\Work\Access-Template\Fig-1.tifSECOND B. AUTHOR** was born in Greenwich Village, New York, NY, USA in 1977. He received the B.S. and M.S. degrees in aerospace engineering from the University of Virginia, Charlottesville, in 2001 and the Ph.D. degree in mechanical engineering from Drexel University, Philadelphia, PA, in 2008.

From 2001 to 2004, he was a Research Assistant with the Princeton Plasma Physics Laboratory. Since 2009, he has been an Assistant Professor with the Mechanical Engineering Department, Texas A&M University, College Station. He is the author of three books, more than 150 articles, and more than 70 inventions. His research interests include high-pressure and high-density nonthermal plasma discharge processes and applications, microscale plasma discharges, discharges in liquids, spectroscopic diagnostics, plasma propulsion, and innovation plasma applications. He is an Associate Editor of the journal *Earth*, *Moon*, *Planets*, and holds two patents.

Dr. Author was a recipient of the International Association of Geomagnetism and Aeronomy Young Scientist Award for Excellence in 2008, and the IEEE Electromagnetic Compatibility Society Best Symposium Paper Award in 2011.