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

Margarido Fonseca. R1, Pereira C.1,2

1Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal. Superior Institute of Engeneering of Coimbra, Rua Pedro Nunes, Alto de São João, 3030-199, Coimbra, Portugal

2Centre for Informatics and Systems of the University of Coimbra, Rua Miguel Bombarda, 3030-790, 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. A total of 145 curated samples (92 healthy, 53 diabetic) from the GMRepo project PRJEB1786 were used to construct a machine learning pipeline incorporating 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 64%, 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 with an accelerating global prevalence. As of 2024, an estimated 589 million adults (aged 20–79) were affected worldwide, with projections reaching 853 million by 2050. [1]. Additionally, over **252 million people** were 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 [3], [4]. 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[5]. Disruptions in microbial composition, or dysbiosis, have been linked to insulin resistance, chronic inflammation, and altered glucose metabolism [5].

With advances in metagenomic sequencing, researchers now have access to species-level microbiome data, which presents a new opportunity for non-invasive diagnostic modeling [6]. However, extracting clinically relevant patterns from this data remains challenging due to its high dimensionality, sparsity, and inter-individual variability [7]. Machine learning offers a promising approach to tackle these challenges, enabling pattern discovery and disease classification at scale [7].

This study evaluates machine learning pipelines—particularly ensemble models—for classifying individuals as healthy or diabetic based on their gut microbiome profiles. The objective was to maximize predictive performance using publicly available metagenomic data. A species-level dataset (PRJEB1786) from GMRepo served as the primary source, with preprocessing steps that included abundance normalization, metadata integration, and univariate feature selection based on mutual information scores [8], [9].

The resulting classification pipeline—built around a stacking ensemble of Random Forest, Support Vector Machine, and Logistic Regression models—achieved 92% accuracy on the internal dataset. However, when samples from a second cohort (PRJNA422434) were introduced, model accuracy dropped to 64%, 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 [10], [11]. Additionally, microbial populations vary significantly across individuals due to factors such as geography, diet, age, and sequencing protocol—further complicating classification efforts [10], [11].

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 [10], [11]. 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.

This study investigates whether machine learning pipelines can reliably classify individuals as diabetic or healthy based on microbiome species abundance, supplemented by host metadata including age and country of origin. In addition to optimizing predictive accuracy, the study evaluates how well these models generalize to independent cohorts, aiming to bridge the gap between microbiome association research and practical diagnostic applications.

RELATED WORK

The gut microbiome has been increasingly recognized as a key player in metabolic health, with numerous studies highlighting its role in the onset and progression of Type 2 Diabetes (T2D). Disruptions in microbial composition—commonly referred to as dysbiosis—have been associated with insulin resistance, systemic inflammation, and altered glucose metabolism.[12]. Młynarska et al. provide a comprehensive review of the mechanisms through which gut microbes may influence T2D, including short-chain fatty acid production, gut barrier integrity, and immune modulation. However, such reviews often stop short of proposing computational models capable of translating microbiome variation into diagnostic tools [13].

In more targeted studies, Dash et al. employed nanopore sequencing to characterize microbial differences between diabetic and healthy individuals. While their analysis identified taxa potentially linked to disease status, the study did not provide access to its dataset nor report standard machine learning performance metrics such as accuracy or F1-score, limiting its reproducibility and practical relevance[12].

The GMRepo database has recently emerged as a curated resource for microbiome-related phenotype data, offering standardized abundance profiles across projects. Despite its accessibility and consistency, only a limited number of studies have leveraged GMRepo for machine learning classification tasks. Many of these works have focused on exploratory taxonomic profiling or differential abundance testing, without advancing predictive modeling frameworks [8], [9].

One notable exception is the mAML framework introduced by Wu et al. which combines a curated microbiome dataset with an automated machine learning pipeline to streamline disease classification tasks. While mAML provides an important step toward reproducibility and automation in microbiome ML studies, it does not focus specifically on GMRepo datasets or assess generalizability across cohorts [14]. Similarly, a 2024 study by Peng et al. applied machine learning algorithms to identify microbial biomarkers associated with obesity, highlighting *Bifidobacterium pseudocatenulatum* as a candidate for therapeutic intervention [15]. While this work demonstrates the potential of microbiome-informed ML pipelines for metabolic diseases, it does not evaluate cross-dataset generalization or use public resources like GMRepo, which limits its reproducibility.

A notable contribution addressing these concerns is the study by **Li et al.** which systematically evaluated the cross-cohort performance of microbiome-based classifiers across 20 different diseases [16]. Their work highlights a common limitation in microbiome ML research: models that perform well within a cohort often fail to generalize when applied to external datasets. Despite using advanced classifiers and large sample sizes, the study found substantial accuracy drops during cross-cohort validation—reinforcing the need for more robust pipelines that account for population and technical variability. However, the classifiers in that study were not explicitly trained on GMRepo datasets, nor were feature selection strategies or ensemble methods as thoroughly optimized and documented as in the present work.

Project PRJEB1786—used as the primary dataset in this study—is listed in the GMRepo database with a brief project description suggesting that microbial features were used to build a mathematical model for classifying T2D status among European women. However, no peer-reviewed publication or publicly available analysis appears to be associated with this project. While the description claims that the model achieved "high accuracy" on the internal dataset and performed poorly when applied to a Chinese cohort, no quantitative performance metrics, external validation data, or source code are provided. As a result, the project offers limited utility for benchmarking or comparative analysis and has not been substantively explored in existing machine learning studies.

This study builds upon the availability of PRJEB1786 by not only utilizing it as the primary dataset, but also expanding the analysis to include PRJNA422434, both of which are publicly accessible through GMRepo. Unlike prior references that offer limited methodological transparency, the present work introduces a fully reproducible machine learning pipeline. Feature selection, class balancing, and ensemble learning techniques are systematically applied, and all preprocessing steps, model parameters, and performance metrics, including cross-cohort generalization results are explicitly documented.

In doing so, this work addresses several key limitations in the existing literature by leveraging fully open and curated microbiome datasets, developing a reproducible and performance-optimized machine learning pipeline, and rigorously evaluating classification accuracy within and across cohorts. By reporting complete preprocessing steps, feature selection strategies, and ensemble configurations, this study promotes transparency and reusability. To our knowledge, it is among the first to benchmark ensemble-based microbiome classifiers across multiple GMRepo datasets while quantitatively assessing generalizability.

WORK CONDUCTED

This section outlines the full methodology used to develop and evaluate a machine learning pipeline for T2D classification 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 entries with the phenotype being set to **“equal** **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 downwards, 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

Following data acquisition, an exploratory analysis was conducted to examine the structure and distribution of the gut microbiome data. This step aimed to assess feature completeness, common microbial patterns across health labels, and guide subsequent preprocessing decisions. Each sample was provided as a tab-delimited file containing species-level relative abundance values and associated taxonomic identifiers such as the one shown in the picture below.

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, sample files were first divided by phenotype (healthy vs. diabetic) facilitating a comparative analysis of microbial composition. Species-level abundance and prevalence were computed by summing relative abundance across samples and counting the number of samples each species appeared in.

A colorful pie chart with text

AI-generated content may be incorrect.

Figure 1(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.

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

As illustrated in **Figure 1**, *Faecalibacterium prausnitzii* and *Eubacterium rectale* were both highly abundant and prevalent in healthy samples. However, subtle compositional shifts are apparent in diabetic samples (**Figure 2**), where *Ruminococcus bromii* and unknown species occupy a greater proportion of the microbial community. (deixo isto?)

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 3 and 4**, which shows the difference in species distribution after removing the beforementioned ambiguous entries. Post-filtering, biologically meaningful taxa become more prominent, offering clearer insights for model development and feature selection.

A colorful pie chart with numbers

AI-generated content may be incorrect.

Figure 3(healthy)

A colorful pie chart with text

AI-generated content may be incorrect.

Figure 4(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

After initial exploration, the healthy and diabetic samples were consolidated into a single dataset. Each row represented an individual sample, while columns captured the relative abundances of specific microbial species, indexed by their corresponding NCBI taxon IDs. A binary target variable was added to indicate phenotype, with 1 denoting healthy and 0 diabetic cases, enabling a supervised classification framework.

A screen shot of a white grid

AI-generated content may be incorrect.

To enrich the dataset with contextual information, **host age** and **location** were manually retrieved from the GMRepo database and persisted for each sample. In this specific project, all participants were female, and body mass index (BMI) data was unavailable, so it was excluded from the feature set. Country of origin was one-hot encoded, while age was treated as a continuous numerical feature and standardized. These two metadata variables were appended to the final feature matrix to provide potential covariates that could capture geographic or age-related 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 (including both species-level features and metadata columns). **Recursive Feature Elimination (RFE)** was also tested as an alternative method to SelectKBest 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.

A detailed comparison of **SelectKBest** performance across multiple scoring functions and K values is provided in Appendix A. Similarly, **RFE accuracy using different base estimators** is summarized in Appendix B.

Finally, all numerical features were standardized using StandardScaler. At this stage, **SMOTE** was applied to the **entire dataset** prior to the train-test split, synthetically augmenting the minority class (diabetic) to mitigate imbalance. The final balanced dataset was then split into training and test sets using an 80/20 stratified split. 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 dataset fully processed, a set of supervised machine learning models was trained to classify samples based on microbial and metadata features. This phase focused on maximizing predictive performance through careful model selection, hyperparameter tuning, and ensemble integration strategies.

(só tem as últimas iterações, não tem uma descrição detalhada de todas as iterações e as acc das mesmas)

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, the models were integrated into a stacking ensemble classifier with Logistic Regression as the meta-learner. This approach was designed to leverage the complementary strengths of the base learners—combining the decision boundaries of the Support Vector Machine, the ensemble power of Random Forest, and the linear separability of Logistic Regression. The final model was implemented using StackingClassifier with passthrough=True, enabling the meta-model to access both raw input features and the predictions of the base models.

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%** under default configurations.

To clearly present this progression, performance results are split into two tables.

Table X reports the accuracy of all individual models across three stages: default configuration, after SMOTE balancing, and after hyperparameter tuning.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Default Accuracy (%) | SMOTE balancing with default accuracy | Hyperparameter Tunning |
| Random Forest | 75 | 73 | 81 |
| Support Vector Machine | 68 | 84 | 86 |
| Logistic Regression | 61 | 81 | 78 |
| Naïve Bayes | 64 | 76 | 76 |
| K-Nearest Neighbors | 61 | 49 | 51 |
| CNN (64,64,128) | 61 | 78 | 78 |
| MLP (128,128) | 68 | 78 | 68 |

Table Y reports the accuracy obtained by various ensemble classifiers under different feature selection strategies. Specifically, it compares the performance of voting and stacking ensembles using features selected via Recursive Feature Elimination (RFE), SelectKBest with mutual information, and their intersection. The results underscore the sensitivity of ensemble classifiers to the input feature space, with the stacking ensemble combined with mutual information-based SelectKBest yielding the highest accuracy (92%).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Default Accuracy | Intersection of RFE and SelectKBest | Only RFE | Only SelectKBest (mutual\_info\_classif, k=200) |
| Voting Classifier (svm, rf, lr), | 84 | - | - | - |
| Voting Classifier (svm, rf, lr), Soft voting and with [3,1,1] weight | 86 | - | - | 89 |
| Stacking Classifier (svm, rf, lr), Logistic Regression as final estimator | 89 | 86 | 89 | 92 |

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.

(Caso necessário posso colocar imagem do dataset)

However, when the trained classifier was applied to this extended dataset, **performance dropped to 64% accuracy**. This result highlights a key model limitation: while it 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

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 the enlarged dataset (**PRJEB1786** + 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 host location, 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

This study demonstrates that machine learning pipelines, particularly stacking ensemble methods, can effectively classify T2D status based on gut microbiome composition. By leveraging curated metagenomic data from GMRepo and incorporating mutual information-based feature selection, SMOTE balancing, and hyperparameter-tuned base learners, the final stacked classifier achieved 92% accuracy on internal data—outperforming both individual models and simpler ensemble approaches.

Nonetheless, the substantial accuracy drop observed when applying the model to a second cohort (64%) underscores a central challenge in microbiome-based diagnostics: limited cross-cohort generalizability. This finding highlights the importance of addressing cohort heterogeneity through domain adaptation, multi-cohort training, or harmonization techniques. Additionally, the results suggest that feature selection strategies significantly affect ensemble model performance, with mutual information-based methods yielding superior results in this context.

Overall, this work contributes a reproducible framework for microbiome-based classification and lays the groundwork for future efforts focused on clinical validation, feature interpretability, and real-world deployment of microbiome-aware diagnostics for metabolic disease.

APPENDIX A

SELECTKBEST PERFORMANCE

|  |  |  |
| --- | --- | --- |
| SelectKBest | score\_func=f\_classif | score\_func=mutual\_info\_classif |
| K = 100 | 70% accuracy | 84% accuracy |
| K = 150 | 86% accuracy | 84% accuracy |
| K = 200 | 78% accuracy | 89% accuracy |
| K = 250 | - | 86% accuracy |

APPENDIX B

RFE PERFORMANCE

|  |  |  |
| --- | --- | --- |
| RFE | N\_features to select | N\_features to select = 200 |
| Default | 200 | - |
| Random Forest | - | 84% accuracy |
| Logistic Regression | - | 84% accuracy |
| SVM | - | 89% accuracy |

REFERENCES

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

[2] “million”.

[3] “Early detection of diabetic nephropathy in patient with type 2 diabetes mellitus: A review of the literature Sopida Thipsawat”, doi: 10.1177/14791641211058856.

[4] A. D. Deshpande, M. Harris-Hayes, and M. Schootman, “Epidemiology of Diabetes and Diabetes-Related Complications,” *Phys Ther*, vol. 88, no. 11, p. 1254, Nov. 2008, doi: 10.2522/PTJ.20080020.

[5] 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.

[6] A. Almeida *et al.*, “A new genomic blueprint of the human gut microbiota,” *Nature 2019 568:7753*, vol. 568, no. 7753, pp. 499–504, Feb. 2019, doi: 10.1038/s41586-019-0965-1.

[7] M. Ahsan, S. A. Luna, and Z. Siddique, “Machine-Learning-Based Disease Diagnosis: A Comprehensive Review,” 2022, doi: 10.3390/healthcare10030541.

[8] S. Wu *et al.*, “GMrepo: a database of curated and consistently annotated human gut metagenomes,” *Nucleic Acids Res*, vol. 48, no. D1, pp. D545–D553, Jan. 2020, doi: 10.1093/NAR/GKZ764.

[9] D. Dai *et al.*, “GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison,” *Nucleic Acids Res*, vol. 50, no. D1, pp. D777–D784, Jan. 2022, doi: 10.1093/NAR/GKAB1019.

[10] P. Przymus *et al.*, “Deep learning in microbiome analysis: a comprehensive review of neural network models,” *Front Microbiol*, vol. 15, p. 1516667, Jan. 2024, doi: 10.3389/FMICB.2024.1516667/XML/NLM.

[11] R. Zhou, S. K. Ng, J. J. Y. Sung, W. W. Bin Goh, and S. H. Wong, “Data pre-processing for analyzing microbiome data – A mini review,” *Comput Struct Biotechnol J*, vol. 21, p. 4804, Jan. 2023, doi: 10.1016/J.CSBJ.2023.10.001.

[12] N. R. Dash *et al.*, “Functional alterations and predictive capacity of gut microbiome in type 2 diabetes,” *Scientific Reports 2023 13:1*, vol. 13, no. 1, pp. 1–12, Dec. 2023, doi: 10.1038/s41598-023-49679-w.

[13] “(PDF) Exploring the Significance of Gut Microbiota in Diabetes Pathogenesis and Management—A Narrative Review.” Accessed: Apr. 30, 2025. [Online]. Available: https://www.researchgate.net/publication/381527572\_Exploring\_the\_Significance\_of\_Gut\_Microbiota\_in\_Diabetes\_Pathogenesis\_and\_Management-A\_Narrative\_Review

[14] F. Yang and Q. Zou, “mAML: an automated machine learning pipeline with a microbiome repository for human disease classification,” *Database*, vol. 2020, Jan. 2020, doi: 10.1093/DATABASE/BAAA050.

[15] H. Wu *et al.*, “Machine learning prediction of obesity-associated gut microbiota: identifying Bifidobacterium pseudocatenulatum as a potential therapeutic target,” *Front Microbiol*, vol. 15, p. 1488656, Feb. 2024, doi: 10.3389/FMICB.2024.1488656/BIBTEX.

[16] M. Li *et al.*, “Performance of Gut Microbiome as an Independent Diagnostic Tool for 20 Diseases: Cross-Cohort Validation of Machine-Learning Classifiers,” 2023, doi: 10.1080/19490976.2023.2205386.

\\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.