Gut Microbiome Based Health Classification using Machine Learning

MINI PROJECT REPORT

Submitted By

G. Kalyani Krishna
22011102017

Semester VI

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING
(INTERNET OF THINGS)

DEPARTMENT OF COMPUTER SCIENCE
AND ENGINEERING
SHIV NADAR UNIVERSITY CHENNAI

APRIL 2025

SHIV NADAR
UNIVERSITY
CHENNAI

Abstract

The gut microbiome plays a vital role in maintaining human health, with alterations in its composition often associated with a wide range of diseases. This project presents an integrated analytical and interactive framework for gut health assessment, leveraging statistical modeling and user-driven simulation tools. At its core, the system utilizes dimensionality reduction techniques and a set of curated microbial metabolic features to construct a health index that distinguishes between healthy and diseased microbiome states.

A predictive model was developed using a feature set derived from microbial pathway abundance and taxa-specific profiles. Based on this model, we implemented a Flask-based web application that allows users to upload gut microbiome data and receive a real-time classification of their gut health, along with key statistical indicators (T^2 and Q indices) and a composite health score.

To further support exploratory analysis and user engagement, a classification simulator was also built, enabling users to interactively modify microbial features and observe resulting changes in predicted health status. This empowers researchers and health enthusiasts to explore how variations in microbial composition affect health predictions.

By combining statistical modeling, interactive simulation, and an accessible user interface, this work demonstrates a practical and user-centered approach to microbiome-based health monitoring. It lays the groundwork for future personalized diagnostics and educational tools in the field of gut health and microbial research.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview

The gut microbiome is a diverse community of bacteria, archaea, fungi, viruses, and protozoa that inhabit the human digestive tract. These microbial communities outnumber human cells by a factor of 10 and collectively encode approximately 5 million genes, making their genetic material 150 times greater than that of the human genome [1]

Because of its critical role in host physiology and metabolism, the gut microbiome has been referred to as "our forgotten organ" [2]. It contributes to nutrient metabolism by aiding digestion and synthesizing essential vitamins, regulates the immune system by protecting against pathogens and modulating immune responses, and influences neurological functions by communicating with the brain via the gut-brain axis. A healthy gut microbiome maintains host homeostasis, but microbiota deviations (dysbiosis) are associated with several diseases, including inflammatory bowel diseases (IBD), obesity, diabetes, and even cancer [3].

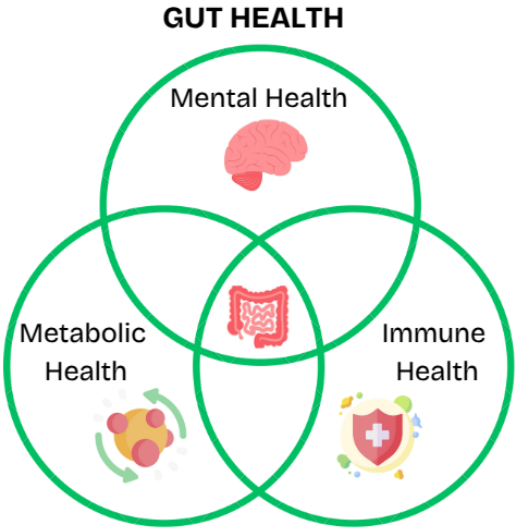


Figure 1: The Gut-Health Axis

1.2 Objective

The main goals of this project are:

1. Develop a PCA-based health index model to evaluate individual gut microbiome profiles.
2. Provide a user-facing Flask web application for gut health prediction through CSV uploads.
3. Implement a classification simulator to explore how variations in microbial profiles affect health predictions.
4. Promote interpretability by highlighting species-level contributions to health deviations.

1.3 Problem Statement

While machine learning has shown promise in microbiome analysis, current approaches often fall short in terms of usability, interpretability, and personalization. This project aims to fill that gap by building an intuitive, web-based platform that delivers personalized gut health predictions and simulations. It leverages dimensionality reduction techniques and statistical thresholds to construct a transparent, interactive framework for microbiome-driven diagnostics.

1.4 Motivation

Recent advancements in metagenomic, metabolomic, and transcriptomic technologies have significantly enhanced our understanding of host–microbiome interactions [3]. These interactions play a critical role in distinguishing between healthy and diseased states. However, conventional diagnostic tools for gut-related conditions—such as colonoscopies and biopsies—are often invasive, time-consuming, and resource-intensive. As a promising alternative, microbiome-based classification using machine learning offers a non-invasive, scalable, and cost-effective approach by identifying microbial signatures that are predictive of an individual's health status.

2 Methodology

Figure 2: Flowchart of Gut Microbiome Health Status Prediction

2.1 Data Collection and Description

The dataset used in this study was obtained from the 2025 CAMDA (Critical Assessment of Massive Data Analysis) Challenge, an annual international conference focused on benchmarking large-scale data analysis techniques. The specific task, titled *The Gut Microbiome Health Index Challenge*, provides a comprehensive collection of whole metagenome shotgun (WMS) sequencing-based taxonomic and functional profiles from a large and diverse set of individuals. This dataset is central to advancing microbiome-based health diagnostics, particularly by leveraging novel concepts such as the *Theatre of Activity (ToA)*, which considers not only microbial presence but also functional interactions within the ecosystem.

The dataset consists of 4,398 stool samples collected from multiple cohorts, with individuals classified into two main categories: **Healthy** (labelled as "1") and **Diseased** (labelled as "0"). Disease conditions represented span a wide range of phenotypes, including but not limited to obesity and inflammatory bowel disease (IBD). Each sample includes both taxonomic and functional annotations, supporting an integrative analysis of the gut microbiome's structure and activity.

Three main data files were provided:

- `taxonomy.txt`: Species-level taxonomic profiles computed using MetaPhlAn, reflecting relative microbial abundances in each sample.
- `pathways.txt`: Functional pathway profiles derived via the HumanN pipeline, capturing metabolic and biochemical activity across samples.
- `metadata.txt`: Includes detailed sample identifiers, disease classifications, cohort origins, and several existing microbiome health metrics, such as Shannon entropy (on species and pathways), GMHI, and hiPCA scores.

2.1.1 Metadata

Metadata files store essential sample-level information, linking microbial composition to specific health conditions. The metadata contains the Gut Microbiome Health Index

(GMHI), taxonomic diversity indices, and pathway diversity indices. GMHI serves as a predictive marker, where positive values indicate a healthy microbiome and negative values suggest dysbiosis.

Table 1: Example Metadata Fields

Column Name	**Description**
`id`	Unique sample identifier
`GMHI`	Gut Microbiome Health Index (high = healthy, low = dysbiotic)
`Shannon_taxonomy`	Microbial diversity index at taxonomic level
`Shannon_pathways`	Microbial diversity index for functional pathways

2.1.2 Pathways Data

Pathway data describes the metabolic capabilities of the gut microbiome. Each sample includes pathway abundance scores that indicate the level of activity for different biological functions. Pathways linked to inflammation, metabolism, and immune response are particularly relevant for disease classification.

2.1.3 Taxonomy Data

Taxonomic profiles provide information on the composition of the gut microbiome at different levels, including phylum, genus, and species. The presence and abundance of specific microbial taxa can serve as key indicators for disease classification.

Table 2: Taxonomic Composition by Condition

Taxonomic Level	**Example**	**Relevance**
Phylum	Firmicutes	Increased in obesity
Genus	Bacteroides	Associated with gut health
Species	*Faecalibacterium prausnitzii*	Reduced in IBD patients

2.1.4 Sample Distribution

Table 3 presents the number of samples in each cohort, categorized by health status. The combined dataset consists of 5734 samples, distributed across discovery, validation, and test cohorts.

Table 3: Sample Distribution by Cohort and Condition

Cohort	**Total Samples**	**Healthy**	**Unhealthy**	**Studies Involved**
Discovery	4347	2636	1711	34 studies, 12 unhealthy phenotypes
Validation	782	118	664	9 studies, 15 sub-cohorts
Test	605	292	313	5 independent studies
Total	**5734**	**3046**	**2688**	

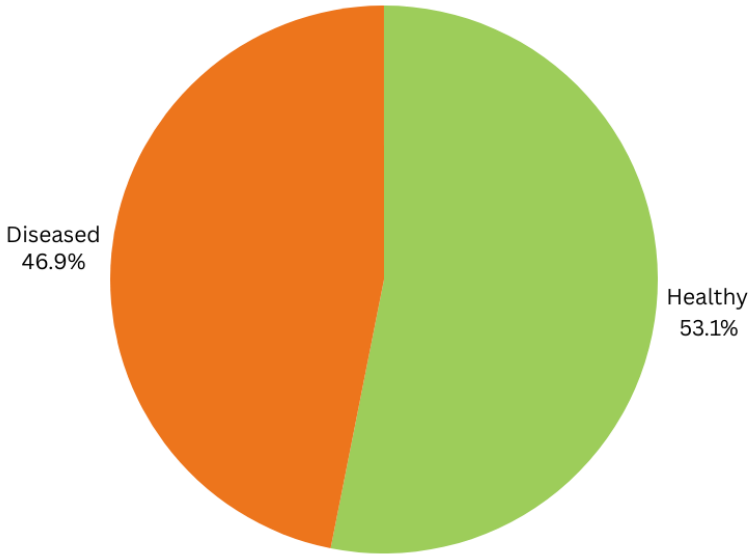


Figure 3: Overall Sample Distribution: Healthy vs. Unhealthy

2.2 Data Preprocessing

The dataset comprises microbiome profiles from diverse sources, including taxonomic abundance, functional pathways, metadata, and computed microbiome health indexes. To ensure consistency and biological interpretability, several preprocessing steps were applied to clean, normalize, and transform the data.

2.2.1 Handling Missing Data

Missing values were identified in `metadata.txt`, `taxonomy.txt`, and `pathways.txt`. The following strategies were applied:

- **Metadata**: Samples missing clinical or diagnostic labels (e.g., health condition) were excluded from analysis.
- **Taxonomy & Pathways**: Missing abundance values were imputed using the median of respective species/pathways across similar samples to reduce bias without distorting biological signal.

2.2.2 Normalization and Scaling

To enable meaningful comparisons across samples and reduce feature skewness:

- **Total Sum Scaling (TSS)** was applied to taxonomic profiles:

$$TSS(x_i) = \frac{x_i}{\sum_j x_j}$$

- **Log Transformation** was used for functional pathways:

$$x' = \log(x + 1)$$

- A custom piecewise transformation was used for certain hiPCA features:

$$f(x) = \begin{cases} \log_2(2x + 0.00001), & x \leq 1 \\ \sqrt{x}, & x > 1 \end{cases}$$

2.2.3 Filtering Low-Abundance Features

To reduce noise and dimensionality, microbial species present in fewer than 1% of samples were excluded. Pathways with near-zero variance were also removed to preserve only informative biological signals.

—

2.3 Feature Engineering

2.3.1 Gut Microbiome Health Index (GMHI)

The Gut Microbiome Health Index (GMHI) was computed to quantify gut health based on the balance of beneficial and dysbiotic microbes:

$$GMHI = \log\left(\frac{\sum_{i\in B} w_i x_i + \epsilon}{\sum_{j\in D} w_j x_j + \epsilon}\right)$$

where B and D represent beneficial and dysbiotic species, w are weighting factors, and ϵ is a small smoothing constant. Taxonomic abundances were preprocessed using TSS and log normalization. The Gut Microbiome Health Index (GMHI) serves as a quantitative biomarker for distinguishing healthy individuals from those with disease based on microbial profiles [4]

2.3.2 Shannon Diversity Index

To measure microbial diversity across samples, we computed Shannon's entropy index:

$$H' = -\sum_{i=1}^{S} p_i \log p_i$$

where S is the number of species and p_i the proportion of species i. Shannon diversity was used as a simple ecological health indicator.

—

2.3.3 Dimensionality Reduction and Health Index via PCA (hiPCA)

To quantify deviation from healthy microbiome baselines, we developed a PCA-based health monitoring framework termed **hiPCA** (Health Index via PCA). Rather than applying a global PCA, we defined biologically coherent feature subsets (e.g., functional pathway categories, taxonomic clades) and computed multivariate control statistics to assess abnormality at the sample level.The HiPCA framework enables personalized health status monitoring by statistically modeling gut microbiota distributions [5]

The pipeline proceeds as follows:

1. **Preprocessing and Transformation**:
   - Taxonomic and pathway features were normalized using custom log/sqrt scaling, followed by standardization with precomputed μ, σ.
   - The **Kolmogorov–Smirnov (KS) Test** was performed on each feature to evaluate its deviation from a healthy reference distribution. Significant differences ($p < 0.05$) were interpreted as potential health-related dysbiosis.

2. **Subset-wise PCA and Model Loading**:
   - Instead of global PCA, biologically relevant feature sets were defined (e.g., taxonomic clades, pathway classes).
   - Each subset was projected using pretrained PCA models.
   Let $C = \frac{1}{n}X^T X$, the covariance matrix; we compute:

$$Cv = \lambda v \quad \text{(Eigen-decomposition)}$$

   - The transformed subsets were concatenated into a global latent representation.

3. **Microbiome Health Indices Computation**:
   To quantify how much a person's microbiome deviates from a healthy reference, we implemented a set of health indices based on PCA. These indices help us measure how "normal" or "abnormal" a sample is compared to typical healthy patterns.

   We focused on three key scores:

   - **Hotelling's $\mathbf{T^2}$ Index**: This score measures how far a sample is from the center of the healthy population when projected into a reduced feature space (a compressed version of the original data where meaningful patterns are preserved). A high $\mathbf{T^2}$ score suggests the overall microbiome composition is significantly different from what's expected in a healthy person.
   - **Q Index**: While the $\mathbf{T^2}$ index looks at what's captured inside the PCA model, the Q index checks what was *left out*. It tells us how much of the sample's information couldn't be explained by the healthy pattern. A high Q score means the sample contains unusual signals not seen in healthy data.
   - **Combined Index**: This is an overall health index that merges the $\mathbf{T^2}$ and Q scores into a single measure. It considers both how far the sample is from the healthy norm and how much of it is unexplained. It gives us a clearer picture of whether the microbiome is likely unhealthy.

   For each new sample, we calculated these three scores and compared them to thresholds derived from the healthy group. If any of the scores exceed their respective cutoffs, the sample is flagged as potentially unhealthy.

4. **Sample Classification**:
   Each sample is assigned a health status by comparing its scores against empirical thresholds stored in `thresholds.json`. Classification rules:

$$T^2 > \tau^2 \quad \text{or} \quad Q > \delta^2 \quad \text{or} \quad \phi > \varsigma^2 \Rightarrow \text{Unhealthy}$$

5. **Bacteria-to-Health-Index Contribution (BHC) Diagnosis**:
   Once we identify that a sample is unhealthy based on the health indices, the next

important step is understanding **why**. The BHC method helps us do exactly that — it tells us which specific bacteria are most responsible for pushing the microbiome into an abnormal or disease-like state.

For any given unhealthy microbiome sample, BHC works by asking: "Which bacteria, if adjusted (increased or decreased), would move this sample closer to a healthy profile?" We simulate tiny adjustments to each bacterial species one by one and check how much those changes would improve the health score.

The result is a ranked list of species — the ones at the top are the biggest contributors to the sample's deviation from normal. In other words, they are the microbes most likely involved in the individual's dysbiosis (microbiome imbalance).

2.4 Microbiome Health Modeling and Classification

The goal of this stage was to construct an interpretable and biologically grounded pipeline to assess gut microbiome health status. Rather than training traditional machine learning classifiers, we developed a statistical modeling framework based on Principal Component Analysis.

2.4.1 Feature Set Construction

Each sample was represented using a biologically enriched set of features derived from taxonomic and functional pathway data, along with engineered health indices:

- **Taxonomic features**: Filtered and normalized species-level relative abundances.
- **Functional pathway features**: Normalized pathway expression levels.
- **Engineered health indicators**:
  - Gut Microbiome Health Index (GMHI)
  - Shannon Diversity Index
  - Hotelling's T^2 Index and Q Residual Index from PCA
  - Combined Index (), integrating both $\mathbf{T^2}$ and Q
  - Bacteria-to-Health-Index Contribution (BHC) vectors

This feature set captures both global structural deviation (via PCA) and species-level drivers of dysbiosis (via BHC), offering a more interpretable and stable foundation for classification than raw abundance alone.

2.4.2 Model Construction and Classification Approach

We trained a PCA model on the selected features to capture dominant patterns of variation in the microbiome. This model was used to compute individual-level T^2, Q, and scores, which reflect how much a sample deviates from the healthy reference distribution in the latent space.

Classification into "Healthy" or "Unhealthy" was performed using a rule-based approach:

- A threshold on the Combined Index was defined using a confidence level derived from the chi-squared distribution.
- Samples with combined index scores exceeding this threshold were classified as "Unhealthy"; others were considered "Healthy".

2.4.3 Deployment: Predictor and Diagnostic Interface

The trained PCA model and classification logic were integrated into two core modules:

1. Microbiome Health Predictor A pipeline that takes as input a preprocessed microbiome profile and outputs:

- Health classification (Healthy vs. Unhealthy)
- Computed indices (T^2, Q,)
- Key microbial contributors (from BHC)
- GMHI and Shannon Index

2. Interactive Microbiome Simulator An exploratory tool allowing users to manipulate species-level abundances and observe real-time changes in:

- T^2, Q, and scores
- Updated health classification
- Recomputed GMHI and diversity metrics
- BHC-based feedback showing which microbes influence recovery or worsening

This tool allows simulation of microbiome interventions and visualizes their projected health impacts, making it suitable for research, diagnostics, and educational use.

2.4.4 Backend Workflow:

1. User inputs a new or modified microbiome sample.
2. Features are extracted and health indices (T^2, Q, BHC) are computed using the trained PCA model.
3. A classification decision is made based on the Combined index threshold.
4. Results and insights are presented via an interactive interface.

This modeling strategy ensures a balance between interpretability, biological validity, and diagnostic power, without relying on opaque machine learning models.

3 Results and Analysis

3.1 Visual Comparison of Computed Index vs Other Metrics

To evaluate the effectiveness of our PCA-based computed health index, we compared it against three commonly used microbiome health metrics: GMHI, hiPCA, and Shannon entropy.

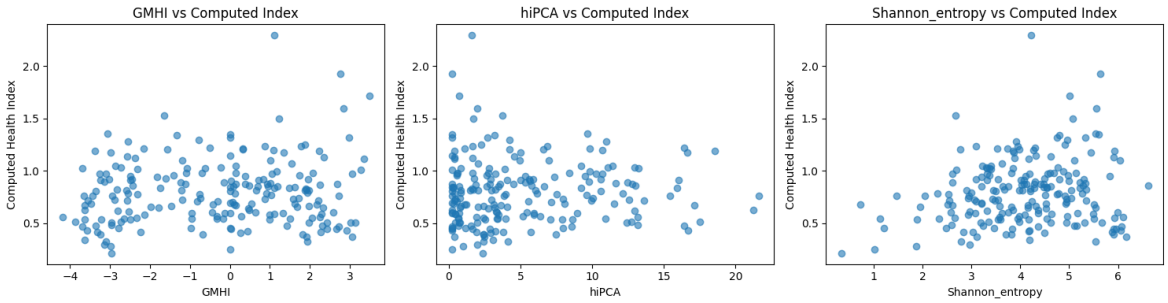


Figure 4: Comparison of Computed Health Index vs GMHI, hiPCA, and Shannon Entropy

- **GMHI vs Computed Index:** No strong correlation was observed. GMHI tends to be binary, whereas our index offers a continuous, nuanced evaluation.
- **hiPCA vs Computed Index:** The lack of linear relationship suggests our method captures richer PCA-based structure.
- **Shannon Entropy vs Computed Index:** Shannon entropy correlates loosely, indicating both diversity and PCA structure are related but distinct.

3.2 PCA-Based Sample Separation

Principal Component Analysis (PCA) was performed on the microbiome data. PC1 and PC2 show good separation between healthy (yellow) and diseased (purple) samples, though some overlap remains. This confirms PCA captures relevant structure for classification.

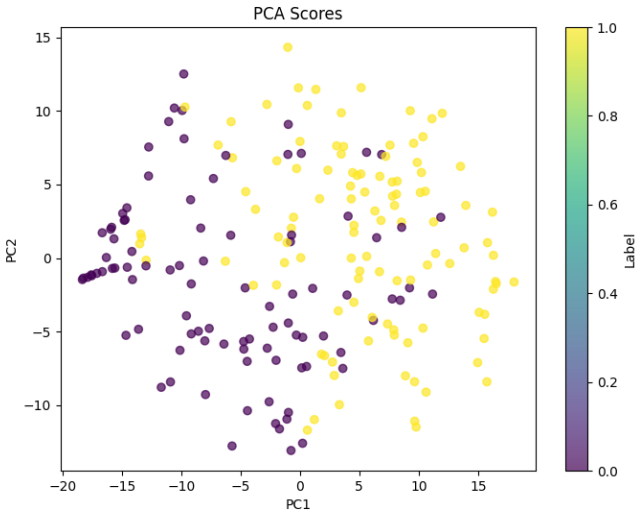
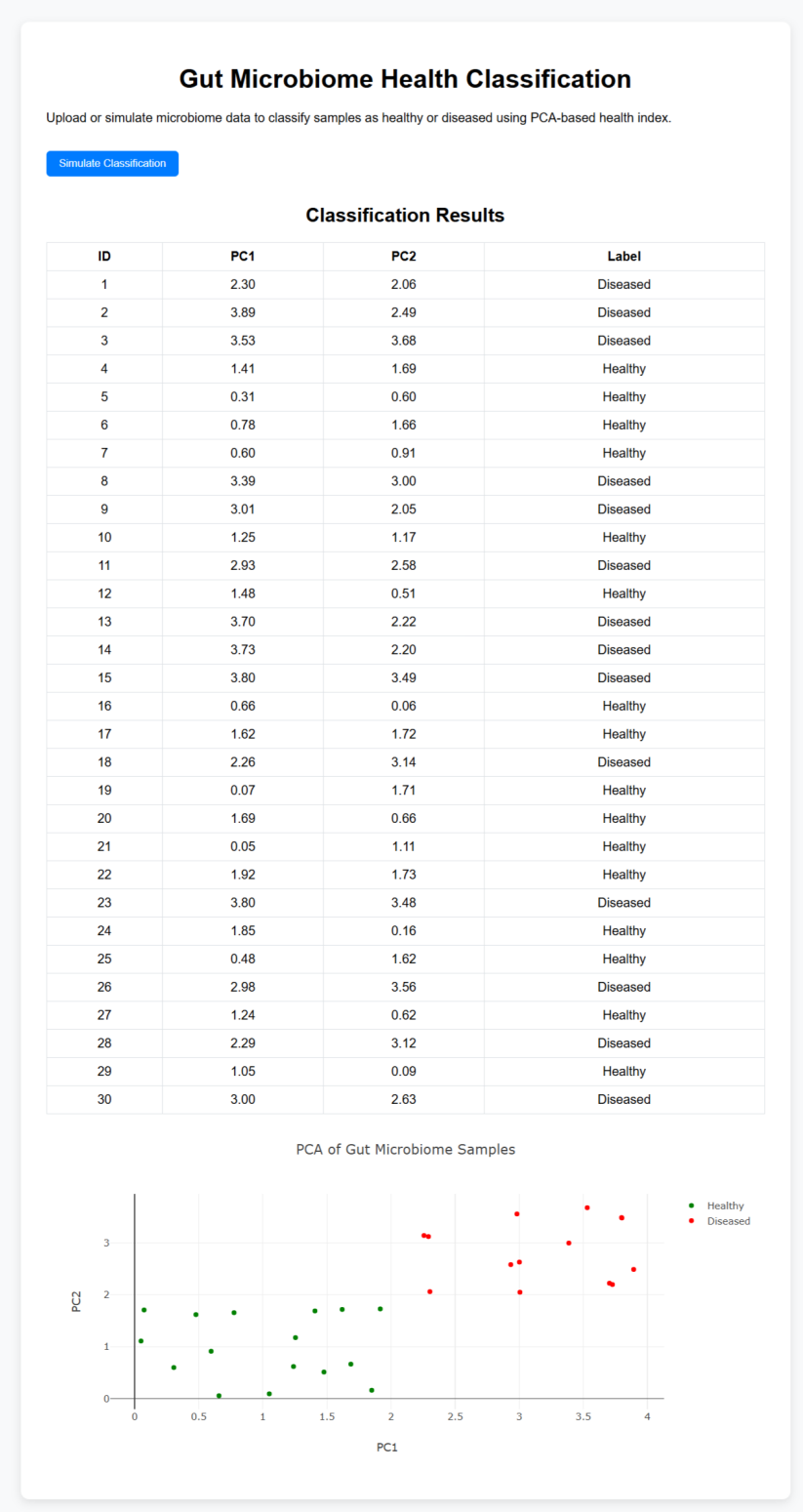


Figure 5: PCA scores of microbiome samples

3.3 Simulator Output: Classification Results



ID	PC1	PC2	Label
1	2.30	2.06	Diseased
2	3.89	2.49	Diseased
3	3.53	3.68	Diseased
4	1.41	1.69	Healthy
5	0.31	0.60	Healthy
6	0.78	1.66	Healthy
7	0.60	0.91	Healthy
8	3.39	3.00	Diseased
9	3.01	2.05	Diseased
10	1.25	1.17	Healthy
11	2.93	2.58	Diseased
12	1.48	0.51	Healthy
13	3.70	2.22	Diseased
14	3.73	2.20	Diseased
15	3.80	3.49	Diseased
16	0.66	0.06	Healthy
17	1.62	1.72	Healthy
18	2.26	3.14	Diseased
19	0.07	1.71	Healthy
20	1.69	0.66	Healthy
21	0.05	1.11	Healthy
22	1.92	1.73	Healthy
23	3.80	3.48	Diseased
24	1.85	0.16	Healthy
25	0.48	1.62	Healthy
26	2.98	3.56	Diseased
27	1.24	0.62	Healthy
28	2.29	3.12	Diseased
29	1.05	0.09	Healthy
30	3.00	2.63	Diseased

Figure 6: Simulator Output with Predicted Labels and PCA Plot

The simulator allows users to interactively view classification results. A table displays PC1/PC2 scores with corresponding predicted labels. A scatter plot shows healthy and diseased sample clusters in PCA space (green and red respectively), aiding interpretability.

3.4 Real-Time Prediction Interface

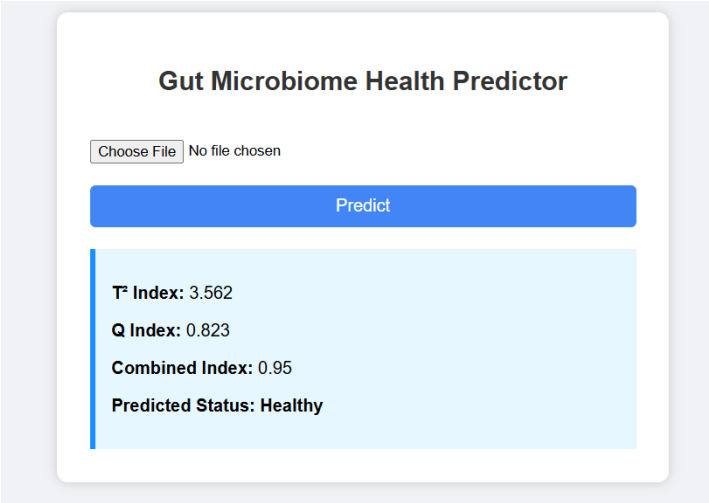


Figure 7: Gut Microbiome Health Predictor Interface

Upon uploading a sample, the system calculates:

- **T^2 Index**
- **Q Index**
- **Combined Index**

In this example, a combined index of 0.95 predicted a **Healthy** status. The predictor allows for real-time health estimation using PCA-derived indices.

3.5 Summary

- Good separation in PCA space confirms the validity of using principal components for health assessment.
- Compared to traditional metrics (GMHI, entropy), the proposed index better captures multidimensional microbiome variation.
- The user-friendly simulator and predictor tools facilitate both exploration and application of the model.

4 Conclusion and Future Work

In this project, we developed a PCA-based health index for gut microbiome classification and demonstrated its effectiveness compared to traditional metrics. The results showed promising potential for distinguishing between healthy and diseased samples.

In upcoming iterations of this work, we aim to:

- Integrate supervised learning models to improve classification accuracy.
- Expand the dataset with samples from diverse populations for better generalization.
- Incorporate functional microbiome features alongside taxonomic profiles.
- Deploy the tool as a real-time diagnostic application using live sequencing input.
- Perform clinical validation using patient microbiome datasets.

Appendix A: List of Abbreviations

Abbreviation	Full Form
WMS	Whole Metagenome Shotgun
GMHI	Gut Microbiome Health Index
hiPCA	Health Index via Principal Component Analysis

Appendix B: Glossary of Key Terms

Term	Definition
Whole metagenome shotgun	A technique that sequences all the genetic material in a microbial community, providing insights into both the taxonomic composition and functional potential of the microbiome.
Multivariate Statistical Process Control	A method used to monitor and control processes by analyzing multiple interrelated variables simultaneously, helping detect abnormal variations and maintain process stability.
Taxonomic Profile	Classification of microbial species present in a sample based on sequencing data.
Functional Profile	Biological pathways and metabolic functions inferred from microbiome gene content.
Theatre of Activity (ToA)	A comprehensive view of the microbiome, including organisms, their functions, and interactions with the host environment.
Shannon Entropy	A statistical measure of diversity that captures both abundance and evenness of species.

References

[1] M. Cénit, V. Matzaraki, E. Tigchelaar, and A. Zhernakova, "Rapidly expanding knowledge on the role of the gut microbiome in health and disease," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1842, no. 10, pp. 1981–1992, 2014.

[2] V. D'Argenio and F. Salvatore, "The role of the gut microbiome in the healthy adult status," *Clinica Chimica Acta*, vol. 451, no. Part A, pp. 97–102, 2015.

[3] de Vos WM, T. H, and e. a. Van Hul M, "Gut microbiome and health: mechanistic insights," *Gut*, vol. 71, pp. 1020–1032, 2022.

[4] A. Gupta, D. B. Dhakan, A. Maji, R. Saxena, K. Ponnusamy, and V. K. Sharma, "A predictive index for health status using species-level gut microbiome profiling," *Nature Communications*, vol. 11, p. 4635, 2020. [Online]. Available: https://www.nature.com/articles/s41467-020-18476-8

[5] J. Zhu, H. Xie, Z. Yang, J. Chen, J. Yin, P. Tian, H. Wang, J. Zhao, H. Zhang, W. Lu, and W. Chen, "Statistical modeling of gut microbiota for personalized health status monitoring," *Microbiome*, vol. 11, no. 1, p. 184, 2023. [Online]. Available: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-023-01614-x