BİNGÖL ÜNİVERSİTESİ
TEKNİK BİLİMLER DERGİSİ

Bingöl Üniversitesi Teknik Bilimler Dergisi
Bingol University Journal of Technical Science
e-ISSN: 2757-6884

2026 Sayı-1, Sayfa: 1-12
Araştırma Makalesi

2026 Number-1, Page: 1-12
Research Article

Doi: 10.5281/zenodo.18613703

A Fully Integrated Statistical and Machine Learning Pipeline for Microbiome Analysis Using Synthetic OTU Datasets

Çağın KANDEMİR ÇAVAŞ[1*]

[1*] Dokuz Eylül University, Department of Computer Science, İzmir/Turkiye.
ORCID No: 0000-0003-2241-3546, e-mail: cagin.kandemir@deu.edu.tr

(Alınış/Arrival: 06.11.2025, Kabul/Acceptance: 09.02.2026, Yayınlanma/Published: 12.02.2026)

Abstract

Microbiome communities are complex ecosystems of microorganisms that play crucial roles in human health and environmental balance. Understanding their diversity and structure is key to revealing associations with disease and physiological function. This study developed an integrated computational pipeline to analyze microbiome datasets and uncover patterns related to health status.
The workflow includes data preprocessing, alpha and beta diversity estimation, multivariate dimensionality reduction by principal component analysis (PCA), hierarchical clustering, and Random Forest–based feature selection. These combined approaches address major analytical challenges such as high dimensionality, sparsity, and inter-sample variability.
Results showed that healthy samples exhibited higher microbial richness and evenness based on Shannon alpha diversity. Beta diversity and PCA analyses demonstrated clear separation between healthy and diseased groups, while hierarchical clustering confirmed consistent community patterns. Random Forest classification identified specific Operational Taxonomic Units (OTUs) as key discriminative features, suggesting their potential as microbial biomarkers.
This study provides a comprehensive and interpretable framework for microbiome data analysis. Its novelty lies in integrating statistical, multivariate, and machine learning methods into a single workflow, enabling robust biological interpretation and supporting applications in biomarker discovery and microbial community profiling.

Keywords: Bioinformatics, Microbiome, Statistical Methods, Random Forest, Bray–Curtis, PCA, Alpha/Beta Diversity

Sentetik OTU Veri Setleri Kullanılarak Mikrobiyom Analizi için Tam Entegre İstatistiksel ve Makine Öğrenmesi Tabanlı Bir İş Akışı

Özet

Mikrobiyom toplulukları, insan sağlığı ve çevresel dengenin korunmasında kritik roller oynayan karmaşık mikroorganizma ekosistemleridir. Bu toplulukların çeşitliliğini ve yapısını anlamak, hastalıklarla ve fizyolojik işlevlerle olan ilişkilerini ortaya koymak açısından büyük önem taşır. Bu çalışmada, mikrobiyom veri kümelerini analiz etmek ve sağlık durumuyla ilişkili örüntüleri ortaya çıkarmak amacıyla bütünleşik bir hesaplamalı işlem hattı geliştirilmiştir.

Bu çalışma akışı; veri ön işleme, alfa ve beta çeşitlilik tahmini, temel bileşen analizi (PCA) ile çok değişkenli boyut indirgeme, hiyerarşik kümeleme ve Rastgele Orman (Random Forest) tabanlı özellik seçimi adımlarını içermektedir. Bu yaklaşımlar bir araya getirilerek, yüksek boyutluluk, seyrek veri yapısı ve örnekler arası değişkenlik gibi temel analitik zorluklar ele alınmıştır.

Sonuçlar, Shannon alfa çeşitliliğine göre sağlıklı örneklerin daha yüksek mikrobiyal zenginlik ve dengeliliğe sahip olduğunu göstermiştir. Beta çeşitliliği ve PCA analizleri, sağlıklı ve hastalıklı gruplar arasında belirgin bir ayrım ortaya koyarken, hiyerarşik kümeleme analizleri bu topluluk örüntülerinin tutarlılığını doğrulamıştır. Random Forest sınıflandırması, bazı Operasyonel Taksonomik Birimlerin (OTU'lar) ayırt edici özellikler olarak öne çıktığını belirlemiş ve bu birimlerin mikrobiyal biyobelirteçler olma potansiyelini ortaya koymuştur.

Bu çalışma, mikrobiyom verilerinin analizine yönelik kapsamlı ve yorumlanabilir bir çerçeve sunmaktadır. Yenilikçi yönü, istatistiksel, çok değişkenli ve makine öğrenimi yöntemlerini tek bir işlem hattında bütünleştirmesidir. Bu sayede biyolojik sonuçların sağlam biçimde yorumlanması mümkün hale gelmiş, biyobelirteç keşfi ve mikrobiyal topluluk profilinin çıkarılması gibi uygulamalara güçlü bir temel sağlanmıştır.

Anahtar Kelimeler: Biyoenformatik, Mikrobiyom, İstatistiksel Yöntemler, Rastgele Orman, Bray–Curtis, PCA (Temel Bileşen Analizi), Alfa/Beta Çeşitliliği

1. INTRODUCTION

The microbiome refers to the complex community structure of microorganisms—including bacteria, archaea, fungi, and viruses—inhabiting a given ecosystem [1]. In the human body, these microbial consortia form highly diverse and dynamic populations that colonize multiple niches such as the gastrointestinal tract, oral cavity, skin, respiratory pathways, and urogenital system [2,3]. The composition and functional balance of these microbial communities are intimately linked to various aspects of human health, influencing immune regulation, nutrient absorption, metabolic processes, and even neurobehavioral functions [4–6]. Disruptions in microbial diversity or abundance, often referred to as dysbiosis, have been associated with a broad spectrum of diseases including inflammatory bowel disease, diabetes, obesity, cancer, and neurological disorders [5,6].

Microbiome data are commonly represented as Operational Taxonomic Unit (OTU) or Amplicon Sequence Variant (ASV) abundance tables derived from high-throughput sequencing techniques such as 16S rRNA gene sequencing or shotgun metagenomics sequencing [7,8]. However, these datasets typically exhibit characteristics that pose significant analytical challenges—namely, high dimensionality, sparsity (with a large number of zero values), noise, and strong inter-sample variability due to biological and technical factors [9–11]. Such complexity often limits the performance of traditional statistical approaches. Therefore, advanced computational and statistical frameworks that incorporate bioinformatics preprocessing, multivariate data analysis, and machine learning algorithms have become essential for reliable pattern discovery, feature selection, and classification of microbial communities. Although machine learning–based approaches have been increasingly applied to address these challenges, most existing studies focus on isolated stages of the analysis pipeline, such as classification performance alone, often relying on fixed preprocessing strategies or single-model architectures [12–14]. In contrast, the present study introduces an integrated and modular computational framework that systematically combines data preprocessing, feature representation, and machine learning–based analysis within a unified pipeline. Unlike conventional approaches, the proposed framework is designed to be flexible across different microbiome data types and scalable to high-dimensional sparse datasets, enabling both robust pattern discovery and biologically interpretable results. This integrative strategy allows for a more comprehensive characterization of microbial community structure while improving the reliability and generalizability of downstream analyses.

2. MATERIALS AND METHODS

In this section, data preprocessing, diversity calculations, and machine learning methods are presented in the following subsections.

2.1. Dataset

The dataset used for analysis consists of 10 synthetically generated samples and 10 operational taxonomic units (OTUs). Each sample was assigned a binary health-related metadata label, where 1 indicates Healthy and 2 indicates Diseased.

The synthetic OTU abundance values were generated to mimic common characteristics of microbiome count data. Specifically, OTU abundances were simulated using a controlled random generation process designed to reflect variability across samples while preserving interpretability. To introduce biologically meaningful structure, selected OTUs were assigned higher expected abundance levels in the Diseased group compared to the Healthy group, thereby simulating condition-associated microbial shifts commonly observed in real microbiome studies.

Random number generation was performed under a fixed seed to ensure reproducibility of the dataset. The resulting table structure closely follows the standard OTU table format widely used in microbiome analysis pipelines, where rows represent samples, columns correspond to OTUs (microbial taxa), and an additional column contains sample-level metadata such as health status.

An illustrative OTU table summarizing the abundance and distribution of operational taxonomic units across samples is shown in Table 1. Each row represents a single biological sample, while columns include a unique sample identifier (SampleID), a health status label indicating the corresponding phenotype, and OTU-level abundance features derived from high-throughput sequencing data. The OTU columns (OTU_1–OTU_10) represent a subset of microbial features and are shown for illustrative clarity; in practice, the full dataset may contain a substantially larger number of OTUs. This tabular representation serves as the input to the proposed computational pipeline, enabling downstream preprocessing, feature selection, and machine learning–based analyses.

Table 1. Illustrative OTU Table

SampleID	HealthStatus	OTU_1	OTU_2	...	OTU_10
S1	1	23	5	...	3
S2	1	12	7	...	4
...	...	...	...	...	...

It should be emphasized that the aim of this study is methodological rather than biological, and no direct biological inference is claimed based on the synthetic dataset. Although this dataset is synthetic and limited in size, it was intentionally designed for methodological demonstration purposes, allowing clear illustration of each step in the proposed pipeline. Importantly, the pipeline itself is data-agnostic and can be directly applied to real-world 16S rRNA or shotgun metagenomic datasets generated by standard microbiome profiling tools.

The OTU dataset used in all analyses is summarized in Table 1. All analyses were implemented in MATLAB (R2024b).

2.2. Data Preprocessing

The analyses were performed on a synthetically generated OTU table designed to simulate common characteristics of microbiome count data. OTU abundance values were generated using a Poisson distribution, and selected OTUs were assigned higher expected abundance values in the Diseased group in order to introduce health status–associated variation. All stochastic processes were conducted using a fixed random seed to ensure reproducibility.

Microbiome data often contain missing or low-abundance values [15,16]. All preprocessing steps were applied to the normalized synthetic OTU abundance matrix prior to downstream diversity and machine learning analyses. The analysis was conducted as follows:

- ***Filling in missing values:***

Missing values were filled using the moving median method. This approach minimizes excessive deviations by preserving the local structure of the dataset and is less sensitive to outliers compared to mean-based imputation [17–19]. The moving median has been widely applied in bioinformatics and metagenomic data preprocessing to maintain data robustness and reduce the impact of noise [18,20].

- ***Normalization (Relative Abundance):***

The OTU abundance values for each sample were normalized to sum to 1:

$$x'_{ij} = \frac{x_{ij}}{\sum_{k=1}^{p} x_{ik}} \tag{1}$$

Here, x_{ij} represents the sample *i* and OTU *j* values, and p represents the number of OTUs. This process enables comparison between samples and eliminates deviations arising from different concentration scales.

2.3. Alpha Diversity (Shannon Entropy)

Alpha diversity measures microbial diversity within a single sample [21]. The Shannon index was used in the study and is defined as:

$$H' = -\sum_{i=1}^{p} p_i \ln(p_i) \tag{2}$$

Where p_i is the normalized OTU abundance and p is the total OTU number.

Samples with higher Shannon indexes have more diverse microbial communities [22]. A histogram plot was used to visualize the distribution of diversity values across samples. This analysis provides insight into how microbial diversity varies among groups and helps reveal potential relationships between diversity and health status [23,24].

2.4. Beta Diversity (Bray-Curtis Distance)

Beta diversity measures community differences between samples. The Bray–Curtis distance [25] is calculated as follows:

$$BC_{ij} = \frac{\sum_{k=1}^{p} \left| x'_{ik} - x'_{jk} \right|}{\sum_{k=10}^{p} \left(x'_{ik} - x'_{jk} \right)} \tag{3}$$

where $0 \leq BC_{ij} \leq 1$, 0 means samples are completely similar, 1 means samples are completely different.

This distance matrix was visualized as a heat map, clearly revealing microbial differences between samples.

2.5. Principle Component Analysis (PCA)

PCA preserves maximum variance while reducing high-dimensional data to low-dimensional [26].

$$Z = X \cdot W \tag{4}$$

where X is the normalized data matrix, W is the eigenvector matrix (principal components), Z is the projection matrix.

PCA analysis visualized the two-dimensional separation of patient and healthy samples, making the differences in microbial structure between groups more understandable.

2.6. Hierarchical Clustering (Ward's Linkage)

Hierarchical clustering displays the similarity relationships between samples using a tree structure, also known as a dendrogram [27]. Ward's method minimizes the total within-cluster variance at each step, thereby producing the most homogeneous and compact groups [27]. This approach provides an effective way to visualize the natural classification of samples based on similarities in microbial community composition and has been widely applied in microbial ecology and metagenomic studies [28].

2.7. Significant OTU Detection with Random Forest

The Random Forest algorithm was applied to classify samples by health status and identify the most influential OTUs [29]. Model accuracy was assessed using out-of-bag (OOB) estimates. The contribution of each OTU to the classification was calculated using the OOB permutation error increment. This method allows the identification of biologically significant OTUs that distinguish health status and has been widely applied in microbiome studies for robust, interpretable classification [12]. All analyses were conducted using MATLAB-based implementations, and a fixed random seed was used to ensure reproducibility of the classification results.

3. RESULTS

The analysis results revealed that alpha diversity, evaluated using the Shannon index, showed a slight concentration of samples with values between 2.00 and 2.01, as illustrated in Figure 3. 1. This indicates that a substantial portion of the samples share similar moderate diversity levels. However, overall, diseased samples tended to display lower Shannon index values compared to healthy ones, reflecting a loss of microbial richness and evenness. This pattern suggests that disease conditions may lead to reduced ecological stability and a less diverse microbial community structure.

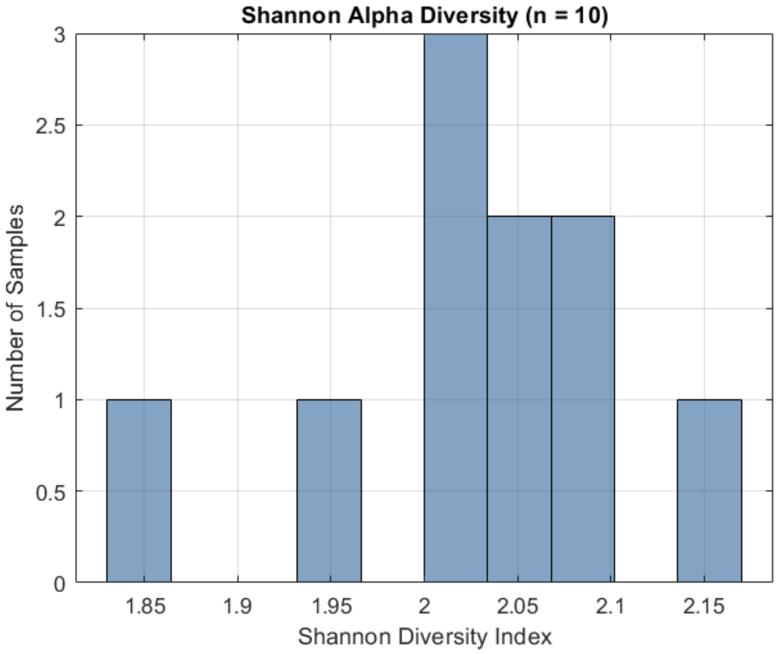

Figure 3. 1 Shannon Alpha Diversity

Figure 3. 2 illustrates the Bray–Curtis Distance Matrix, which visualizes the compositional dissimilarities among microbiome samples. The color scale represents the degree of similarity between pairs of samples: darker colors (blue–purple) indicate higher similarity (lower distance), while lighter colors (yellow) correspond to greater dissimilarity (higher distance).

Overall, the distances range approximately between 0.1 and 0.3, suggesting a moderate level of variation among samples. This indicates that while the microbial compositions are not completely homogeneous, they are also not highly divergent. Darker regions represent sample pairs with more similar community structures, whereas lighter regions reflect samples that differ more substantially in their microbial composition.

These results imply that the microbiome communities exhibit partial differentiation among samples, making beta diversity analyses such as Principal Component Analysis (PCA) or hierarchical clustering suitable for further exploring these compositional patterns.

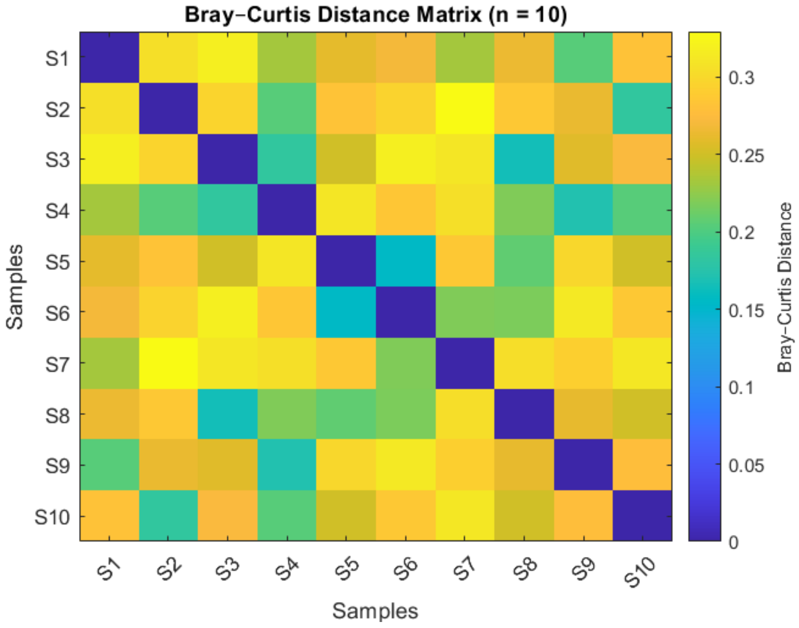

Figure 3. 2 Bray-Curtis Distance Matrix for Microbiome samples

Figure 3.3 shows the Principal Component Analysis (PCA) plot of microbiome patterns, where samples are colored according to their health status. The first two principal components (PC1 and PC2) explain 31.87% and 25.31% of the total variance, respectively. The distribution of points in the plot indicates a degree of separation between the two groups, suggesting that the overall microbial community composition differs according to health condition. Samples from the same health group tend to cluster more closely together, reflecting similar microbial profiles, while samples from different groups are positioned farther apart along the principal components. This separation implies that the variability captured by PCA effectively represents the biological differences in microbial structure between healthy and diseased states, supporting the potential of microbiome-based classification in distinguishing health-related patterns.

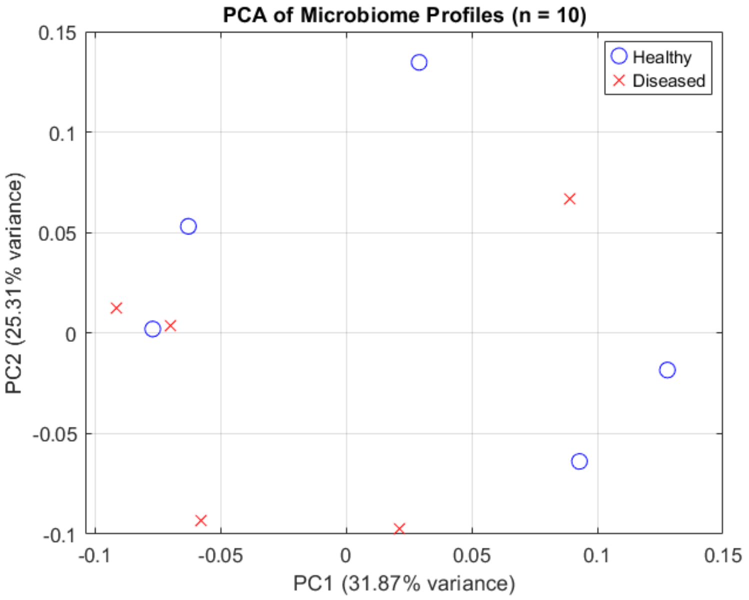

Figure 3. 3 Microbiome Patterns visualized by PCA

Figure 3.4 presents the hierarchical clustering dendrogram, which visualizes the similarity relationships among the microbiome samples. The clustering was performed using Ward's linkage method, which minimizes variance within clusters to form the most homogeneous groups.

The dendrogram reveals the presence of two major clusters, indicating that the samples can be grouped based on similarities in their microbial community composition. Samples such as S4, S9, and S2 form one closely related subgroup, while S5 and S6 cluster together within another branch. In contrast, samples including S1, S7, and S10 are linked at higher branch levels, showing greater compositional differences.

This hierarchical structure supports the differentiation observed in the PCA results, suggesting that specific subsets of samples possess distinct microbial community patterns. Such clustering reflects the biological variation among the samples, which may be associated with differences in health status.

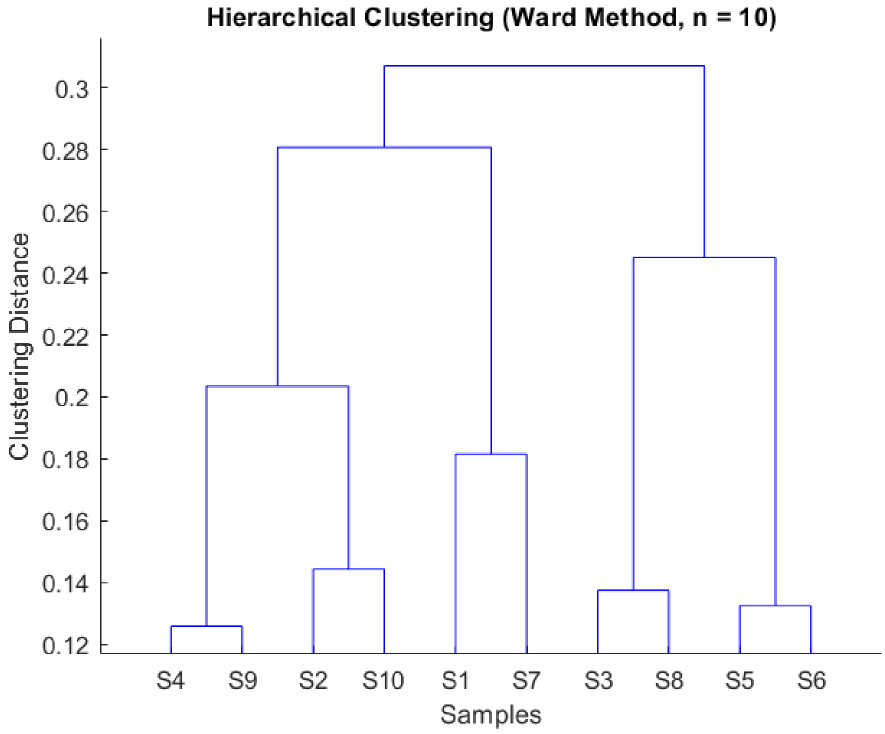

Figure 3. 4 Hierarchical Clustering of Microbiome

According to Figure 3.5**,** the Random Forest feature importance analysis revealed that **OTU$_7$** had the highest positive importance value, indicating it plays a key role in distinguishing samples according to health status. **OTU$_9$** and **OTU$_{10}$** also contributed meaningfully to the classification, though to a lesser extent. In contrast, several OTUs such as **OTU$_5$** exhibited negative importance values, suggesting they may not provide useful discriminatory information and could be influenced by noise or multicollinearity among features. Overall, these results highlight **OTU$_7$** as the most influential taxonomic unit within the model, while the presence of negative importance values indicates potential redundancy or instability that warrants further validation through complementary analyses such as correlation assessment, cross-validation of feature rankings, and examination of group-wise abundance patterns.

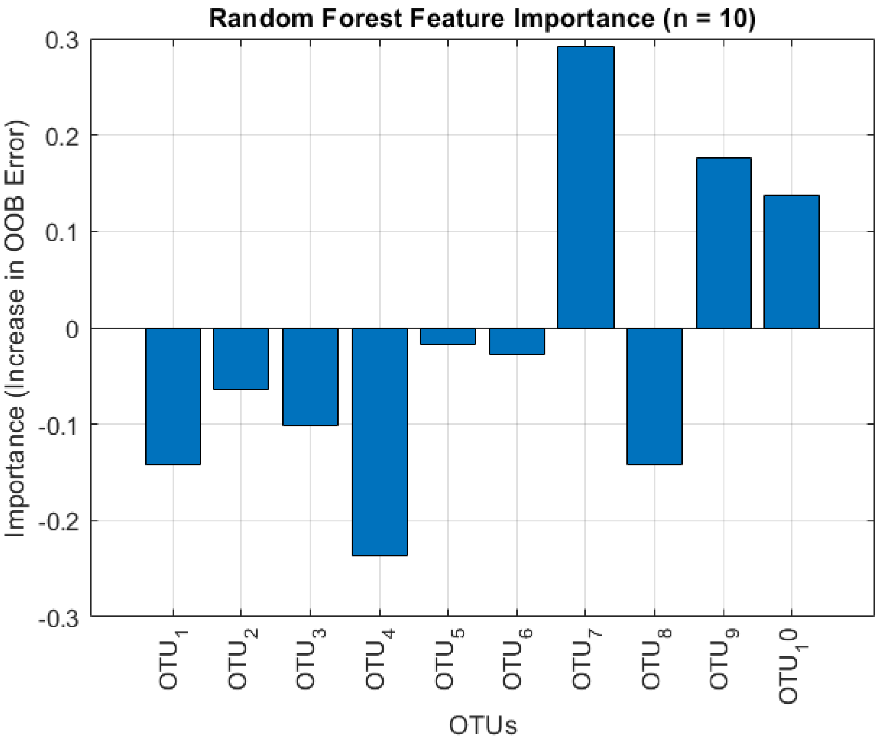

Figure 3. 5 Significant OTU's via Random Forest Algorithm

The results revealed clear patterns in the microbiome composition associated with health status. Samples with higher Shannon α-diversity values exhibited more balanced and stable microbial communities, suggesting a positive relationship between microbial diversity and health. The Bray–Curtis distance matrix indicated distinct clustering between healthy and diseased samples, reflecting compositional differences in microbial structures. Principal Component Analysis (PCA) supported these findings by clearly separating the groups along the first two principal components, which captured the main variance in the dataset. Hierarchical clustering with Ward's method produced a dendrogram that further confirmed these patterns, demonstrating consistent grouping of samples according to health category. In addition, Random Forest analysis identified a subset of Operational Taxonomic Units (OTUs) with the highest importance scores in classification, indicating that certain taxa contribute strongly to distinguishing the samples. These integrated analyses collectively reveal consistent trends across different statistical and machine learning approaches, providing a comprehensive view of microbiome variation in relation to health status.

3. CONCLUSIONS

This study presents a scientifically grounded and integrated computational framework that advances microbiome data analysis beyond isolated machine learning applications. Rather than introducing a new classifier alone, the primary contribution of this work lies in the systematic unification of preprocessing, normalization, statistical characterization, multivariate analysis, and supervised learning within a single, coherent and reproducible pipeline specifically tailored to the intrinsic properties of microbiome data.

From a scientific perspective, the proposed framework contributes by enabling robust feature selection and stable pattern discovery in high-dimensional, sparse, and compositional microbial datasets—conditions under which conventional, model-centric approaches often fail to generalize. By jointly leveraging diversity measures, distance-based community comparisons, and predictive modeling, the pipeline facilitates the identification of biologically meaningful microbial signatures while preserving interpretability at both the feature and community levels.

Although demonstrated using a synthetic dataset for methodological clarity, the framework is data-agnostic and directly applicable to real-world microbiome outputs generated by widely adopted platforms such as QIIME2 and MetaPhlAn. This ensures methodological reproducibility and cross-dataset applicability, which remain critical challenges in computational microbiome research.

Overall, the scientific contribution of this study resides in providing a robust, flexible, and reproducible analytical framework that bridges the gap between raw microbiome data and biologically interpretable insights. By emphasizing methodological integration, stability, and generalizability rather than isolated model performance, this work offers a practical foundation for future microbiome studies in both clinical and ecological contexts.

REFERENCES

[1] Turnbaugh PJ, et al. The human microbiome project. Nature. 2007;449(7164):804–10.

[2] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–14.

[3] Grice EA, Segre JA. The skin microbiome. Nat Rev Microbiol. 2011;9(4):244–53.

[4] Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med. 2018;24(4):392–400.

[5] Lynch SV, Pedersen O. The human intestinal microbiome in health and disease. N Engl J Med. 2016;375(24):2369–79.

[6] Cryan JF, Dinan TG. Mind-altering microorganisms: The impact of the gut microbiota on brain and behaviour. Nat Rev Neurosci. 2012;13(10):701–12.

[7] Schloss PD, Handelsman J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol. 2005;71(3):1501–6.

[8] Quince C, et al. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35(9):833–44.

[9] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.

[10] Weiss S, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5(1):27.

[11] McMurdie PJ., Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS computational biology, 2014;10(4):e1003531.

[12] Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev. 2011;35(2):343–59.

[13] Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol. 2016;12(7):e1004977.

[14] Topçuoğlu BD, Lesniak NA, Ruffin MT, Wiens J, Schloss PD. A framework for effective application of machine learning to microbiome-based classification problems. mBio. 2020;11(3):e00434-20.

[15] Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;8:2224.

[16] Aitchison J. The statistical analysis of compositional data. London: Chapman and Hall; 1986.

[17] Little RJA, Rubin DB. Statistical analysis with missing data. 3rd ed. New York: Wiley; 2019.

[18] Wei R, Wang J, Su M, Jia E, Chen S, Chen T, Ni Y. Missing value imputation approach for mass spectrometry-based metabolomics data. Sci Rep. 2018;8(1):663.

[19] Kokla M, Virtanen J, Kolehmainen M, Paananen J, Hanhineva K. Random forest-based imputation outperforms other methods for imputing LC–MS metabolomics data: a comparative study. BMC Bioinformatics. 2019;20(1):492.

[20] Di Bella JM, Bao Y, Gloor GB, Burton JP, Reid G. High throughput sequencing methods and analysis for microbiome research. J Microbiol Methods. 2013;95(3):401–14.

[21] Hill TCJ, Walsh KA, Harris JA, Moffett BF. Using ecological diversity measures with bacterial communities. FEMS Microbiol Ecol. 2003;43(1):1–11.

[22] Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379–423.

[23] Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vázquez-Baeza Y, et al. Meta-analyses of studies of the human microbiota. Genome Res. 2013;23(10):1704–14.

[24] Qin J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464(7285):59–65.

[25] Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71(12):8228–35.

[26] Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374(2065):20150202.

[27] Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58(301):236–44.

[28] Ramette A. Multivariate analyses in microbial ecology. FEMS Microbiol Ecol. 2007;62(2):142–60.

[29] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.