IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 6, Issue 1, February 2026

Impact Factor: 8.2

Enhancing Microbiome Based Disease Prediction with SuperTML and Data Augmentation

Ms. Manisha Eswari. M[1] and Reyaa. K[2]

Assistant Professor, Department of Computer Science and Engineering[1]
Student, Department of Computer Science and Engineering[2]
Vandayar Engineering College, Thanjavur, India

Abstract: *Accurate disease prediction using microbiome data remains a major challenge due to high dimensionality, compositional complexity, and limited sample sizes. Traditional machine learning models frequently struggle to capture the subtle interactions and patterns within such data, leading to reduced predictive performance and poor generalization to unseen cases. To address these limitations, this project introduces an advanced framework that integrates Super Tabular Machine Learning with data augmentation techniques. SuperTML enables deep feature representation to capture subtle microbial patterns. By transforming raw microbiome data into a structured, learnable representation, SuperTML allows the model to uncover hidden correlations and discriminate between healthy and diseased conditions with higher accuracy. Data augmentation techniques expand the effective size and diversity of the training dataset, mitigating the risk of overfitting and improving the model's ability to generalize to new, unseen samples. This ensures the model remains robust even when faced with sparse or noisy data. By integrating SuperTML with data augmentation, this project not only improves the predictive accuracy of microbiome based disease classification but also enhances interpretability, allowing researchers and clinicians to better understand the microbial signatures associated with different diseases.*

Keywords*:* ML, SuperTML, Data Augmentation, Microbiome

I. INTRODUCTION

Accurate disease prediction using microbiome data is a growing area of interest in biomedical research and precision medicine, as the human microbiome plays a vital role in influencing health and disease. Although the availability of large microbiome datasets through advanced sequencing technologies, analysing this data effectively remains a major challenge. Microbiome data is highly dimensional, sparse, and compositional, making it difficult for traditional analytical and machine learning methods to extract meaningful patterns. These models often fail to capture the subtle interactions among microbial taxa that are important indicators of disease conditions. In addition, limited sample sizes and noisy measurements increase the risk of overfitting and reduce the generalization capability of predictive systems. This creates a need for advanced computational approaches that can better handle the complexity of microbiome data. To address this problem, the proposed project introduces a framework that integrates SuperTML with data augmentation techniques for improved disease classification. SuperTML converts raw microbial data into structured representations and learns deep features that reveal hidden correlations between microbes and disease. Data augmentation increases the diversity of the training data, reduces overfitting, and improves robustness when dealing with sparse or noisy samples. By combining these methods, the system enhances prediction accuracy, reliability, and interpretability. This integrated approach provides a dependable tool for microbiomebased disease prediction and supports the goals of precision medicine and personalized healthcare.

II. LITERATURE REVIEW

Numerous studies have applied machine learning techniques to microbiome data for disease prediction and understanding. [1] Pasolli et al. (2016) demonstrated the use of machine learning classifiers, including random forests and support vector machines, to predict host phenotypes from gut microbiome data, establishing the viability of

DOI: 10.48175/568

microbiome-based diagnostics. [2] Reiman et al. (2020) applied deep learning models to 16S rRNA sequencing data, showing that convolutional neural networks could outperform traditional methods in classifying inflammatory bowel disease states by learning higher-order microbial interactions. [3] Marcos-Zambrano et al. (2021) systematically reviewed the application of machine and deep learning on microbiome data, highlighting data sparsity and the "curse of dimensionality" as persistent challenges that require novel preprocessing and modeling strategies. [4] Chen et al. (2019) introduced SuperTML, a novel method for transforming tabular data into a two-dimensional format, enabling the application of powerful computer vision models to non-image classification tasks, which has shown success in various biomedical datasets. [5] Topçuoğlu et al. (2020) specifically addressed the issue of data augmentation in microbiome studies, implementing synthetic oversampling techniques to improve model performance on small, imbalanced cohort data, proving its necessity for robust predictive modeling.

III. MACHINE LEARNING

Machine Learning is a technique that enables computer systems to learn from data without being explicitly programmed for every individual task. Instead of following fixed rules, the system learns from examples and increases its performance as more data is provided. By combining data with statistical methods and intelligent algorithms, Machine Learning can identify hidden patterns and relationships within large datasets. These patterns are then used to make accurate predictions and support decision-making.Here Machine Learning is applied to analyze microbiome data for disease prediction. The algorithms process microbial abundance information to detect subtle patterns that may not be easily recognized through manual analysis. Similar to how recommendation systems suggest content based on user preferences and history, Machine Learning models can predict potential diseases by learning from variations in microbial composition across samples. There are two primary categories of machine learning. They are as followed,

A. Supervised learning:

An algorithm, much like a smart student, learns from examples rather than relying on step-by-step instructions from humans. It studies training data and uses feedback to understand how different inputs are related to a desired output. Over time, the algorithm identifies hidden patterns, trends, and relationships within the data that may not be obvious to humans. For example, a marketer might provide data such as marketing expenses, seasonal trends, and weather forecasts to help the algorithm predict the sales of canned products. By analyzing this information repeatedly, the algorithm learns how these factors influence sales and becomes capable of making accurate predictions when new data is provided.

B. Unsupervised Learning:

Unsupervised learning is a type of machine learning in which the algorithm is given data without any labels or predefined outcomes. The system is not told what to predict; instead, it examines the data and tries to find structure within it on its own. It groups similar data points together, identifies patterns, and detects relationships that may not be immediately visible. This approach is useful for tasks such as clustering, pattern recognition, and discovering hidden trends in large datasets. Unlike supervised learning, where the model learns from known input–output pairs, unsupervised learning allows the algorithm to explore freely and draw conclusions based solely on the data it observes.

IV. SYSTEM ARCHITECTURE

In this figure, a traditional machine learning model serves as a baseline, trained on microbiome data that has already been preprocessed. To train a CNN model, the same data is processed simultaneously through data augmentation and transformed into 2D formats using the SuperTML. This approach allows a direct comparison between deep learning methods that use 2D formats and conventional algorithms. At the end of the process, a final evaluation is conducted to determine which model provides better biological insights and performance.

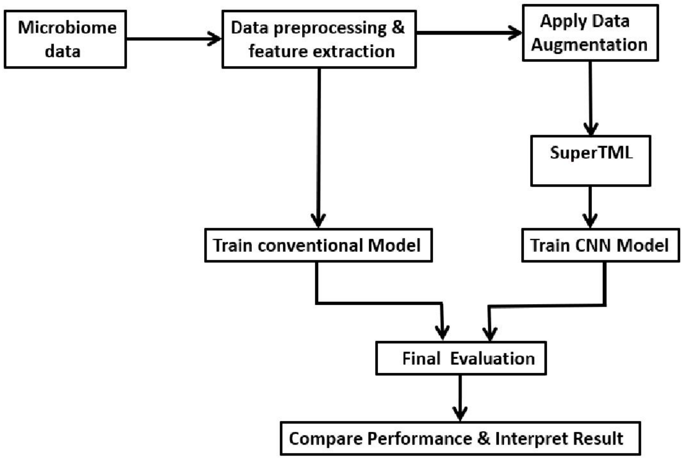

Fig 1: System Architecture

V. PROPOSED METHODOLOGY

Data Collection and Data Preprocessing:

Patient samples are used to collect the initial microbiome dataset, which is then processed. Information about the types of microorganisms present in these samples forms the basis for predicting diseases. The dataset includes various microbiome characteristics such as diversity levels, the quantity of different microbes, and related biological properties. After the dataset is collected, it goes through a preprocessing step. During this process, any missing, conflicting, or unnecessary data entries are detected and removed. The cleaned data is then arranged and categorized according to the needs of the proposed system for further analysis.

Data Augmentation:

The SMOTE (Synthetic Minority Over-sampling Technique) approach was used to increase the number of samples in the microbiome dataset and reduce the imbalance between classes. Instead of simply copying existing samples, SMOTE creates new synthetic samples by combining data from nearby samples in the dataset. This method maintains the original data structure while introducing additional variation. By generating more samples for the smaller classes, the dataset becomes more balanced and diverse, which supports the learning process of the SuperTML–CNN model and leads to better prediction accuracy.

SuperTML for Feature Transformation:

The SuperTML approach is applied to transform preprocessed tabular microbiome data into a two-dimensional format. SuperTML interprets tabular data as similar to images by converting numerical features into structured visual patterns that look like heatmaps. The system leverages visual feature learning, where Convolutional Neural Networks can automatically identify spatial patterns without requiring manual feature design due to the 2D representation. By combining image-based and language-based deep learning models with tabular data, SuperTML enhances the system's overall ability to extract features and make predictions.

Convolutional Neural Network:

The microbiome data, after undergoing transformation through SuperTML, is used as input for a Convolutional Neural Network (CNN) that is based on the ResNet18 architecture. Before it enters the network, the 2D data goes through preprocessing steps such as resizing, normalization, and converting to a single-channel format. The CNN, which is specifically ResNet18, uses its convolutional layers to automatically detect and extract important spatial patterns from the heatmaps. These extracted features are then passed through fully connected layers to perform classification. During the training phase, the model's parameters are updated through backpropagation, and the version of the model that performs best on validation data is kept.

Here's how the CNN works:
Step 1: The input image is processed and fed into the network.
Step 2: Convolutional layers apply filters to the image to identify basic features like textures and patterns.
Step 3: Activation functions and pooling layers are used to reduce the image size and highlight relevant features.
Step 4: As the network moves through deeper layers, it combines identified features to recognize more patterns.
Step 5: The final fully connected layer uses the learned features to classify the image into the correct category.

Disease Prediction:

The patterns of the microbiome identified in the input data are used by the trained CNN model to predict the presence of a disease. The system correctly categorizes different diseases by examining the patterns it has learned. This supports better medical decisions and helps in identifying diseases at an earlier stage.

Performance Evaluation:

The performance of the proposed system is evaluated using different measurement standards to fully check how well it can make correct predictions. Accuracy shows how correct the model is by calculating the percentage of samples that were predicted correctly compared to the total number of samples. Precision tells us, among all the samples the model marked as belonging to a certain category, how many of those were actually correct. Recall, on the other hand, shows how well the model can find all the real samples that belong to a specific category, highlighting its ability to catch all relevant cases. The F1-Score gives a balanced view by combining both precision and recall into a single value, which is especially useful when the dataset has an unequal number of samples in different categories. In addition, confusion matrices are used to visually present the results of classification, showing both correct and incorrect predictions for each category. Together, these measurements help evaluate the system's dependability, consistency, and effectiveness in predicting diseases based on microbiome data.

VI. CONCLUSION

This project tackles the issues of complex and limited microbiome data by using SuperTML, data augmentation, and a CNN. SuperTML helps in transforming the microbiome data into 2D formats, while data augmentation increases the variety of samples to make the training process more effective. The CNN model then uses the data from SuperTML to make accurate predictions about diseases, which enhances the overall accuracy and reliability of the predictions.

VII. FUTURE SCOPE

This research holds great promise for future growth by improving disease prediction systems that rely on the microbiome, particularly through the use of bigger and more varied datasets. As microbiome studies advance, including more detailed microbial characteristics can significantly improve the accuracy and trustworthiness of predictions. Future efforts might involve combining genomic and metabolomic data to achieve a better understanding ofbiological processes. Employing sophisticated deep learning structures and combined techniques can further enhance system performance.

REFERENCES

[1]. A. Chaussard, M. Monshizadeh, and Y. Liu, "A taxonomy aware augmentation strategy for microbiome data prediction," -2025.

[2]. P. Przymus, M. K. Johnson, and A. J. Turner, "Deep learning in microbiome analysis: A comprehensive review of neural network models," Frontiers in Microbiology-2025.

[3]. J. Han, Y. Zhang, and Z. Liu, "Techniques for learning and transferring knowledge in microbiome-based disease prediction," Nature Communications 2025.

[4]. G. Tazza, D. Ruggeri, and L. Vidács, "Improving microbiome-based disease prediction with SuperTML and data augmentation," IEEE -2025.

[5]. D. Sharma, S. K. Gupta, and S. S. S. R. Depuru, "phylaGAN: Data augmentation through conditional GANs and autoencoders for improving disease prediction accuracy using microbiome data," Bioinformatics-2024.

[6]. Q. Wang, Y. Zhang, and Z. Liu, "PM CNN: Microbiome status recognition and disease detection," Bioinformatics Advances- 2024.

[7]. M. Monshizadeh, Y. Liu, and P. S. K. R. Reddy, "Incorporating metabolic activity, taxonomy, and clinical metadata for microbiome-based disease prediction," Journal of Translational Medicine-2024.

[8]. G. Roy, P. S. K. R. Reddy, and M. S. K. R. Reddy, "Deep learning methods in metagenomics: A review," PubMed Central-2024.

[9]. M. David, M. S. K. R. Reddy, and P. S. K. R. Reddy, "Revealing general patterns of microbiomes that transcend study systems,"- 2022.

[10]. B. Sun, L. Yang, W. Zhang, M. Lin, P. Dong, C. Young, and J. Dong, "SuperTML: TwoDimensional Word Embedding for the Precognition on Structured Tabular Data", IEEE-2019

DOI: 10.48175/568