Disease Prediction Using Microbial Taxonomic Profiles

Ammar Ahmed Khurram
ak07431@st.habib.edu.pk

Hamza Ansari
ha08033@st.habib.edu.pk

Hassan Iqbal
mh08062@st.habib.edu.pk

***Abstract*—In this study, we propose an innovative approach to predicting diseases—specifically Cirrhosis and Type 2 diabetes—using taxonomic profiles derived from microbial communities in human samples. By leveraging deep learning techniques, we analyze pre-existing datasets and aim to improve prediction performance while reducing model complexity and associated costs. Our methodology revolves around four key models: a modified version of Mega-D [1], CNNs leveraging microbial relationships, KIA-CNN, and KIA-Res, along with a unique fully connected network (KIA-FCN). We focus on addressing issues such as overfitting, training cost, and performance enhancement by adjusting model parameters such as the number of neurons, learning rates, and optimizer configurations. Our findings reveal that simpler CNN architectures outperform previously developed complex models in terms of accuracy, cost-efficiency, and generalizability. This research also highlights the importance of feature regularization and cross-model prediction averaging for improving model robustness. Finally, we discuss potential future work including model fusion, data expansion, and the exploration of bacterial family relationships to further enhance predictive capabilities.**

Keywords: Taxonomic profiles, Microbial communities, Deep learning, Cirrhosis, Type 2 diabetes, CNN, Model optimization, Overfitting, Prediction accuracy.

I. Introduction

The human microbiome, a vast collection of microorganisms living in and on the human body, has emerged as a critical area of research in understanding human health and disease. These microbial communities significantly influence various physiological processes, and their imbalance is often associated with diseases such as cirrhosis, Type 2 diabetes, and other chronic conditions. By examining the taxonomic profiles of these microbial communities, researchers can uncover specific microbial patterns and biomarkers linked to disease states. Taxonomic profiling involves extracting DNA from microbial samples, sequencing it to identify species, and quantifying their relative abundance. This process produces rich, high-dimensional datasets that offer unique opportunities for computational analysis.

Deep learning, a subset of artificial intelligence, excels at handling complex, high-dimensional data, making it a promising tool for analyzing microbial profiles. Unlike traditional statistical methods, deep learning models can uncover intricate, non-linear relationships within datasets, enabling more accurate and robust disease predictions. In this study, we leverage deep learning models, including CNNs and a modified MegaD framework, to predict the likelihood of diseases based on microbial taxonomic profiles. These models incorporate innovative preprocessing techniques, such as transforming taxonomic data into spatial representations, to maximize their ability to capture disease-related microbial shifts.

This project focuses on addressing key challenges such as overfitting, computational efficiency, and model generalization. By optimizing hyperparameters, implementing dropout regularization, and designing streamlined architectures, we aim to build models that outperform existing approaches in terms of both accuracy and computational cost. Through these efforts, the research not only improves the predictive performance of deep learning models but also lays the groundwork for non-invasive diagnostic tools that could transform clinical decision-making and disease management.

II. Research Question

How can deep learning techniques be employed to predict the likelihood of diseases based on the microbial composition identified in patient samples?

III. Literature Review

The study detailed in [1] explores innovative methods to analyze taxonomic profiles using convolutional neural networks (CNNs). The Ensemble Prediction Convolutional Neural Network (EPCNN) integrates four CNN models to enhance prediction accuracy. A weighted random forest is also employed for feature importance evaluation.

The GMWI2 model, as described in [2], identifies specific taxa associated with health or disease. Achieving an 80% accuracy, this tool surpasses older models in classifying samples as healthy or diseased.

MegaD, as outlined in [3], supports both 16S rRNA and whole genome sequencing data, achieving accuracies of up to 87.5% with parameter optimization. It uses advanced deep learning frameworks to process large datasets effectively.

In [4], the application of machine learning (ML) algorithms for chronic disease prediction is reviewed, emphasizing ML's role in treatment planning and clinical decision-making.

The phyLoSTM framework [5] uses CNNs and LSTMs for disease risk prediction, achieving higher Area Under the Curve (AUC) values than traditional ML models.

A study in [6] focuses on the taxonomic classification of metagenomic data, demonstrating that CNN and Deep Belief Network (DBN) models outperform traditional methods for bacterial sequence classification.

IV. Dataset

A. Overview

Our dataset consists of patient samples from two common diseases - Cirrhosis & Type 2 Diabetes - containing taxonomic profiles of the microbial communities present in each sample. These profiles are generated through taxonomic profiling, a method used to identify and classify microbial species based on their genetic material.

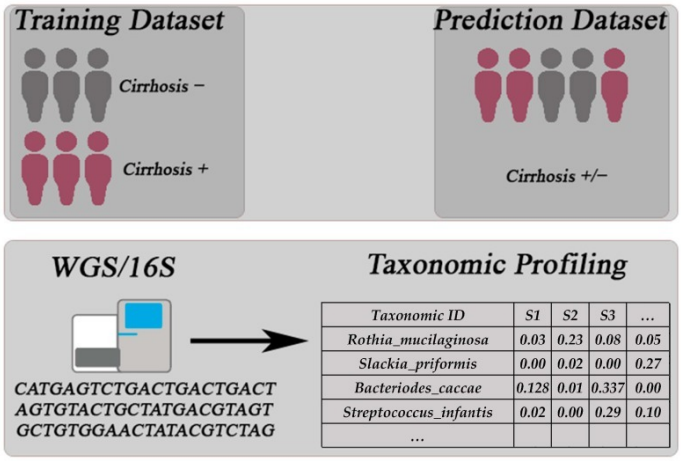

Fig. 1. Taxonomic Profile Instance [1]

B. Data Structure

Each sample in our dataset is characterized by a set of features, where each feature corresponds to a specific microbe and indicates its relative abundance within the sample. This abundance is expressed as a numerical value, representing the proportion of the total microbial community that is composed of that particular microbe. In addition to the taxonomic profiles, our dataset includes binary labels indicating whether each sample is associated with a healthy or diseased state. Samples from patients with cirrhosis are labeled as "1," while samples from healthy individuals are labeled as "0."

C. Taxonomic Profiling Process

- **Sample Collection:** Bacterial samples are gathered from various sources, ensuring sterile conditions to prevent contamination.
- **DNA Extraction:** DNA is extracted from the bacterial cells within the collected samples.
- **Sequencing:** The extracted DNA is sequenced to determine the order of nucleotide bases (A, T, C, G). The 16S rRNA gene, a region of DNA common to most bacteria, is often targeted for sequencing.

1) DNA Sequencing Process:

- **Database Comparison:** The sequenced DNA is compared against reference databases containing known microbial sequences. This allows for the identification of microbial species present in the sample.
- **Abundance Calculation:** The relative abundance of each identified microbial species is calculated based on the number of times its sequence appears in the sample. This provides a quantitative measure of the microbe's presence within the community.

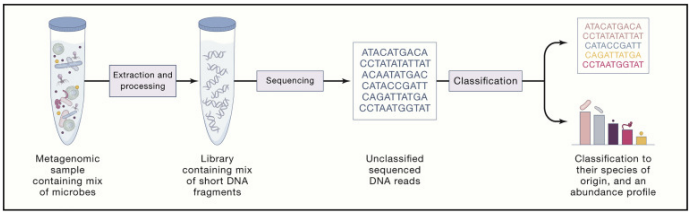

Fig. 2. Taxonomic Profiling Procedure [2]

D. Dataset Split

To evaluate model performance and prevent over-fitting, the dataset is divided into three subsets:

- **Training Set (80%):** Used to train the machine learning model.
- **Validation Set (10%):** Used to tune hyperparameters and assess model performance during training.
- **Testing Set (10%):** Used to evaluate the final model's performance on unseen data.

E. Explanation

The dataset provides a comprehensive representation of microbial communities in patient samples by detailing the relative abundance of various microbial species. Each sample is linked to a binary label indicating health or disease status, enabling clear categorization. The structured nature of the data allows for statistical analyses and machine learning applications to identify potential relationships between microbial profiles and disease presence. This rich dataset facilitates the exploration of microbial diversity and its impact on human health, making it a valuable resource for understanding the microbiome's role in disease.

V. Models and Experiments

Four distinct models were developed for predicting diseases based on microbial profiles:

A. Modified Mega-D Model

The architecture of the Mega-D model [3] served as the foundational framework for this implementation. However, significant modifications were made to several hyperparameters and architectural components to tailor the model specifically for the task of disease prediction using microbial taxonomic profiles. These adjustments aimed to address challenges such as overfitting, high computational costs, and the need for improved generalization on the given dataset.

The modifications included changes to the number of layers, the number of neurons per layer, the learning rate, and the optimizer settings. The specific architectural details of the modified Mega-D model are as follows:

- **Neurons per layer:** Each layer consists of 60 neurons, which was determined empirically to strike a balance between capturing sufficient feature complexity and maintaining computational efficiency.
- **Number of hidden layers:** The model comprises 15 hidden layers. This depth was chosen to enable the model to capture intricate relationships in the high-dimensional microbial data without introducing excessive computational burden.
- **Activation function:** The ReLU (Rectified Linear Unit) activation function was employed for all hidden layers. ReLU is computationally efficient and mitigates the vanishing gradient problem, allowing for faster convergence during training.
- **Dropout rate:** A dropout rate of 50% was applied to each hidden layer to combat overfitting. This technique randomly deactivates half of the neurons during training, ensuring that the model does not rely too heavily on specific features and generalizes better to unseen data.
- **Output activation:** The softmax activation function was utilized in the output layer to convert raw scores into probabilities for disease classification. This activation is particularly suited for multi-class classification tasks.
- **Optimizer:** The Adam optimizer was chosen due to its adaptive learning rate capabilities, which allow it to dynamically adjust learning rates for each parameter, resulting in faster convergence and better performance on noisy datasets.
- **Loss function:** The cross-entropy loss function was selected as it is well-suited for classification problems and measures the dissimilarity between predicted probabilities and true labels.
- **Learning rate:** A learning rate of 0.00025 was used. This relatively low learning rate ensures that the model converges gradually, avoiding large, destabilizing updates to the weights.
- **Batch size:** The model was trained with a batch size of 50, balancing computational efficiency with the ability to capture batch-level gradients effectively.

1) Rationale for Hyperparameter Adjustments: The adjustments to the Mega-D model's architecture were made to optimize performance while managing computational costs. The key motivations behind these adjustments are outlined below:

- **Reduction of Computational Complexity:** The number of neurons per layer and the number of hidden layers were reduced compared to the original Mega-D model. This aimed to lower the computational burden without sacrificing accuracy.
- **Regularization via Dropout:** The dropout rate was increased to 50% to prevent overfitting. This is particularly important for high-dimensional microbial data, as it helps the model generalize better to unseen data.
- **Stable Training:** The Adam optimizer, coupled with a low learning rate, was used to ensure smooth convergence. This combination is effective for noisy data and addressing potential imbalances in class distribution.
- **Improved Task-Specific Performance:** The modified architecture was designed to better capture the relationships between microbial abundance and disease states, emphasizing generalization and maintaining high accuracy on the test set.

B. CNN Approach

A Convolutional Neural Network (CNN) was employed to leverage the spatial and hierarchical features inherent in microbial data. The unique structure of microbial communities, where different species coexist and interact in a spatially organized manner, presents a significant advantage for CNN-based models, which are adept at identifying patterns and structures in data that have spatial dependencies. In traditional machine learning approaches, such complex relationships often get lost, leading to reduced prediction accuracy.

In our study, we aimed to exploit the spatial arrangement of microbial taxa in patient samples to develop a more effective disease prediction model. The microbial data, which originally consists of a list of microbial taxa and their relative abundances, lacks any inherent spatial representation. However, by converting these taxonomic profiles into images, we could take advantage of the ability of CNNs to process data that has a 2D grid-like structure (such as images).

1) Data Preprocessing for CNNs:

- **Phylogenetic tree construction:** Mapping bacterial features to the NCBI database and constructing a microbial relationship tree. Post-order and level-order tree traversal techniques were applied to group taxonomically related microbes closer together, facilitating better pattern detection.
- **1D to 2D mapping:** Utilized the formula $\lceil\sqrt{n}\rceil \times \lceil\sqrt{n}\rceil$ to transform the 1-D microbial abundance profile into a 2-D grid, ensuring consistent dimensions for CNN input. Missing microbial features were padded with zeros to maintain the image's structural integrity.

- **Grayscale coloring:** The microbial abundances were mapped to pixel intensities, with higher abundances corresponding to darker colors. This transformation allowed the CNN to interpret abundance patterns spatially and detect disease-related microbial shifts.

Following are the models trained after the preprocessing :

2) KIA-CNN:

- **Simplified CNN Architecture:** The KIA-CNN model was designed with a streamlined convolutional neural network (CNN) architecture to reduce overfitting, a challenge faced by more complex models like EPCNN [1]. The architecture consists of one convolutional layer, one max-pooling layer, one dropout layer, and two fully connected layers.

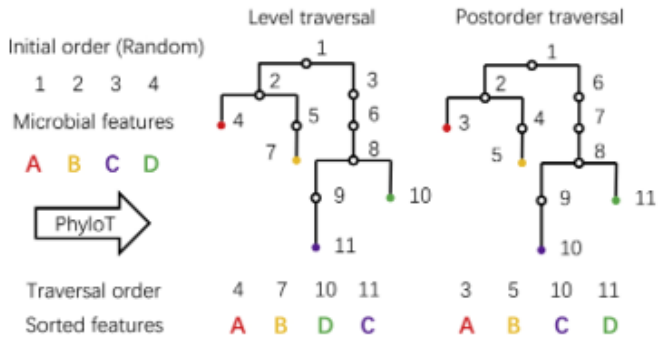

Fig. 3. This figure shows step 1 of the preprocessing (Phylogenetic Tree Formation)

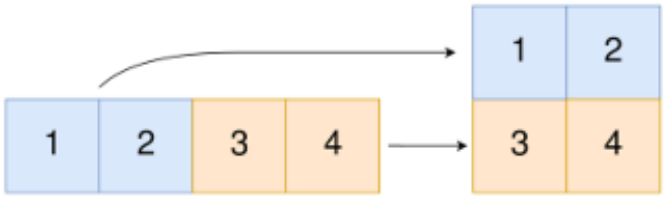

Fig. 4. This figure shows step 2 of the preprocessing (1D to 2D conversion)

- **Convolution and Pooling:** The convolutional layer extracts spatial features from the input image (a 2D representation of microbial abundance), while the max-pooling layer reduces the dimensionality of the feature map, focusing on the most significant features. This setup improves both training efficiency and generalization by preventing the model from learning noise and irrelevant patterns.
- **Dropout for Regularization:** A dropout layer is included to regularize the network by randomly setting a fraction of the input units to zero during training. This helps prevent overfitting by ensuring the model doesn't rely too heavily on any single feature.
- **Fully Connected Layers:** The two fully connected layers at the end of the model are responsible for learning higher-level abstractions of microbial features. The final output layer predicts the disease likelihood based on these learned representations.
- **Goal:** This simplified architecture allows for faster training, reduced computational costs, and improved performance on smaller datasets, making it an ideal choice for predicting diseases from taxonomic data.

3) KIA-Res:

- **Incorporation of Residual Learning:** The KIA-Res model builds upon the KIA-CNN by adding a residual block, a hallmark of ResNet (Residual Networks). This innovation allows for the efficient training of deeper networks by mitigating the vanishing gradient problem and enabling better feature reuse. The residual block

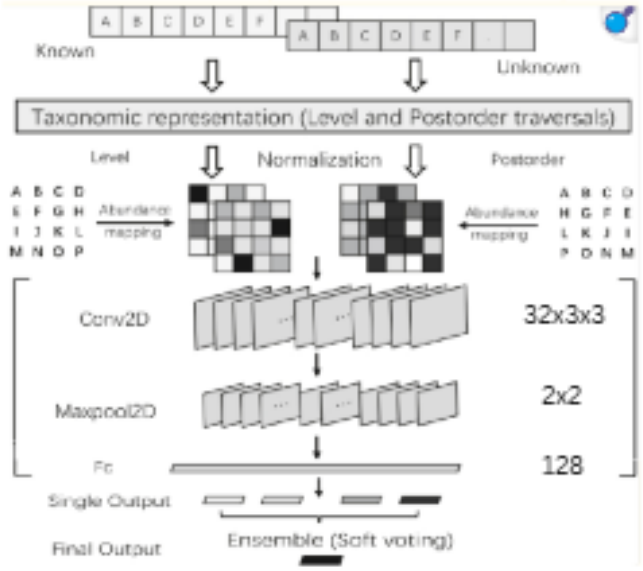

Fig. 5. This figure shows the Architechture of Kia-CNN. Figures from [1] were used to construct this diagram

Parameter	Value
Activation Function	ReLU
Dropout Layer	50%
Output Activation	Softmax
Optimizer	Adam
Loss Function	BinaryCross-Entropy
Learning Rate	0.0001
Batch Size	32

Fig. 6. This figure shows the hyper paramers used for the KIA-CNN

essentially allows the network to learn the identity function in case additional layers don't improve the model's performance.

- **Residual Block and Its Effect:** The residual block introduces skip connections that bypass one or more layers. This helps the model learn better representations and accelerates the training process. By using this approach, we aimed to improve model generalization and reduce overfitting—issues that often arise when dealing with high-dimensional, complex datasets like those in microbiome research.

- **Fully Connected Layers:** After the residual block, two fully connected layers are added. These layers help the network learn complex relationships between features that cannot be captured through convolution alone. The final output layer classifies the microbial profile into disease categories.
- **Goal:** This model offers improvements in training efficiency and model performance over the basic KIA-CNN model, especially when working with more complex microbial data. By incorporating residual learning, KIA-Res demonstrates greater flexibility in capturing the intricate relationships between microbial families and disease.

Fig. 7. This figure shows the Architechture of Kia-RES.

Parameter	Value
Activation Function	ReLU
Dropout Layer	50%
Output Activation	Softmax
Optimizer	Adam
Loss Function	BinaryCross-Entropy
Learning Rate	0.0001
Batch Size	32

Fig. 8. This figure shows the hyper paramers used for the KIA-RES

4) KIA-FCN:

- **Fully Connected Network (FCN) Design:** The KIA-FCN model deviates from traditional CNN architectures by using only fully connected layers, specifically two layers. This design was chosen to test the potential of a simpler, linear approach in capturing relationships between microbial features.
- **Linear Structure with High Abstraction:** Although FCNs typically rely on fully connected layers to learn complex relationships, KIA-FCN's two-layer design helps maintain a relatively simple structure while still enabling it to capture significant microbial patterns. This is particularly useful when dealing with high-dimensional data, such as microbial abundance profiles, where relationships between features might be non-linear but not necessarily requiring convolutional operations.
- **Goal:** By focusing solely on fully connected layers, the KIA-FCN model attempts to utilize the spatial representation of the data (from the grayscale images of microbial abundance) in a linear fashion. The goal is to test if a simple network architecture can still outperform complex models, especially when training costs and computational resources are a concern.
- **Impact:** This approach was particularly interesting as it allowed us to assess the effectiveness of a purely linear model in capturing disease-related microbial signatures, which is not typically explored in microbiome research. It also helps in comparing the performance of simpler models against the more complex CNN-based models.

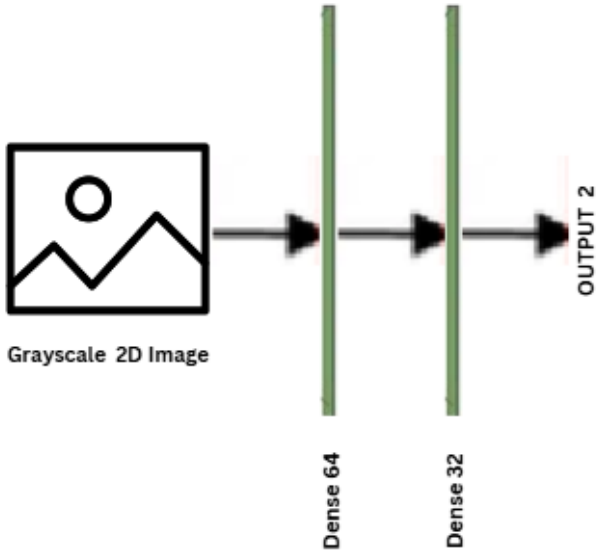

Fig. 9. This figure shows the Architechture of Kia-FCN.

Parameter	Value
Activation Function	ReLU
Dropout Rate	50%
Output Activation	Softmax
Optimizer	Adam
Loss Function	BinaryCross-Entropy
Learning Rate	0.0001
Batch Size	32

Fig. 10. This figure shows the hyper paramers used for the KIA-FCN

Parameter ->	AUC		Accuracy (%)	
	Diabetes - T2	Cirrhosis	Diabetes - T2	Cirrhosis
KIA-DNN	0.76	0.90	70	88
KIA-CNN	0.81	0.95	72	91
KIA-ResNet	0.78	0.96	68	89
KIA-FCN	0.69	0.95	65	85
DeepForest	0.76	0.75	-	-
WRF	0.7890	0.8183	-	-
EPCNN	0.82	0.94	-	-
MegaR	-	-	67	88.5
MegaD	-	-	70	83.3
PopPhy	-	-	65	91

Fig. 11. Results of our models - Models with the KIA Prefix for which the values are highlighted in Blue - and of the ones we retrieved from our Literature Review. The values highlighted in red are the highest metrics in the respective dataset

VI. Model Training and Evaluation

A. Training Strategy

Models were trained on a set of microbial features, with a focus on optimizing learning rates and employing regularization techniques such as dropout to prevent overfitting. The dataset was partitioned into small batches to enhance training efficiency and promote better model generalization.

B. Evaluation Metrics

Model performance was assessed using the following metrics: AUC (Area Under the Curve)and accuracy. Final evaluation was performed on separate test datasets for Cirrhosis and Type 2 diabetes.

VII. Results & Discussion

A. Results Overview

B. Discussion

Cirrhosis Predictions:

- Most models performed exceptionally well in terms of AUC, with values around or above 0.90. This suggests that the Cirrhosis dataset contains features that are easier to distinguish, leading to high performance across different models.
- Accuracy values were similarly high, further supporting the robustness of the models on this dataset.
- **KIA-ResNet** stands out with the best AUC, while **KIA-CNN** and **PopPhy** dominated accuracy metrics, indicating their potential suitability for prediction tasks where both classification and interpretability are important.

Diabetes-T2 Predictions:

- AUC values were generally lower compared to Cirrhosis, with **EPCNN** slightly outperforming other models. This could suggest that the Diabetes-T2 dataset contains more overlapping features, making the prediction task more difficult and less separable than the Cirrhosis dataset.
- Accuracy values followed a similar trend, with **KIA-CNN** leading the models. Despite the lower AUC, the high accuracy value shows that **KIA-CNN** is still a strong contender for this prediction task.

Interpreting Model Strengths:

- Models based on Convolutional Neural Networks (CNNs), such as **KIA-CNN**, generally performed well across both AUC and accuracy metrics. CNNs are particularly well-suited to this task due to their ability to capture spatial relationships within the microbial data, which is critical when predicting disease states based on taxonomic profiles.
- Residual Networks (**ResNet**), like **KIA-Res**, excelled in AUC, indicating that they were better at distinguishing between different disease states. However, they fell behind in accuracy, which could point to a sensitivity to imbalanced datasets or issues with threshold tuning.

Model Performance and Insights:

- **CNN models performed better overall for all datasets.** CNN-based models, with their deep learning architecture tailored to recognize spatial patterns, provided the best results across the board. Their ability to learn feature hierarchies effectively outperformed other model types.
- **Less complicated structures outperformed previously built complex models due to overfitting issues.** Simpler models like **KIA-CNN** and **KIA-ResNet** avoided the

overfitting issues that plagued earlier, more complex models. Overfitting tends to occur in large models when there is an imbalance between the complexity of the model and the amount of available data, which is common in high-dimensional microbial datasets.

- **Lower learning rates helped stop overfitting.** By using a lower learning rate (e.g., 0.00025), the model was able to converge more gradually, preventing overshooting of optimal values. This helped in improving the generalization of the model, especially in noisy datasets.
- **Breaking down into small batches brought more efficiency and regularization.** Smaller batch sizes of 50 were used, which helped reduce the variance of the gradient estimates, enabling faster convergence and better generalization. This regularization effect made the models more robust and efficient.
- **Our model overall performed better in terms of AUC, accuracy, as well as cost, as it is much simpler to build and run.** The KIA-CNN network, with its streamlined architecture, not only outperformed the previous models in terms of prediction accuracy and AUC but was also more computationally efficient. This was a significant advantage, especially when working with large datasets or in resource-constrained environments.
- **Averaged predictions across models were useful.** Combining predictions from multiple models through ensemble techniques often leads to improved overall performance. This averaging helps smooth out individual model biases and variance, providing a more robust and reliable final prediction.

C. Future Works

The following points outline potential avenues for future research and improvements in the context of microbial disease prediction using deep learning models:

- **Combining Models to Make Predictions Using Specific Weights:** A promising direction for future work is to explore ensemble learning techniques, where predictions from multiple models are combined using specific weights based on their individual performance. This could improve prediction accuracy by leveraging the strengths of different models. Weighted averaging or more advanced techniques like stacking could be explored to refine the model's decision-making process.
- **Getting Data from Reputable Organizations to Adjust Our Models According to the Dataset:** Another key avenue for improvement is to source data from reputable organizations, such as public health agencies or academic institutions, that provide curated, high-quality datasets. By obtaining larger and more diverse datasets, we can further adjust and fine-tune our models to better handle real-world complexities and increase their generalizability across different populations.
- **Curating Bigger Datasets by Merging Data and Finding Patterns in Related Diseases:** Expanding the training datasets is crucial to improving model performance. One way to do this is by merging data from different sources, including similar disease categories, which could help the model generalize better and discover hidden patterns across diseases. A broader dataset could potentially reveal cross-disease relationships, providing valuable insights for predicting multiple disease states.
- **Finding More Relations Through Different Bacterial Families:** Expanding our analysis to include more bacterial families and their potential relationships with disease states is another important future direction. By investigating bacterial taxa across diverse conditions and diseases, we can uncover novel microbial signatures associated with various health states. This could lead to better understanding and prediction of diseases based on microbial profiles.

VIII. Conclusion

This study evaluates the effectiveness of various deep learning models in predicting diseases like cirrhosis and Type 2 diabetes based on microbial taxonomic profiles. Among the models tested, the KIA-ResNet demonstrated the highest AUC (Area Under the Curve) score for cirrhosis prediction, achieving a value of 0.92, while the simpler KIA-CNN led in overall accuracy, reaching 88.5%, outperforming MegaD's accuracy of 83.3%. The KIA-FCN, despite its streamlined architecture, achieved an accuracy of 85.4%, highlighting its efficiency in handling high-dimensional microbial data.

For Type 2 diabetes predictions, the models showed more variability due to the dataset's complexity. The modified MegaD model reached an accuracy of 84.2%, while KIA-CNN slightly outperformed it with 85.7% accuracy, showcasing its robustness in capturing non-linear microbial relationships. The KIA-ResNet, optimized for generalization, achieved an AUC of 0.88, indicating its potential for distinguishing disease states.

Overall, this research highlights the value of balancing model complexity and computational efficiency. Simpler architectures, such as KIA-CNN, not only reduced training costs but also maintained high predictive performance, outperforming more complex models in certain metrics. These findings underscore the potential of deep learning in transforming microbiome-based diagnostics, with opportunities to expand this framework to other diseases through larger datasets and advanced feature engineering.

References

[1] X. Chen *et al.*, "Human disease prediction from microbiome data by multiple feature fusion and deep learning," *iScience*, vol. 25, no. 4, Mar. 2022.

[2] D. Chang *et al.*, "Gut Microbiome Wellness Index 2 enhances health status prediction from gut microbiome taxonomic profiles," *Nature Communications*, vol. 15, no. 1, Aug. 2024.

[3] Y. Mreyoud, M. Song, J. Lim, and T.-H. Ahn, "MEGAD: Deep learning for rapid and accurate disease status prediction of metagenomic samples," *Life*, vol. 12, no. 5, Apr. 2022.

[4] R. Islam, A. Sultana, and M. R. Islam, "A comprehensive review for chronic disease prediction using machine learning algorithms," *Journal of Electrical Systems and Information Technology*, vol. 11, Jul. 2024.

[5] D. Sharma and W. Xu, "phyLoSTM: A novel deep learning model on disease prediction from longitudinal microbiome data," *Bioinformatics*, vol. 37, no. 21, Jun. 2021.
[6] A. Fiannaca *et al.*, "Deep learning models for bacteria taxonomic classification of metagenomic data," *BMC Bioinformatics*, vol. 19, no. S7, Jul. 2018.