Harnessing Generative AI to Build a Foundation Model for Human Microbiome Analysis and Precision Medicine

Nicholas A. Medearis[1,2,3] and Ali R. Zomorrodi[2,3]

MassGeneral Hospital for Children
Mucosal Immunology and Biology Research Center

[1]Massachusetts Institute of Technology Department of Electrical Engineering and Computer Science, [2]Mucosal Immunology and Biology Research Center, Department of Pediatrics, Massachusetts General Hospital, [3]Harvard Medical School

1. Abstract

While various tools have been developed to study the microbiome, each tool tends to be specialized for a specific task. To overcome this limitation, we report on the development of a foundation model pretrained on 13,573 human microbiome metagenomic samples.

We investigated how well a foundation model pretrained on shotgun metagenomic data can **predict host clinical status.**

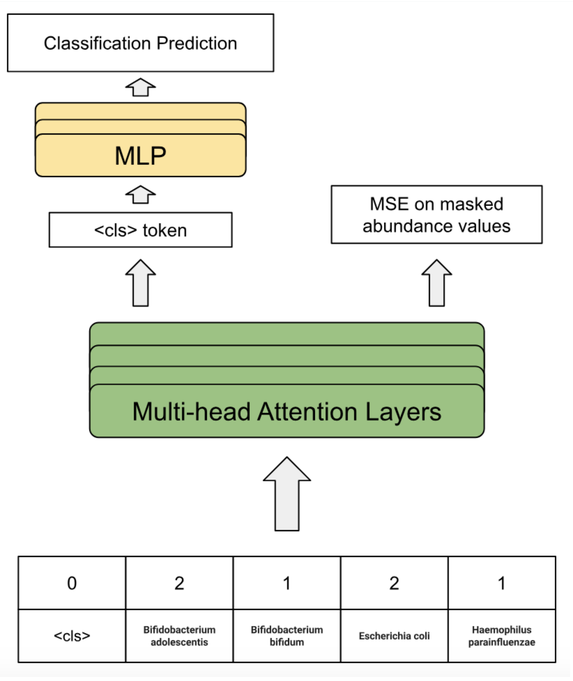
We tested:

1. Binary classification of healthy vs. diseased microbiome samples.
2. Binary and multiclass classification of 33 disease types

2. Dataset

- Compiled a dataset of 14,500 human microbiome shotgun metagenomic samples from ~80 studies.
- Training dataset with 13,573 samples: healthy + 33 diseases
- Validation dataset with 927 samples: healthy + 5 diseases
- MetaPhlAn profiling [F. Beghini et. al, eLife (2021)]

3. ML Model



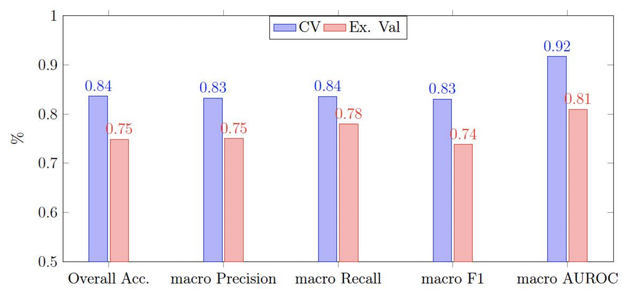
4. Embedding Approach

Organism name	Abundance
Bifidobacterium adolescentis	0.36
Bifidobacterium bifidum	0.04
Escherichia coli	0.50
Haemophilus parainfluenzae	0.10

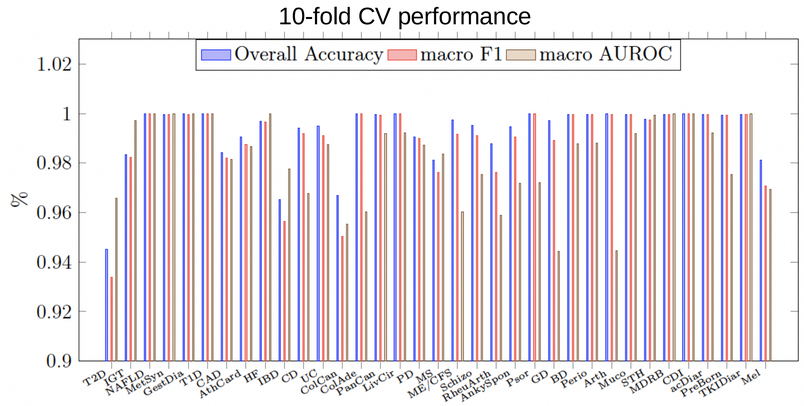
→ Binned abundance vector (with number of bins=2): 2,1,2,1

5. Results

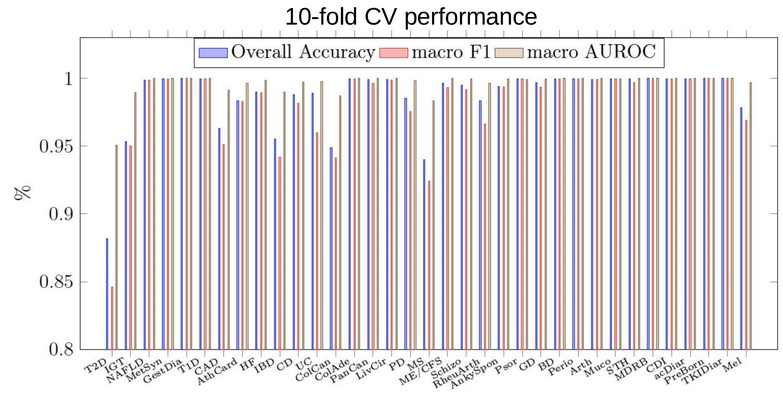
Task 1: Healthy vs. Disease Classification



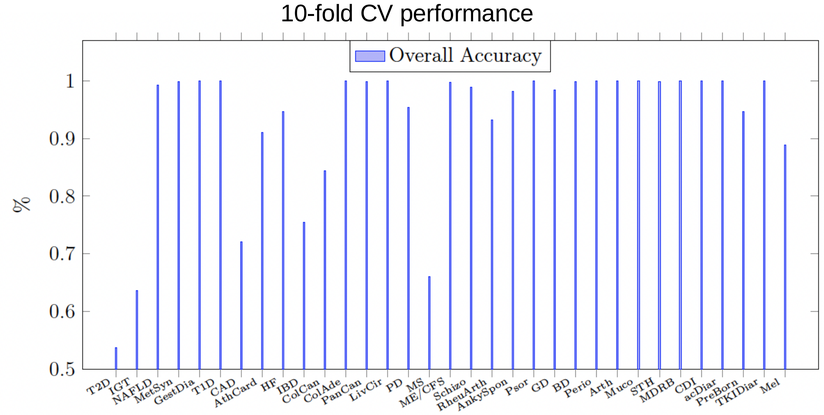
Task 2: Disease X vs. Healthy Classification



Task 3: Disease X vs. All Other Diseases Classification



Task 4: Multiclass Classification of all Diseases



6. Summary and Conclusions

- Developed a foundation model for human microbiome analysis using large-scale metagenomic data.
- Task 1 and Task 4: The model is better able to differentiate diseased samples from other diseased samples than diseased samples from healthy samples.
- Task 2 and Task 3: Diseases with more samples (e.g., T2D) tend to have lower performance, likely from greater intra-class variance.