Multimodal Deep Learning applied to classify healthy and disease states of human Microbiome*

NUT 235 - Final Project

Hermine Tranie

* Lee, S.J. and Rho, M., 2022. Multimodal deep learning applied to classify healthy and disease states of human microbiome. *Scientific Reports*, *12*(1), p.824.

Table of Content

Introduction

- Deep learning methods to **classify healthy and disease states** of the human microbiome:
  - IBD (inflammatory bowel disease)
  - T2D (type 2 diabetes)
  - CRC (colorectal cancer)
  - LC (liver cirrhosis)

- Microbiome plays a **critical role in human health and disease**: modeling how it is affected by different factors can lead to better understanding

- Novel use of deep learning as it is able to **learn complex pattern**s in the data that other methods may miss

Background

- Human microbiome: **trillions of microorganisms** that live in and on the human body and **can impact health** in a variety of ways
- Analyzing microbiome data is **challenging** because it is high-dimensional, noisy, and **highly variable between individuals**
- Deep learning methods can help address these challenges by using multiple layers of neural networks to **learn hierarchical representations of the data**, and by leveraging large amounts of data to improve performance

Data Sourcing

Samples of gut microbiome sequencing data

IBD

NIH Common's Fund HMP Program
- 100 controls
- 100 IBD patients

LC

European Nucleotide Archive
- 83 controls
- 94 LC patients

T2D

NCBI Sequence Read Archive
- 47 controls
- 101 T2D patients

CRC

European Nucleotide Archive
- 60 controls
- 59 CRC patients

Data Pre-Processing

- Filtered out **low-abundance features,** normalized the data, and **scaled the features** to a standard range:
  - For each downloaded raw sample, trimmed of paired-end sequencing reads for quality control

- **Removed:**
  - Low-quality reads (Phred quality score < 20), with Sickle333
  - All reads containing Ns in their sequences
  - Host contaminations by mapping reads to the UCSC human reference genome (GRCh37, hg19, established in February 2009) using Bowtie (ver. 2.3.4.1)34. (taking technological imperfections in extracting gut microbiome into account)
  - Reads with a mismatch and soft-clip length under 10% and 30% of the read length, during mapping results

Input Features

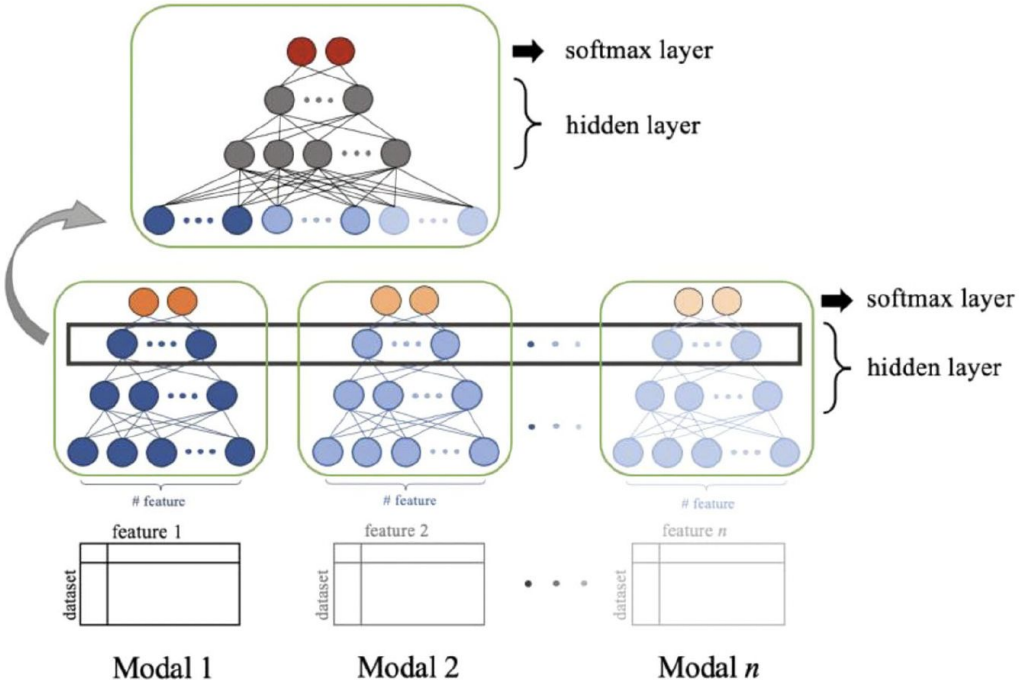
Features		IBD	T2D	LC	CRC
Taxonomic composition	Phylum	12	11	11	11
	Class	20	19	18	17
	Order	26	32	27	26
	Family	52	62	54	52
	Genus	116	141	121	122
	Species	327	388	361	313
Genomic contigs[a]	2 Refs	27	50	23.7	69
	40 Refs	279.3	366	207.3	479.3
Functional proportion		6,147	5,333	7,381	7,220

Deep Learning Methodology

- **Multi-modal convolutional neural network** (MCNN) to classify the microbiome samples:
  - Train each data separately
  - Freeze second to last layer and extract features
  - Concatenate all features and retrain
  - Output result

- Evaluated the **performance of the model** using several metrics, including accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC)

Deep Learning Architecture



A multimodal deep learning model aims for combining features from different modalities.

1) Each feature generated by different methods is first fed to the classifier. The nodes of the last hidden layer are considered as embedded representations of each feature.
2) Embedded representations are concatenated into a new shared representation inheriting original features.
3) Combined feature representation is fed to the classifier for final classification.

Impact of framework on data representativity in small dimensions (t-SNE)

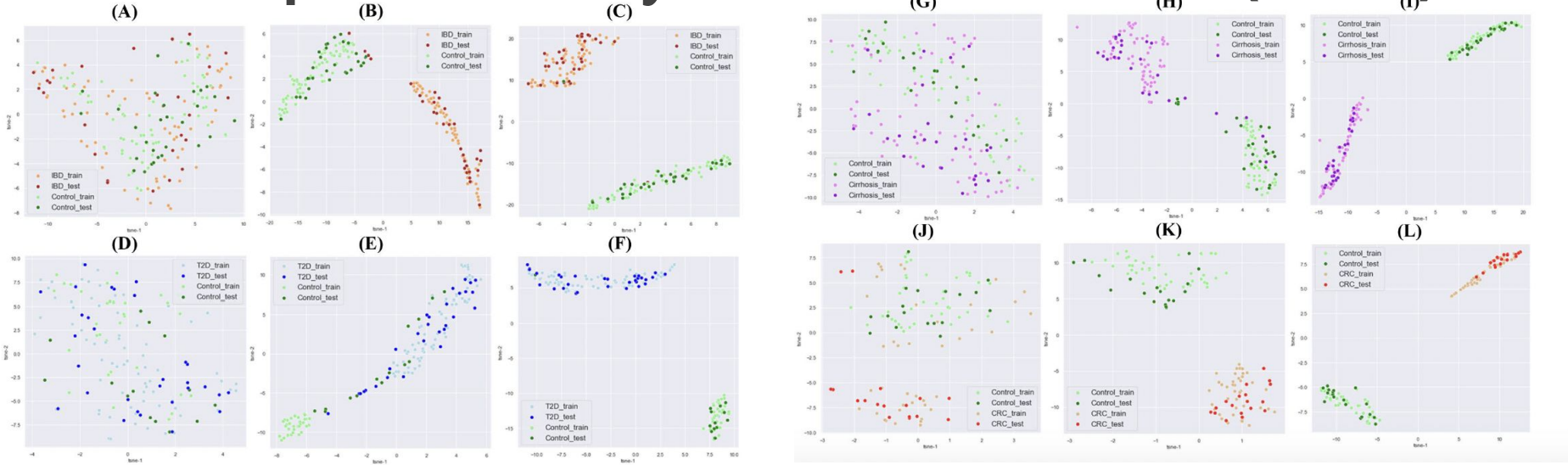
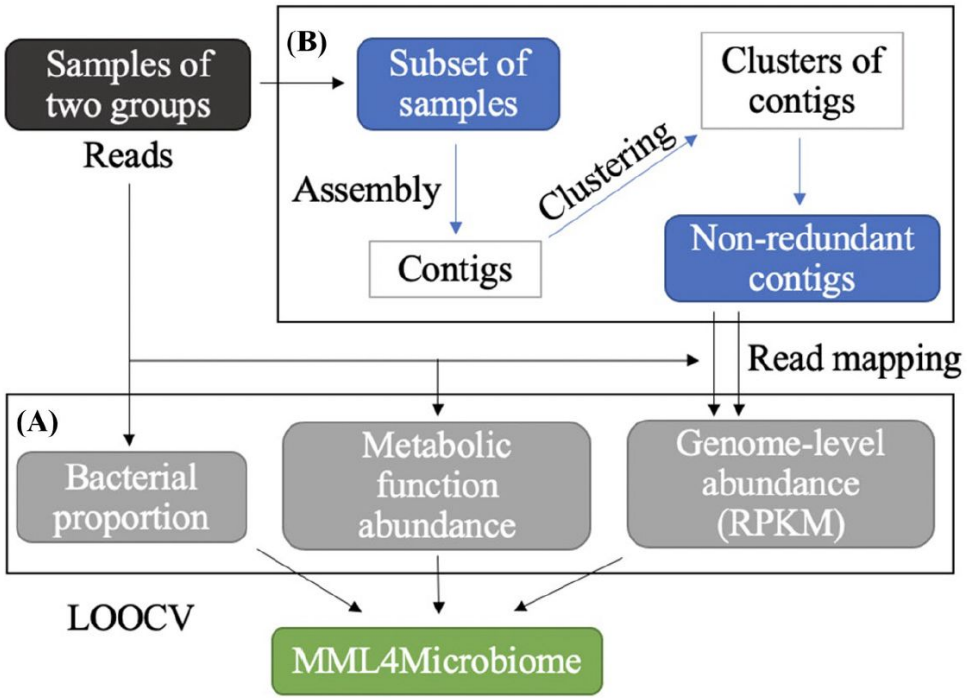


Figure 4. Visualizations of the IBD, T2D, and LC datasets before and after training using t-SNE. Each column, (**A–C**), (**D–F**), (**G–I**), and (**J–L**) was generated with IBD, T2D, LC, and CRC datasets, respectively. (**A,D,G,J**) are from the data before training MDL4Microbiome with all features combined (simply concatenated). (**B,E,H,K**) are from the data in the last hidden layer when the classifier was trained with 70% of the dataset (as light colors). The remaining 30% retained for testing were predicted using the classifier (as dark colors). (**C,F,I,L**) are the result of one-fold of LOOCV. All samples, except for one, in the dataset were used for training, and all samples were included in prediction for visualization.

Pipeline Overview



(A) **Three methods** were used to generate different types of features, viz., conventional taxonomic profiles, metabolic functional features, and genome-level abundance. Different features are fed into the **multimodal deep learning model**. The model was evaluated by the leave-one out cross-validation method.

(B) Specific steps of **extracting non-redundant contigs** of known and unknown microorganisms. A subset of samples is selected randomly. After contigs are assembled with the reads of the selected samples, they are **clustered to collect a set of non-redundant representative contig**s. Entire sample reads are mapped to non-redundant contigs to measure the relative abundance of genomic fragments.

Results

- 98% accuracy for IBD dataset
- Feature importance: Metabolic functions in the T2D dataset
- Comparison of ROC curves across diseases for data separately and combined

- Comparison across classical models:

	RF	XGBoost	PCR	lasso	SVM	Ensemble[a]	MDL4Microbiome
IBD	0.98	0.98	0.99	0.99	0.73	0.99	0.98
T2D	0.68	0.72	0.68	0.72	0.68	0.74	0.76
LC	0.81	0.82	0.83	0.82	0.82	0.84	0.84
CRC	0.94	0.94	0.95	0.96	0.87	0.94	0.97

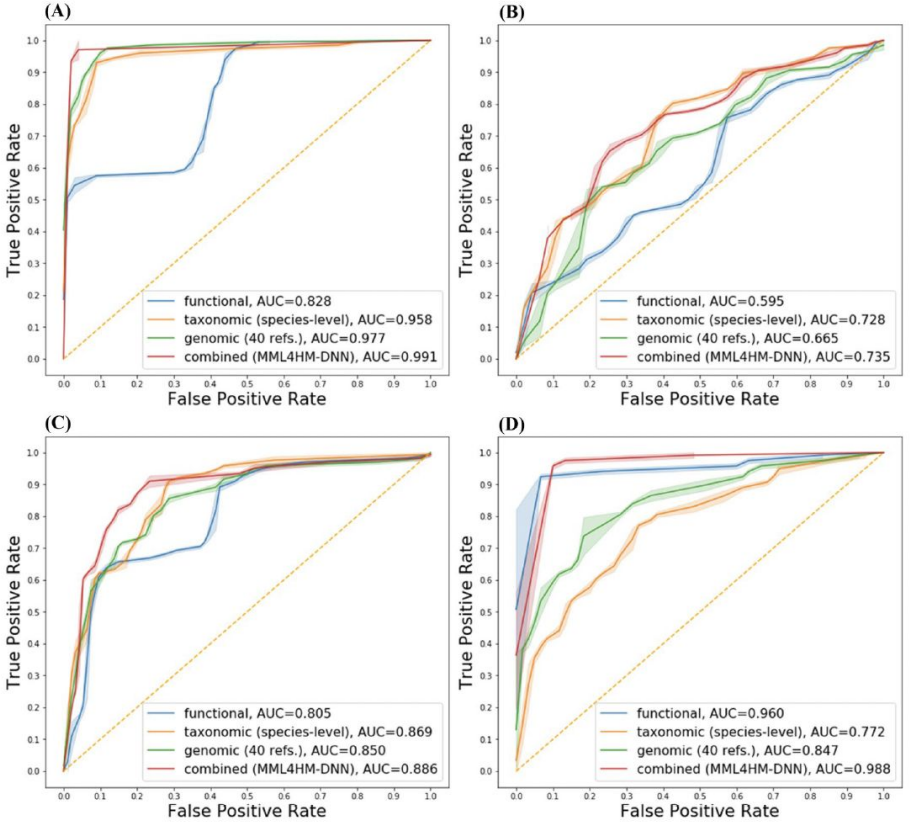


Figure 3. ROC curves and AUCs for MDL4Microbiome with each feature and all features combined. For ROC curves, thresholds were selected as the means between any two consecutive values observed in the data. ROC curves and AUCs for the (A) IBD, (B) T2D, (C) LC, and (D) CRC datasets.

Discussion

- **Limitations:** small datasets, may not generalize well to other populations or disease types
- Potential of **deep learning methods for microbiome research** and may lead to new insights and discoveries in the field

Summarization of the paper

- Deep learning methods can be effective for classifying healthy and diseased states of the human microbiome:
  - Allowed the combination of features of different aspects of microbiomes
  - Resulted in an overall high accuracy of classifying host phenotypes

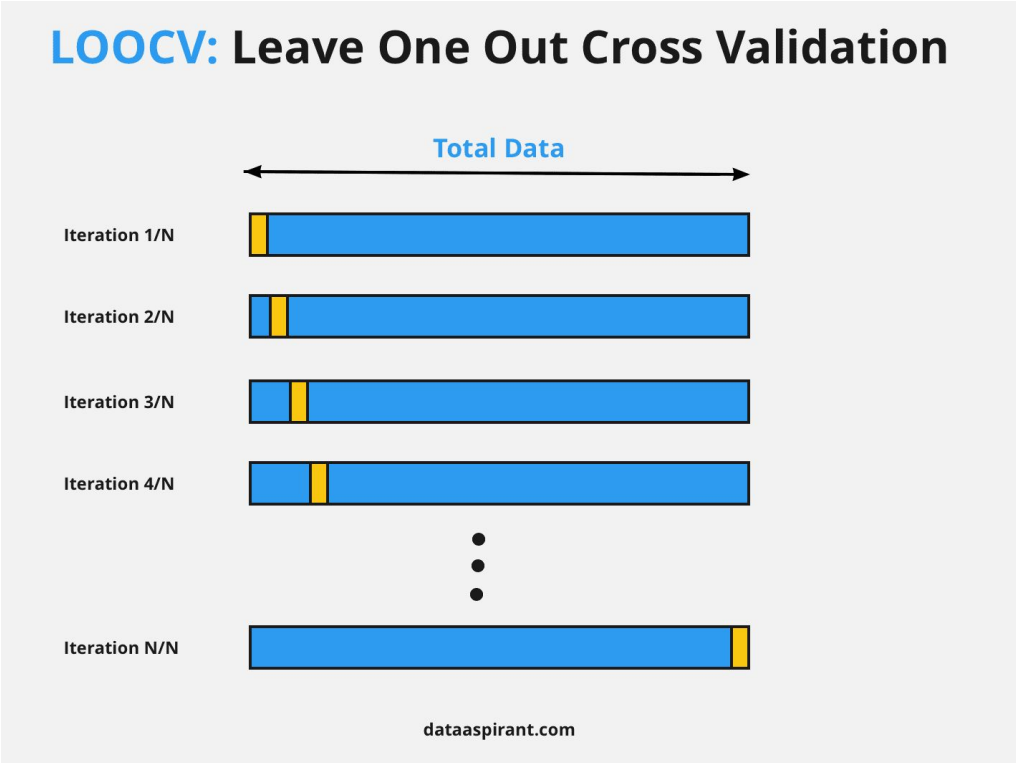
- Lots to explore with deep learning in microbiome research!!

LOO Cross Validation

- High computational cost: re-learn everything n times (number of samples)

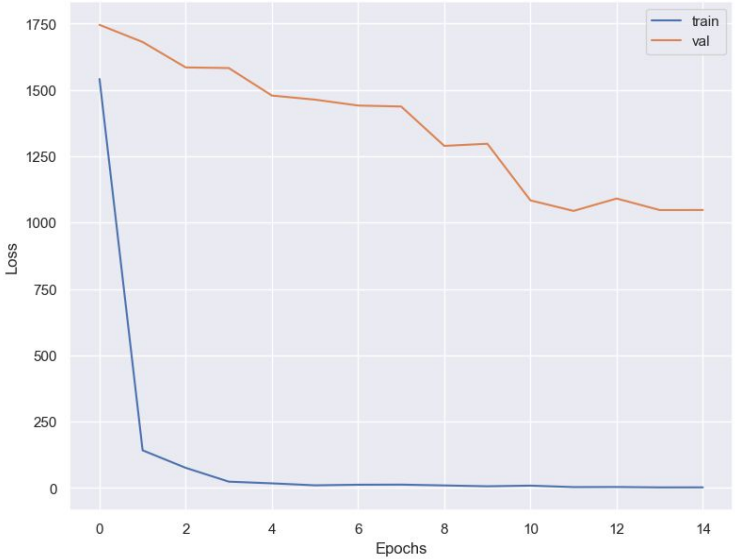
- Predict each instance, by training on all (n-1) instances

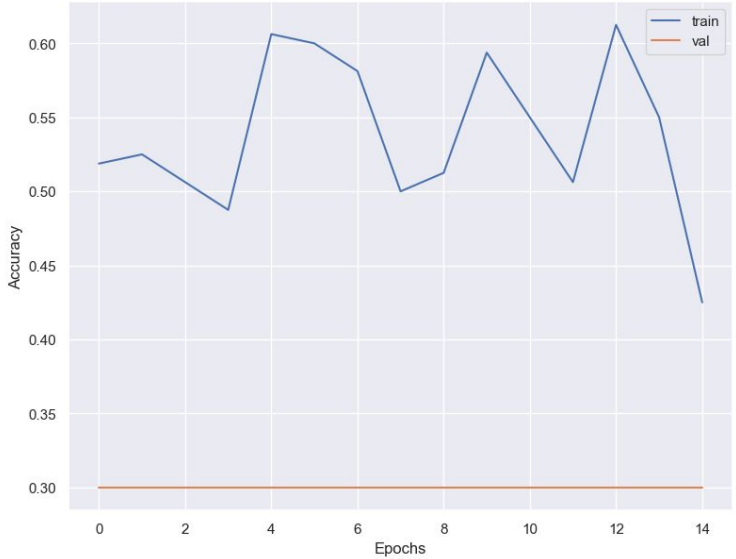
It’s hard to do on one personal machine



Reproducing Results

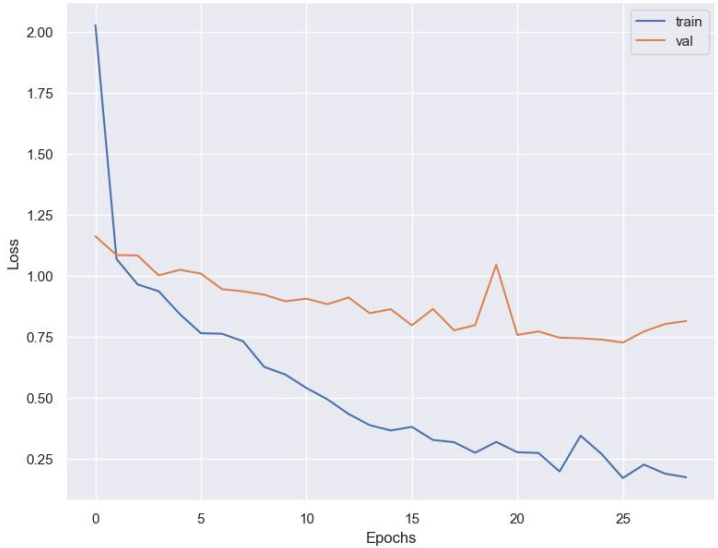
Functional Data

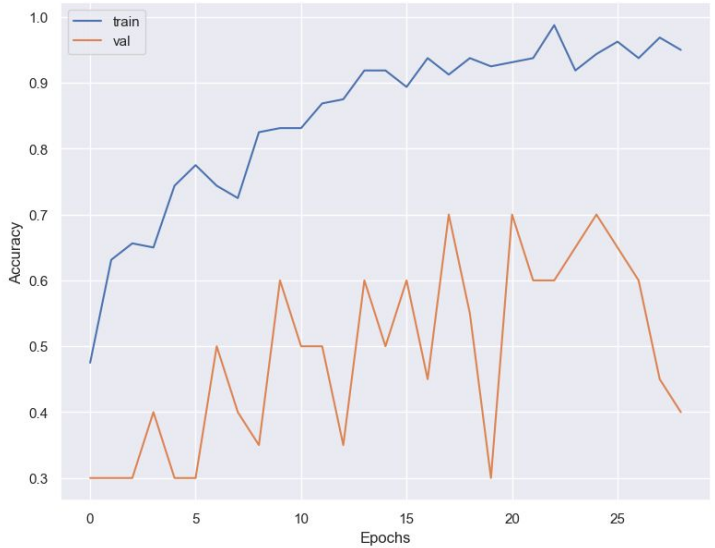




Reproducing Results

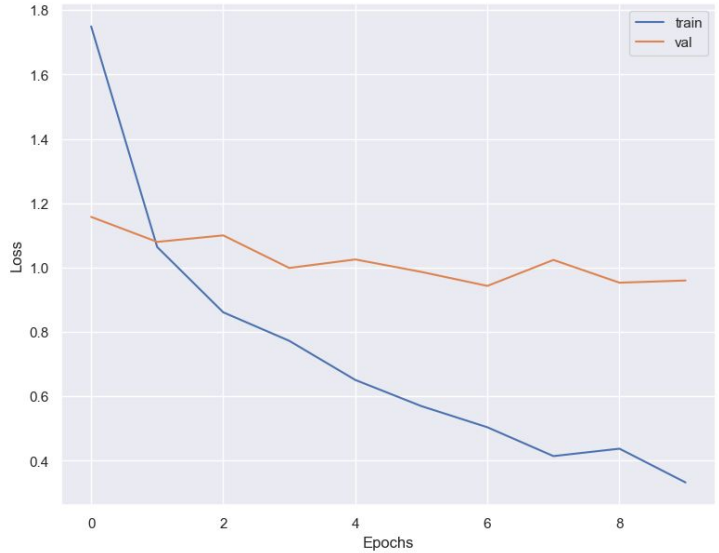
Genomic Data

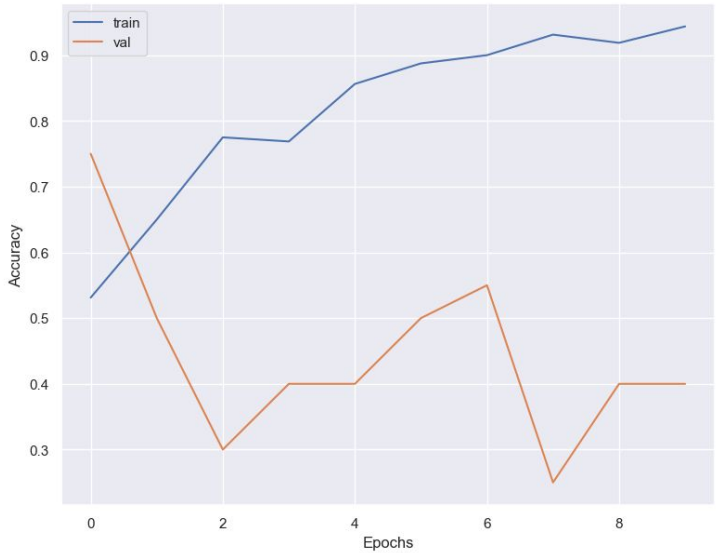




Reproducing Results

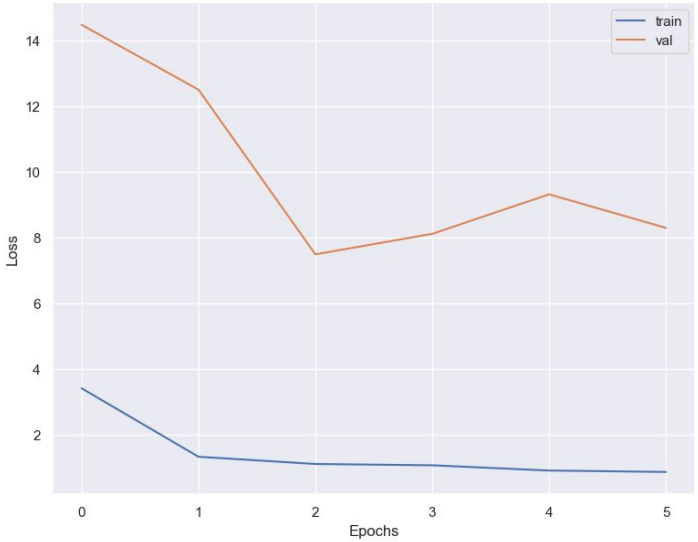
Taxonomic Data

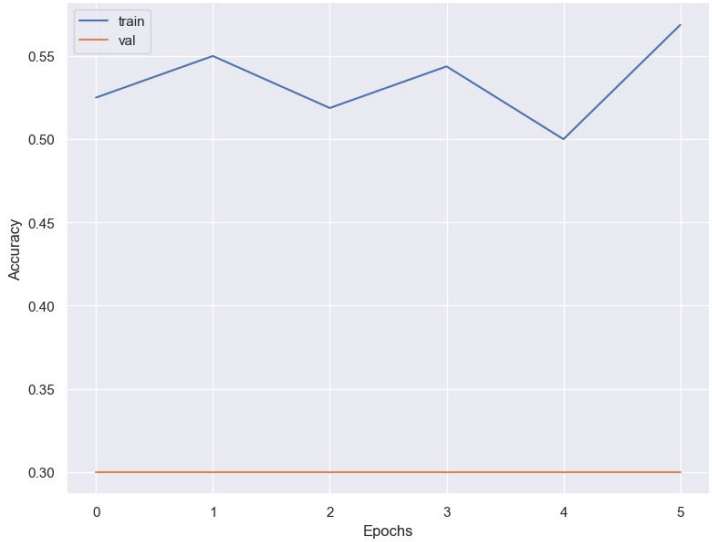




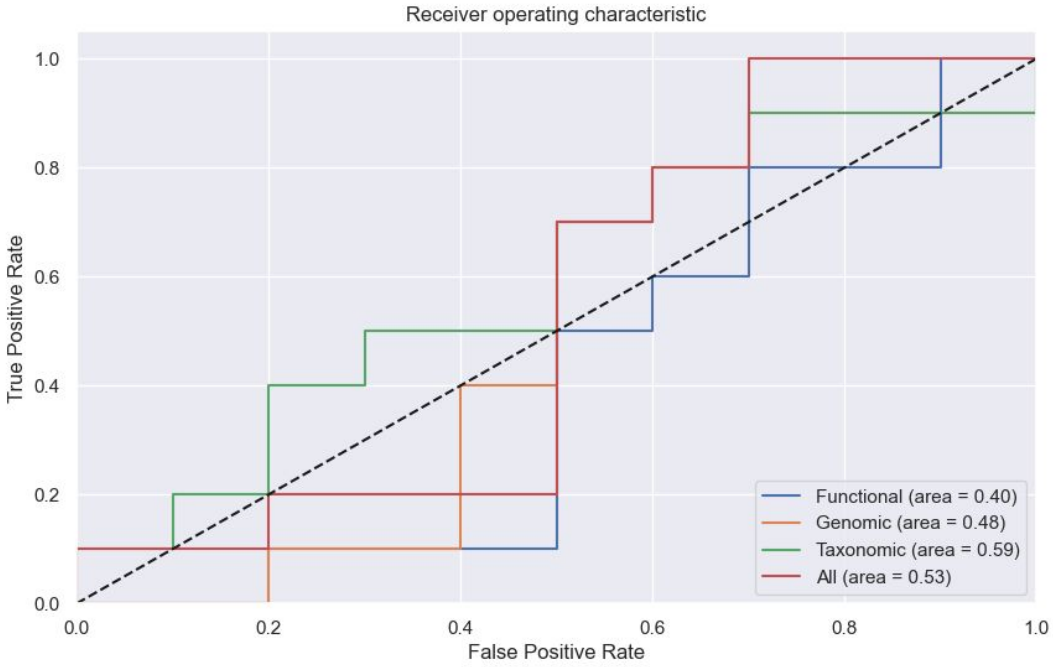
Reproducing Results

Altogether





AUC Curves (without LOOCV)



Conclusion

- Reproducing paper results is not straight forward
  - Out of date/Inexact coding
  - Dataset is small
  - Limited GPU access (training one dataset can take multiple hours on my machine)

- Improved the initial baseline by 6% (from my first run of the paper code without LOOCV) thanks to using regularization techniques:
  - Dropout layer
  - L2 regularization with Dense layers

Post presentation comment/questions

Comments

- T2P: type 2 diabetes is not particularly known to have impact on the gut microbiome, is might be one of the reasons why it's not performing well for that disease classification (in contract IBD is and achieve 98% accuracy)
- Assembly is not very stable (taxonomic data so it might be a reason why results are hard to generalize

Questions

- What is the functional data?
  - metabolic data, 6000 features, sparse and very noisy
- Did you use different architecture?
  - Yes, different architectures for the feature embedding and the final prediction (available on my GitHub)

General impression and insights on difference between classical statistics and Deep Learning

	Classical Statistical Analysis	Deep Learning Methods
What it does?	Univariate + multivariate methods: exploration of relationships between variables	Neural Nets: learn complex patterns in the data
Pros	Useful for identifying differences between groups, exploring correlations between variables, and visualizing complex data sets	Useful for tasks such as classification, prediction, and clustering
Cons	May not capture all complex patterns and interactions in large and complex data sets	May be computationally intensive and require large amounts of data to train

- Classical statistics may be appropriate for **exploring relationships** between microbiome diversity and disease status
- Deep learning methods may be more appropriate for **identifying complex interactions** between microbial taxa and host factors or predicting disease state

Choice of method may depend on **specific research question,** size and complexity of the data, and available computational resources.