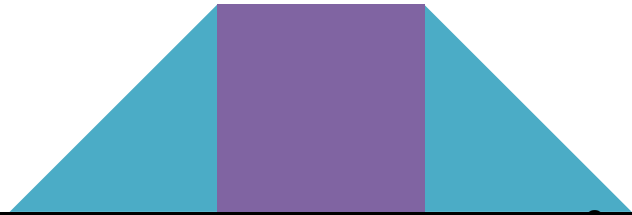
Phenotype Prediction Using Metagenomic Reads

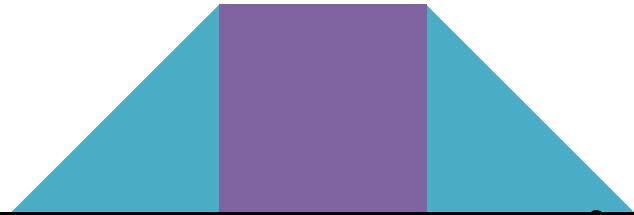
Contents



Problem and Motivation

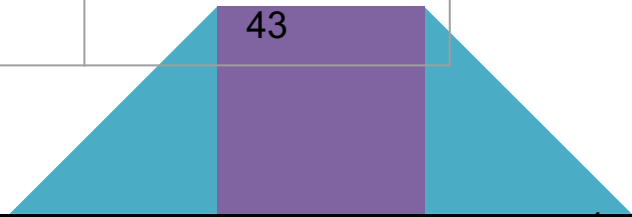
- The human microbiome -
  - Crucial for human health and for the development and maintenance of the immune system and for several metabolic activities
- The potential use of the microbiome as a diagnostic tool is a promising line of investigation

We assessed the prediction power of metagenomic data in linking the gut microbiome with diseases such as obesity, type-2 diabetes



Dataset

Dataset Name	Body Site	Disease	Positive Samples	Negative Samples
Cirrhosis	Gut	Liver Cirrhosis	118	114
Colorectal	Gut	Colorectal Cancer	48	73
IBD	Gut	Inflammatory Bowel Diseases	25	85
Obesity	Gut	Obesity	164	89
T2D	Gut	Type 2 Diabetes	170	174
WT2D	Gut	Type 2 Diabetes	53	43

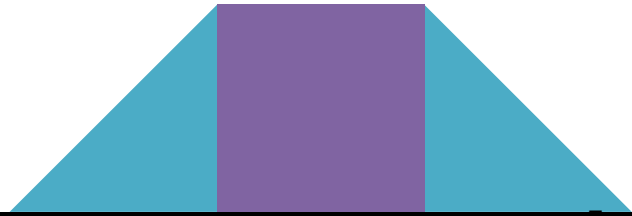


Features

- All metagenomic samples were processed with MetaPhlAn2 [1] for quantitative species- and subspecies-level taxonomic profiling

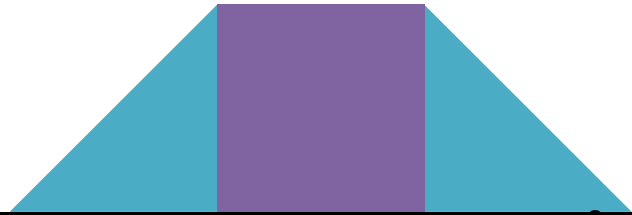
- **MetaPhlAn2** : computational tool for profiling the composition of microbial communities (Bacteria, Archaea, Eukaryotes and Viruses) from metagenomic shotgun sequencing data with species-level and strain-level microbial profiling
  - It relies on reference datasets of marker genes identified from known genomes

- Features used :
  - Species abundances
  - Strain-specific marker abundances



Related Work and Our Take

- [2] Developed a tool, MetAML for metagenomic-based prediction tasks on the same six datasets as ours
- This paper assessed viability of metagenomic data for various tasks:
  - Disease prediction
  - Non-disease classification: gender discrimination, body-site prediction etc
- We focus on disease prediction

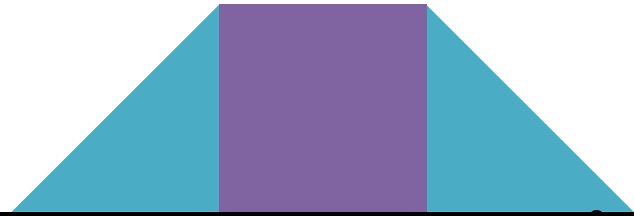


Related Work and Our Take

	MetAML [2]	Our Method
Features	1. Species abundances 2. Strain-specific marker presence	1. Species abundances 2. Strain-specific marker abundances 3. Combination of 1,2
Feature Extraction	MetaphlAn2	MetaphlAn2
Pre-processing	Normalization across entire data	1. Normalization across every sample 2. Binning on normalized values on a logarithmic scale based on the abundance distribution 3. Fill up images
Classifier	SVM, RDF, Elastic Net, Lasso	NN, CNN, RDF, XgBoost, SVM

Method I

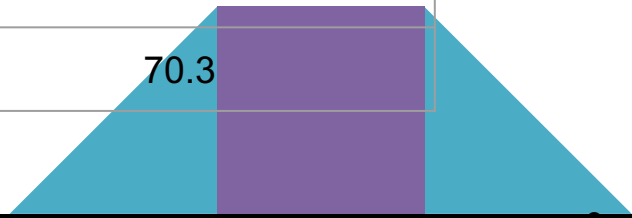
- Features used: Species Abundances
- Pre-processing
  - Normalization: For each sample, species abundance is represented as a real number and the total abundance of all species sums to 1
  - Binning on normalized values on a logarithmic scale based on the abundance distribution
- Classifier
  - RDF
  - XgBoost



Method I

After normalization

Dataset Name	Best Classifier (among RDF, XgBoost)	Accuracy (%)	MetAML [2] (Best classifier %)
Cirrhosis	RDF	87.32	87.7
Colorectal	XgBoost	83.78	80.5
IBD	XgBoost	91.2	80.9
Obesity	RDF	64.94	64.4
T2D	RDF	69.52	66.4
WT2D	XgBoost	79.31	70.3



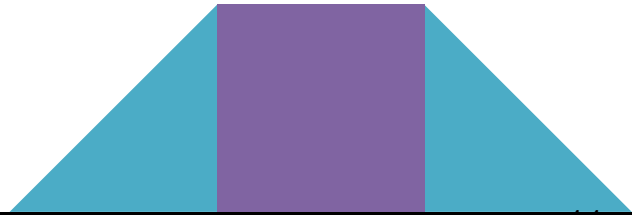
Method I

After binning

Dataset Name	Best Classifier (among RDF, XgBoost)	Accuracy (%)	MetAML [2] (Best classifier %)
Cirrhosis	RDF	88.73	87.7
Colorectal	XgBoost	78.38	80.5
IBD	XgBoost	88.24	80.9
Obesity	RDF	64.94	64.4
T2D	RDF	73.3	66.4
WT2D	RDF	82.76	70.3

Method II [3]

- Features used: Species Abundances
- Pre-processing
  - Normalization: For each sample, species abundance is represented as a real number and the total abundance of all species sums to 1
  - Binning on normalized values on a logarithmic scale based on the abundance distribution. A set of colors is chosen and applied to different bins
  - Fill-up images are created by arranging abundance values into a matrix in a left-to-right order by row. The image is square and empty bottom-left of the image are set to zero (white)
- Classifier
  - CNN



Method II [3]

Fig 1: Fill up image for Cirrhosis dataset sample

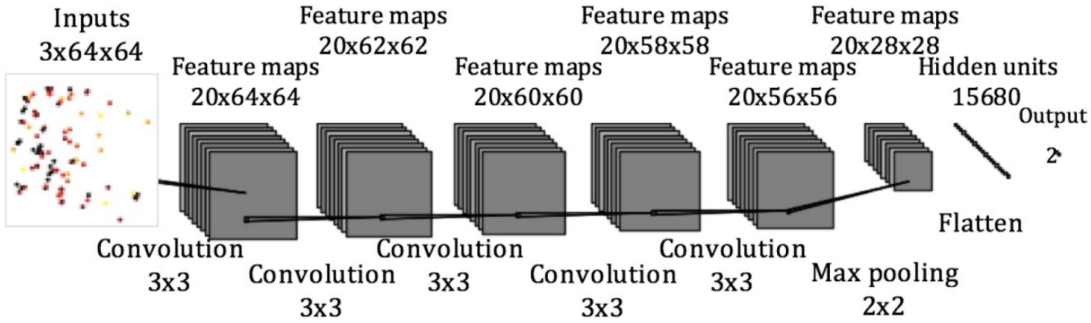


Fig 2: Architecture

Method II [3]

Accuracy for WT2D dataset 82.7%
Accuracies for rest of the datasets were ~ 60%

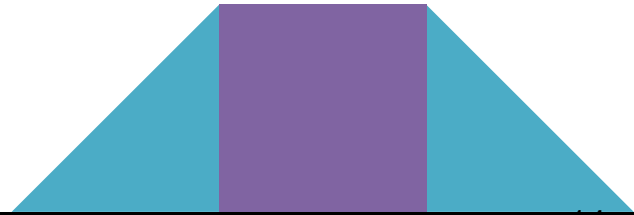
Implication: CNN doesn't generalize well for datasets of small sizes and large variance

Fig 1: Fill up image for Cirrhosis dataset sample

Method II

Let’s try a small 2 layer fully connected neural network

- Features used: Species Abundances
- Trains in a minute
- Pre-processing
    - Normalization: For each sample, species abundance is represented as a real number and the total abundance of all species sums to 1



Method II

Cirrhosis Dataset Results (200 epochs)

Our Method	Best Classifier of MetAML (RDF) [2]	CNN [3]
89.8%	87.7%	89.1%

Confusion Matrix

	Healthy predicted	Disease predicted
Healthy actual	34	3
Disease actual	4	28

Implication: Occam's Razor. Simplicity is the best policy.

Method II

IBD dataset was unbalanced! So to overcome that we fed a random batch size of only 7 samples in each epoch and for 100 epochs. This prevented overfitting and improved generalizability as shown.

Our Method	Best Classifier of MetAML(RDF) [2]	CNN [3]
91.3%	80.9%	83.6%

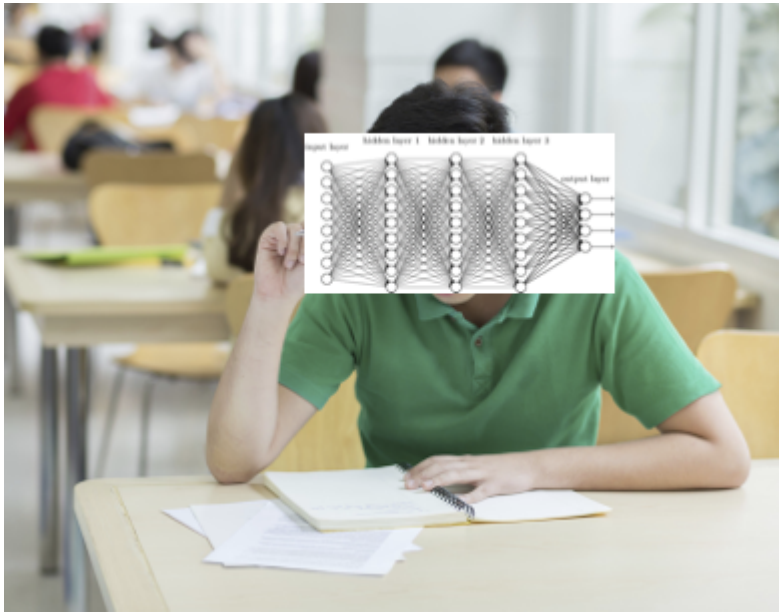
Confusion Matrix

	Healthy predicted	Disease predicted
Healthy actual	14	1
Disease actual	2	6

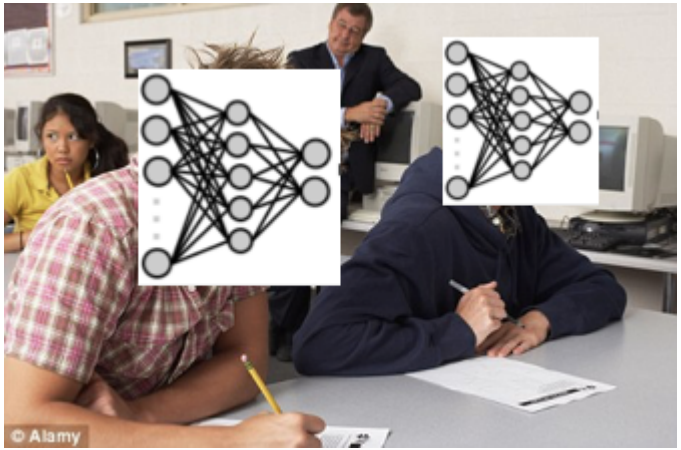
Implication: Overfitting can be prevented by epoch thresholding and keeping randomly sampled small batch sizes

Method II

Obesity Dataset



Does not realize that it can communicate with other Neural Networks to solve the exam, so has to work harder to score !



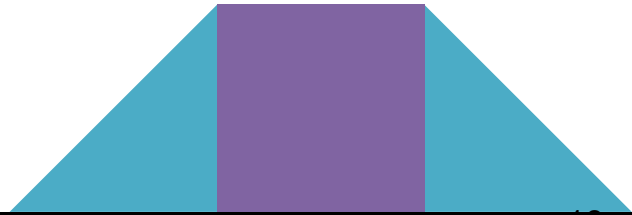
Leverages on the fact that cheating and communication is allowed during the exam, so don't have to work that hard !

Method II

Obesity

The Global minima was at 57% for a conventional neural network architecture.
Implied:

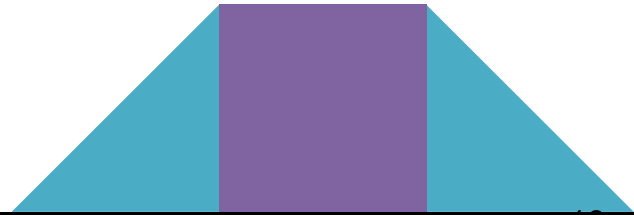
1. Non-Gradient Based Approaches
2. **Induce Generalizability in Gradient Based approaches**
    - Reduce errors by keeping many weak learners
    - "**We want our ensemble to consist of highly correct networks that disagree as much as possible.**" [6]
    - Other theoretical approaches - Unlearning
3. Used an ensemble of 2 shallow networks (trains in 15 seconds)
    - Accuracy of 70.6%



Method III

Hypothesis: Complex diseases are associated with the presence of specific strains or subspecies rather than only species-level abundances.

- Features used: Strain-specific marker abundances
- Dataset: Type 2 Diabetes
- Pre-processing
  - Normalization: For each sample, species abundance is represented as a real number and the total abundance of all species sums to 1
- Classifier:
  - RDF (since it was the best performing)



Method III

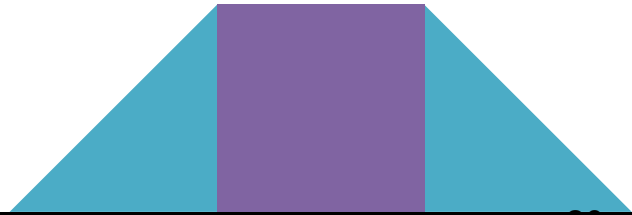
Consistent with this hypothesis, better predictions were obtained from markers:

Accuracy using species abundance (Case 1): 66.9%
Accuracy using Strain-specific marker abundances (Case 2): 69.33%

However,
features (Case 1): 1444
features(Case 2): 127,247

Therefore, time (Case 2)>>>>time(Case 1)
Also, since number of features in Case 2 is pretty high, potentially lot of noise

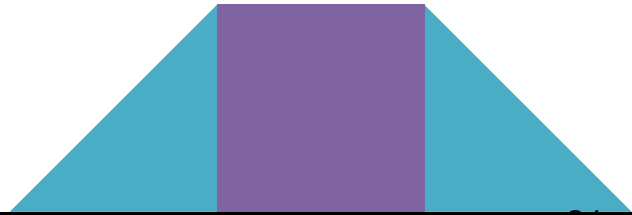
Solution: **Feature selection**



Feature Selection

We used the Feature importances found by RDF for this purpose:

- It is sometimes called "gini importance" or "mean decrease impurity"
- Defined as the total decrease in node impurity (weighted by the probability of reaching that node (which is approximated by the proportion of samples reaching that node) averaged over all trees of the ensemble.



Method III

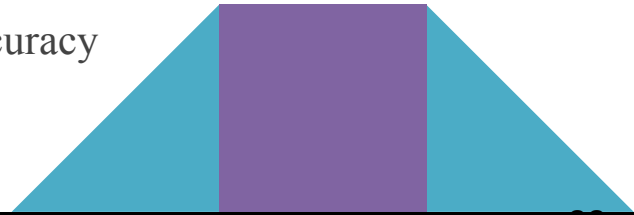
On taking top 50 features for Case 1:

Accuracy: 0.70
F1: 0.69
AUC: 0.77

On taking top 100 features for Case 2:

Accuracy: 0.71
F1: 0.71
AUC: 0.79

Our hypothesis was correct, taking less features actually improved accuracy
Time taken now is substantially less (**~60 times less**)

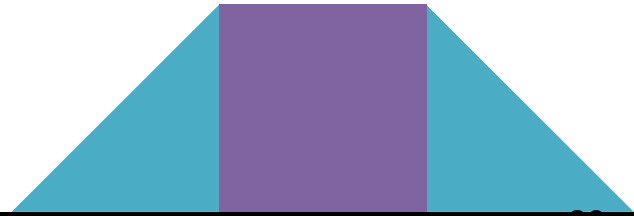


Method III

How about combining Case 1 and Case 2 ?
(Given the reduced features, it is possible to combine them efficiently)

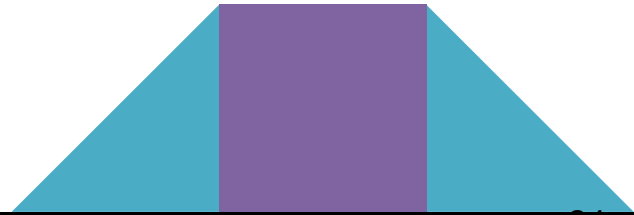
Accuracy: 0.73
F1: 0.73
AUC: 0.79

Yes, even better results

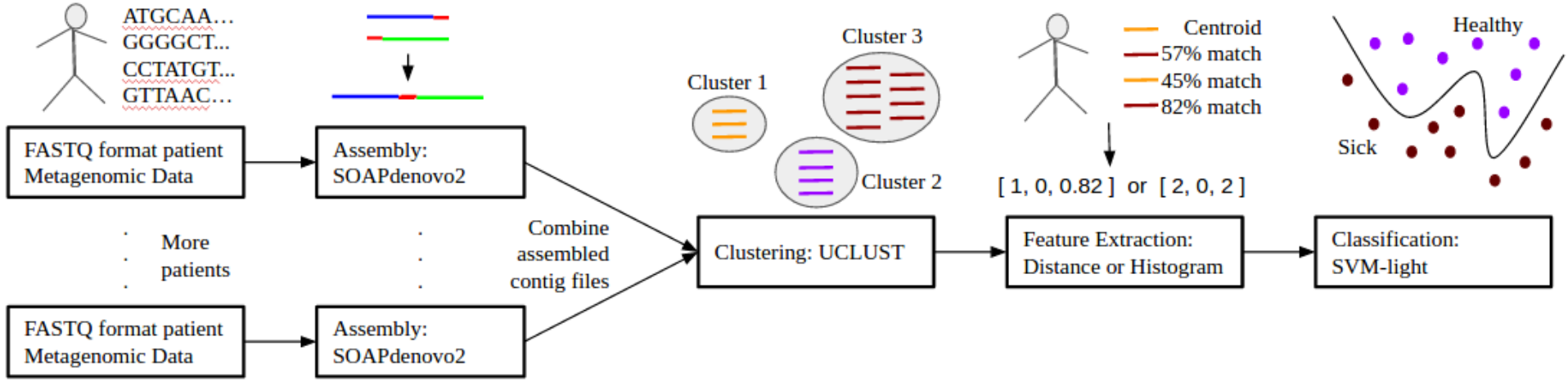


Method IV [4]

- One of the drawback of using features from MetaPhlAn2 (species-level relative abundances and presence of strain-specific markers) is that such feature representation **requires reference datasets of marker genes identified from known genomes.**
- If a large fraction of sequences are novel then information from those sequences which could potentially differentiate a sample with medical conditions from healthy ones will be lost in the process.
- Hence we tried a **de-novo (unsupervised) approach requiring no reference datasets** for taxonomic assignments.



Pipeline

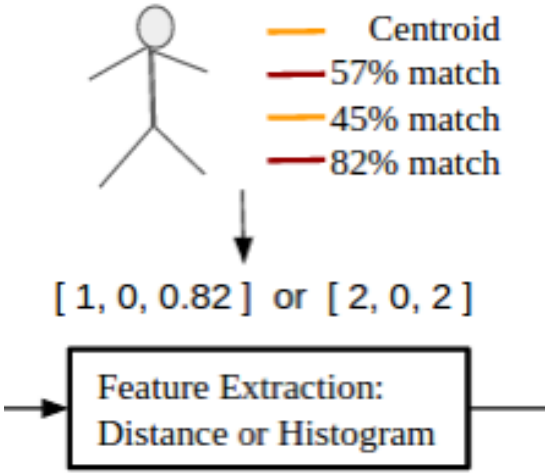


FASTQ Data → Assembly → Clustering → Feature Extraction → SVM Classification

Feature Extraction

- MIL: “Bag of Words” (Amores)*
  - 1. Cluster instances
  - 2. Map to feature vectors
  - 3. Standard classifier

    - Distance Bag of Words (D-BoW)
    - Histogram Bag of Words (H-BoW)

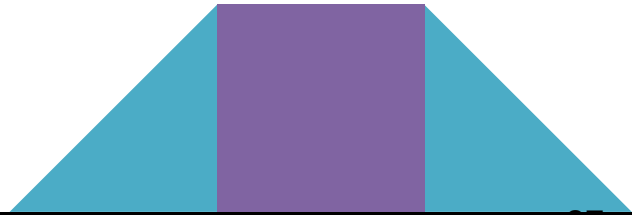


* J. Amores, “Multiple instance classification: Review, taxonomy and comparative study,” Artificial Intelligence, vol. 201, no. 1, pp. 81–105, 2013.

Experiment Setup

- Assembly + Clustering
  - SOAPdenovo2 + UCLUST
  - Run it on Linux server (24-core, 64 GB)

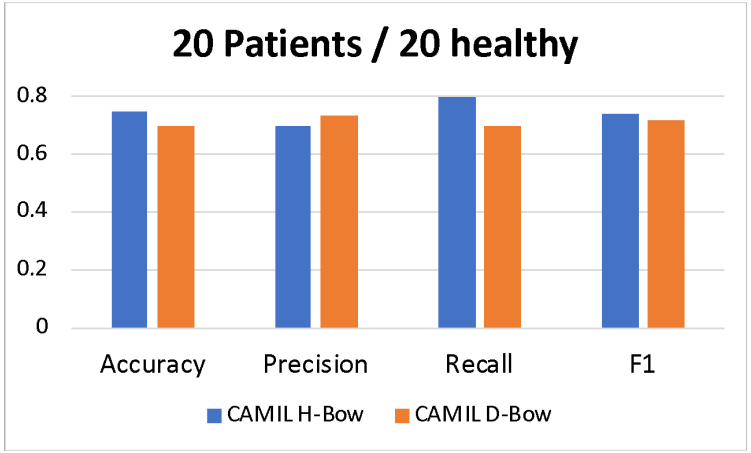
- Feature Extraction + SVM Classification
  - H-Bow, D-Bow
  - Run it on MacOS (Macbook Pro 2014)



Preliminary Results - Diabetes

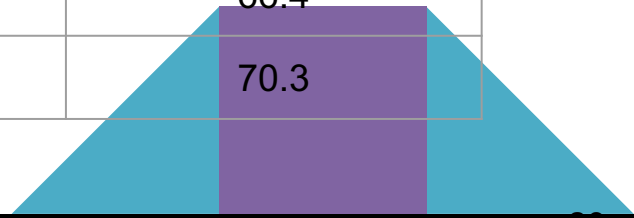
- Small Datasets (70/30 split)
  - 140 samples (70 healthy, 70 diabetes)
  - 100 for training, 40 for testing
- Training: 50 healthy people / 50 diabetes
- Overhead
  - Assembly + clustering (8 hours)
  - Feature extraction + classification (30 mins)

	CAMIL H-Bow	CAMIL D-Bow
Accuracy	0.75	0.7
Precision	0.695652174	0.736842105
Recall	0.8	0.7
F1	0.744186047	0.717948718



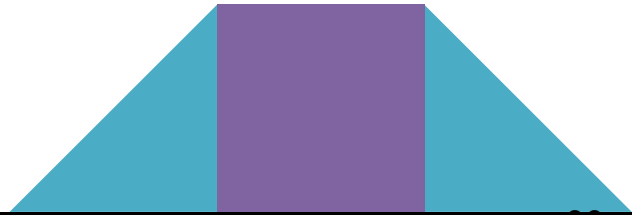
Discussion

Dataset	Our Method (%)	CNN [3] (%)	MetAML [2] (Best classifier %)
Cirrhosis	89.8 (NN)	89.1	87.7
Colorectal	83.78 (XgBoost after normalization)	74.2	80.5
IBD	91.3 (NN)	83.6	80.9
Obesity	70.6 (Ensemble based NN)	66	64.4
T2D	75 (Method IV)	62.6	66.4
WT2D	82.76 (RDF after binning)	58.9	70.3



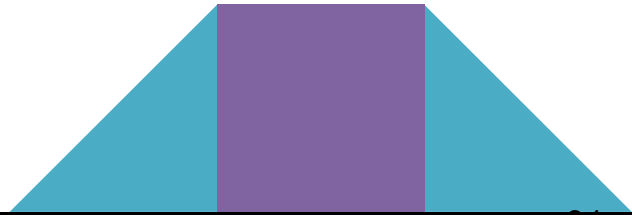
Discussion

1. The dataset size was small and had high variance. Thus, Deep Learning based approaches won't perform too well
2. Random Decision Forests and shallow neural networks perform much better for this data, both in terms of performance and time complexity
3. Ensemble based neural networks should be explored for highly biased datasets
4. Long features like marker's abundances of the order of ~127K for a small dataset would only add more noise, making feature selection important
5. Normalization of dataset across samples helps more than normalization across entire dataset. This will emphasize relative abundances of the strains in a sample
6. Epoch thresholding, random subsampling of small batch sizes helps in controlling overfitting for small datasets



Future Work

- Larger datasets
- Alternatives to reduce feature extraction time
- Perform non-disease classification: gender discrimination, body-site prediction
- Cross-study analysis



References

[1] Truong, Duy Tin, et al. "Erratum: MetaPhlAn2 for enhanced metagenomic taxonomic profiling." Nature Methods 13.1 (2016)

[2] Pasolli, Edoardo, et al. "Machine learning meta-analysis of large metagenomic datasets: tools and biological insights." PLoS computational biology 12.7 (2016)

[3] arXiv:1712.00244

[4] Rahman, Mohammad Arifur, Nathan LaPierre, and Huzefa Rangwala. "Phenotype Prediction from Metagenomic Data Using Clustering and Assembly with Multiple Instance Learning (CAMIL)." IEEE/ACM transactions on computational biology and bioinformatics (2017)

[5] Amores, Jaume. "Multiple instance classification: Review, taxonomy and comparative study." Artificial Intelligence 201 (2013): 81-105.

[6] D. W. Opitz and J. W. Shavlik, “Generating Accurate and Diverse Members of a Neural-Network Ensemble,” Adv. Neural Inf. Process. Syst., vol. 8, pp. 535–541, 1996.

THANK YOU!

Any Questions?