ML4FG Project Midpoint Report: Using Transfer Learning to Enhance Microbiome Analysis

George Austin, gia2105

March 2021

1 Abstract

The goal of this project is to improve microbiome analysis through transfer learning techniques. This is to be done through extending the results of DeepMicro, a study applying various deep learning methods to 6 different fecal gut datasets. So far, we have been able to reproduce their major results. Additionally, we have completed explorations of the datasets and created a plan to optimize our transfer learning approaches.

Figure 1: Dataset Summaries [1]

	Nationality	Year	Positive Samples	Negative Controls
IBD	European	2010	25	85
EW-T2D	European Women	2013	53	43
C-T2D	Chinese	2012	170	174
Obesity	Danish	2013	164	89
Cirrhosis	Chinese	2014	118	114
Colorectal	French	2014	74	47

2 Data Exploration

2.1 Summary

In total, we are working with six datasets, spanning five different prediction tasks: inflammatory bowel disease (1), type 2 diabetes(2)(3), obesity(4), liver cirrhosis(5), and colorectal cancer(6). The datasets are from six distinct studies, all of which provide publicly available metagenomic data. All datasets were processed during the DeepMicro (7) study using MetaPhlAn2 (8) and MetAML (9) to create two representations:

1. the Markers data, which illustrate the strain-level markers that appeared in the samples. These datasets have $O(100,000)$ distinct strains per sample
2. the Abundance data, which show the species-level abundances for each sample. These datasets tend to have around 500 taxa per sample.

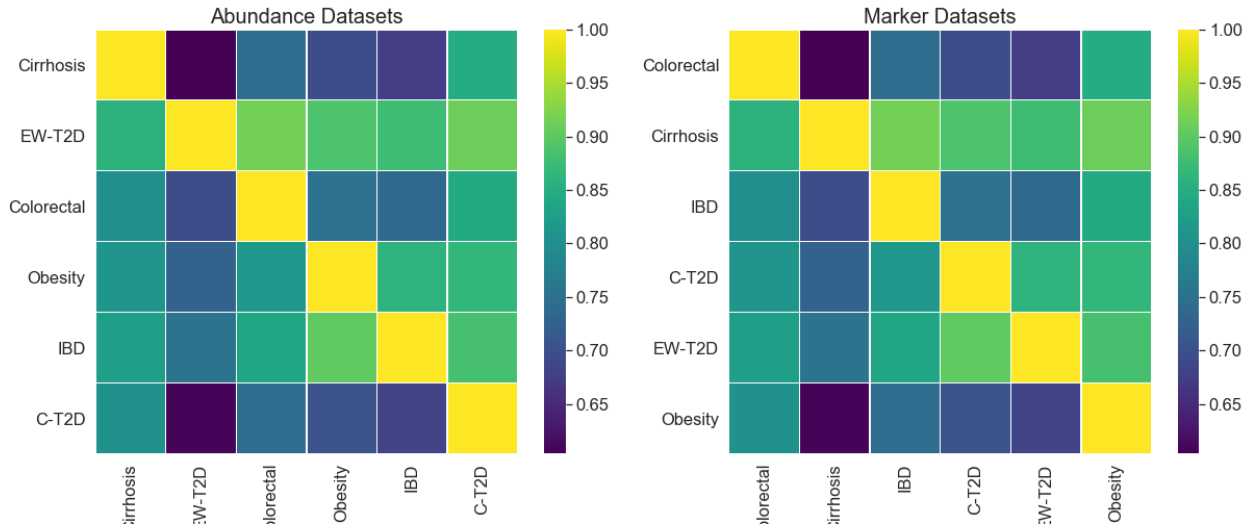
All analyses are performed separately on each dataset, although DeepMicro focus more on the marker datasets, as those high-dimensional cases lead to comparatively better results from their autoencoders.

2.2 Similarities Across Datasets

From our initial analysis, many of the datasets are similar enough to suggest that we can make use of transfer learning techniques like hard parameter sharing (10) and MAML (11). However, there are still some noticeable batch effects in the data, as illustrated by Figures 2 and 3. The PCA plot clearly shows the Cirhhosis and C-T2D datasets forming their own cluster. This might stem from the fact that these datasets originate from China (Figure 1), which could contribute to differences in the underlying microbial samples, or differences in processing techniques. Due to these different batches, it makes sense to start the transfer learning approaches by only considering either the European datasets, or the Chinese datasets. While for our purposes, we don't need to perfectly eliminate the batch effects to get some use out of the transfer learning, we can assume that we won't see many improvements by incorporating data representations that are completely different. For the time being, it does not appear that we will need to do any data transformations to further reduce batch effects, since the PCA representations of just the European data don't produce distinct clusters. Furthermore, it would be better to avoid transformations, if we can help it, so we can isolate the impact of just the transfer learning methods in comparison to DeepMicro's published results.

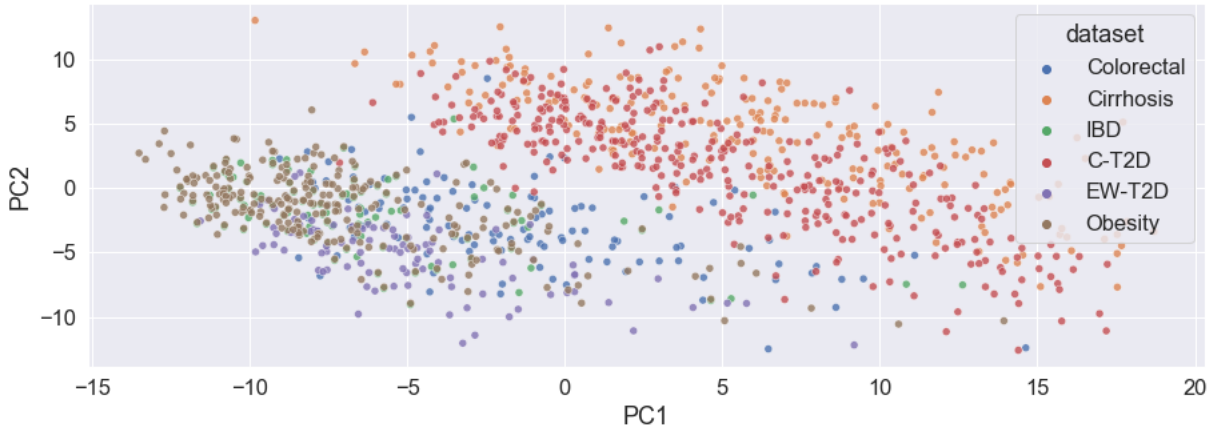
[1] Our Colorectal positive/negative numbers are different from what DeepMicro uses. This is because we classify our adenomas as positive, as these detections might warrant a colonoscopy, even if they are not cancerous. We might switch this back to DeepMicro's definitions.

Figure 2: Percent of Elements Overlapping Datasets



Element (i,j) of each heat map illustrates what percentage of elements from dataset i appear in dataset j. In general, there is a high rate of overlap across groups, with some exceptions. We also see very similar patterns across the marker and abundance datasets.

Figure 3: PCA of Markers Datasets

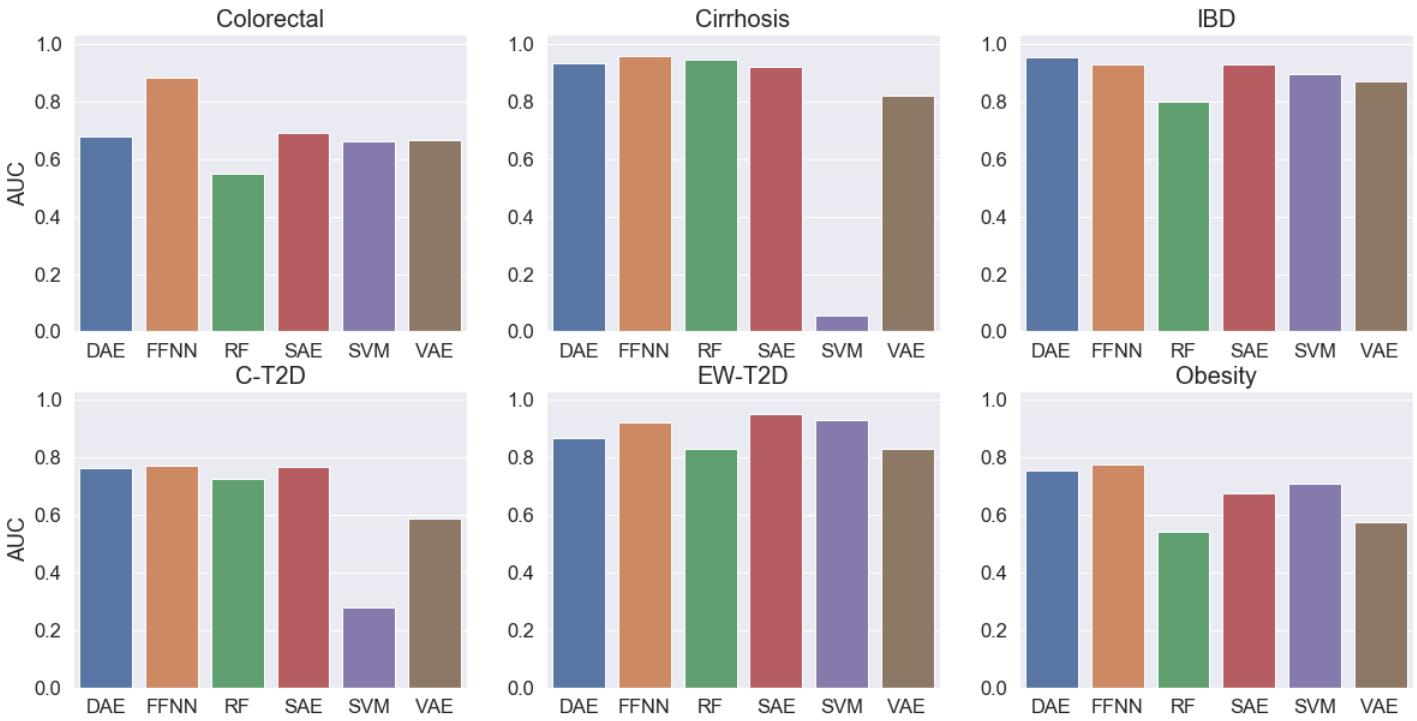


Similarly to Figure 2, we see that the Cirrhosis and the C-T2D datasets are separate from the other four groups. PCOA with Bray-Curtis similarity and abundances datasets give similar results.

3 Baseline Results

So far, we have for the most part been able to reproduce results within the uncertainty ranges reported by DeepMicro (https://www.nature.com/articles/s41598-020-63159-5/figures/2). Figure 4 shows results for our best-performing models of each type, which comes from running Facebook's AX tuning (12), a computationally cheaper approach than DeepMicro's grid search scheme. While we did do the train/valid/test scheme DeepMicro propose, we have so far only run it for one test group per dataset, which could account for some of our discrepancies. Our encoders also do slighlty worse for the colorectal data, which could stem from our classification of adenomas as positive. While this does allow for more potential improvements via transfer learning, we are considering adjusting this to keep our results comparable to DeepMicro's. We also see SVM occasionally fail, which will warrant a few additional tests.

Figure 4: Baseline Marker Results from Tuned Models



One surprising detail from our baseline results is that we often see our deep FFNN model outperform DeepMicro's autoencoder structures. This is unexpected, since it wasn't as big of a focus in the original paper, although it could bode well for this project, since the FFNN structure can lend itself a little more naturally to the transfer learning techniques proposed in this project. While we will still explore hard parameter sharing and MAML for all the neural networks, these baseline results suggest our final results might focus a little more on out FFNN results. For all code, results and other figures obtained so far, see our github repository:
`https://github.com/gaustin15/W4671_project`

4 Next Steps

In general, the project is going well. All the goals for the midway point have been completed, and we have a plan in place to apply the proposed transfer learning techniques. Once those have been implemented, we will have a few options for how to proceed, including exploring transformations to reduce batch effects, and interpreting what signals are easier to detect using our methods, which would involve approaches like LIME (13) and Deeplift (14).

References

[1] Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, et al. 2010. "A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing." Nature 464 (7285): 59–65. https://doi.org/10.1038/nature08821.

[2] Karlsson, Fredrik H., Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed. 2013. "Gut Metagenome in European Women with Normal, Impaired and Diabetic Glucose Control." Nature 498 (7452): 99–103. https://doi.org/10.1038/nature12198.

[3] Qin, Junjie, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, et al. 2012. "A

Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes." Nature 490 (7418): 55–60. https://doi.org/10.1038/nature11450.

[4] Le Chatelier, Emmanuelle, Trine Nielsen, Junjie Qin, Edi Prifti, Falk Hildebrand, Gwen Falony, et al. 2013. "Richness of Human Gut Microbiome Correlates with Metabolic Markers." Nature 500 (7464): 541–46. https://doi.org/10.1038/nature12506.

[5] Qin, Nan, Fengling Yang, Ang Li, Edi Prifti, Yanfei Chen, Li Shao, Jing Guo, et al. 2014. "Alterations of the Human Gut Microbiome in Liver Cirrhosis." Nature 513 (7516): 59–64. https://doi.org/10.1038/nature13568.

[6] Zeller, Georg, Julien Tap, Anita Y Voigt, Shinichi Sunagawa, Jens Roat Kultima, Paul I Costea, Aurélien Amiot, et al. 2014. "Potential of Fecal Microbiota for Early-stage Detection of Colorectal Cancer." Molecular Systems Biology 10 (11): 766. https://doi.org/10.15252/msb.20145645.

[7] Oh, Min, and Liqing Zhang. 2020. "DeepMicro: Deep Representation Learning for Disease Prediction Based on Microbiome Data." Scientific Reports 10 (1). https://doi.org/10.1038/s41598-020-63159-5.

[8] Truong, Duy Tin, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. "MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling." Nature Methods 12 (10): 902–3. https://doi.org/10.1038/nmeth.3589.

[9] Pasolli, Edoardo, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata. 2016. "Machine Learning Meta-Analysis of Large Metagenomic Datasets: Tools and Biological Insights." Edited by Jonathan A. Eisen. PLOS Computational Biology 12 (7): e1004977. https://doi.org/10.1371/journal.pcbi.1004977.

[10] Sebastian Ruder. An Overview of Multi-Task Learning in Deep Neural Networks, 2017; arXiv:1706.05098.

[11] Chelsea Finn, Pieter Abbeel and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks, 2017; arXiv:1703.03400.

[12] Eytan Bakshy, Lili Dworkin, Brian Karrer et at, AE: A domain-agnostic platform for adaptive experimentation; https://konstantinkashin.com/assets/papers/AE-NeurIPS2018.pdf

[13] Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, 2016; arXiv:1602.04938.

[14] Avanti Shrikumar, Peyton Greenside and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences, 2017, PMLR 70:3145-3153, 2017; arXiv:1704.02685.