Pacific Symposium on Biocomputing 25:55-66(2020)

Multiclass Disease Classification from Microbial Whole-Community Metagenomes

Saad Khan[1] and Libusha Kelly[1,2†]

1) Department of Systems & Computational Biology, 2) Department of Microbiology & Immunology
Albert Einstein College of Medicine,
Bronx, NY, USA
[†]*E-mail: libusha.kelly@einstein.yu.edu*

The microbiome, the community of microorganisms living within an individual, is a promising avenue for developing non-invasive methods for disease screening and diagnosis. Here, we utilize 5643 aggregated, annotated whole-community metagenomes to implement the first multiclass microbiome disease classifier of this scale, able to discriminate between 18 different diseases and healthy. We compared three different machine learning models: random forests, deep neural nets, and a novel graph convolutional architecture which exploits the graph structure of phylogenetic trees as its input. We show that the graph convolutional model outperforms deep neural nets in terms of accuracy (achieving 75% average test-set accuracy), receiver-operator-characteristics (92.1% average area-under-ROC (AUC)), and precision-recall (50% average area-under-precision-recall (AUPR)). Additionally, the convolutional net's performance complements that of the random forest, showing a lower propensity for Type-I errors (false-positives) while the random forest makes less Type-II errors (false-negatives). Lastly, we are able to achieve over 90% average top-3 accuracy across all of our models. Together, these results indicate that there are predictive, disease-specific signatures across microbiomes that can be used for diagnostic purposes.

Keywords: Microbiome; Machine learning; Metagenomics

1. Introduction

In the past few years, there has been an immense interest towards developing statistical methods to predict phenotypes, such as disease, from metagenomic sequencing of a microbiome.[1] One of the challenges in achieving this goal is the problem of separating out signals for different diseases from each other. Many studies that have looked for signatures of individual diseases in the microbiome have produced conflicting results,[2] and there is evidence that there are general signatures of dysbiosis common to all diseases.[3] Thus, the standard protocol of comparing samples from a disease of interest against healthy controls, with the goal of identifying features of predictive for that disease, may instead be identifying more general features that signal a diseased or healthy microbiome. This problem can arise if the microbiome signals associated with dysbiosis are stronger than those specific to a given disease. The classifier will then have no mechanism to discriminate between general dysbiosis and the specific signatures of the disease. This is problematic if we want to understand the differences between diseases

on a microbial level and/or make specific diagnoses. For example, in a clinical setting, a classifier for a certain disease that has not been trained against a diverse range of conditions may produce false positives when applied on a patient who has a different disease.[4]

We propose that approaching this problem as a multiclass classification can alleviate this issue by forcing the classifier to find features in the input that are specific for discriminating between a given class and every other class in the dataset. Additionally, this approach allows us to use a larger dataset containing samples from more conditions, potentially alleviating biases due to batch effects. Making accurate predictions becomes much harder in this setting, because output is much more specific (random guesses are now correct with probability $\frac{1}{n}$ rather than $\frac{1}{2}$). Here, we quantify the performance of three machine learning approaches to address this problem: random forests due to their popularity and ease-of-use, deep neural networks as a baseline for a deep-learning approach, and a novel graph convolutional network architecture which incorporates microbiome phylogeny to improve performance over the baseline deep net. We use these models to demonstrate that the multiclass disease classification problem is tractable given the amount of publicly available metagenomic data, and we posit this tractability will only improve over time as more data becomes available.

2. Previous Work

Two previous works have attempted to encode the graph structure of phylogenetic trees in order to enhance microbiome-disease predictions using publicly available data from healthy and diseased patients. Reiman et al. implemented a CNN by embedding phylogenetic trees into $\mathbb{R}^2$ and used two-dimensional convolutional layers to construct a body-site classifier.[5] Fioravanti et al. developed a model to discriminate between subtypes of inflammatory bowel disease (IBD) and healthy by projecting samples into a two-dimensional space using Multi-Dimensional Scaling with the patristic distance between phylogenetic trees as the distance metric.[6] However, both papers mapped phylogenetic data to a Euclidean domain to perform convolutions instead of operating in the original tree topology, as we do in this study.

We also show that this problem is tractable in a multi-class setting, whereas most previous studies have focused on discriminating between individual diseases and healthy. For example, there have been meta-analyses[3,7,8] that have tried to identify disease-specific signatures which generalize beyond individual studies, but the results presented in these papers have all been for disease vs. healthy scenarios. Our major contribution here is to present three machine learning models that can make multiclass disease predictions, including a novel convolutional architecture which exploits the tree structure of bacterial phylogenies.

3. Problem Setup

There are many moving targets that make machine learning in bioinformatics particularly challenging, and one major problem is the paucity of standardized datasets. Data pre-processing, particularly in the relatively new microbiome field, involves numerous components that are each active areas of research, and thus being continually improved; for metagenomics data, this includes new assembly methods, decontamination algorithms, sequencing libraries, annotation methods, reference genomes and so on.[9–11] Additionally, new studies are published

every month, resulting in an ever-increasing catalog of potential data points to utilize for model training. In an effort to promote usage of a 'standard dataset' instead of constructing our own from scratch, we drew our training data from a recently published database containing annotated metagenomic data from multiple studies called curatedMetagenomicData.[12]

3.1. *Dataset Construction*

Disease	**Count**	**Site**	**Studies**
Atopic Dermatitis (AD)	38	Skin	Chng[13]
Adenoma	143	Stool	Thomas,[7] Feng,[14] Hannigan,[15] Zeller[16]
Bronchitis	18	Stool	Yassour[17]
C. difficile Infection (CDI)	33	Stool	Vincent[18]
Colorectal Cancer (CRC)	273	Stool	Vogtmann,[2] Yu,[19] Feng,[14] Zeller,[16] Hannigan[15]
Fatty Liver	94	Stool	Loomba,[20] Feng[14]
Hepatitis B Virus (HBV)	99	Stool	Qin[21] (2014)
Healthy	3808	All	25 Studies
Hypertension	169	Stool	Thomas,[7] Feng,[14] Li[22]
Inflammatory Bowel Disease (IBD)	148	Stool	Nielsen[23]
Impaired Glucose Tolerance (IGT)	49	Stool	Karlsson[24]
Infectious Gastroenteritis	20	Stool	David,[25] Yassour[17]
Metabolic Syndrome	50	Stool	Vrieze[26]
Otitis	107	Stool	Yassour[17]
Periodontitis	48	Oral	Shi[27]
Psoriasis	74	Skin	TettAJ[28]
Rheumatoid Arthritis (RA)	194	Stool	Chengping[29]
Type 1 Diabetes (T1D)	55	Stool	HeintZ-Buschart,[30] Kostic[31]
Type 2 Diabetes (T2D)	223	Stool	Qin[32] (2012), Karlsson[24]

Table 1: Overview of dataset samples.

We constructed a dataset containing 5643 samples with 4885 from stool, 403 from skin, 254 from oral cavity, 93 from nares (nasal cavity), and 8 from maternal milk (healthy babies from Asnicar et al.[33]) by including diseases that had at least 15 unique samples (table 3.1). One of the challenges we faced is that some samples have multiple disease labels due to the way the original studies were run. We approached this problem as a multiclasss (one correct label) as opposed to a multilabel (k possible correct labels) problem and thus sought to avoid conflicts due to multiple labeling. Multiple labeling was present in four of our disease sets: Atopic Dermatitis (atopic rhinitis (28), asthma (12)); *C. difficile* (pneumonia (15), cellulitis (2), osteoarthritis (1), ureteral stone (1)); Adenoma (fatty liver (28), hypertension (19), Type 2 Diabetes (6), Hypercholesterolemia (2), metastases (1)); and Hepatitis B (Schistosoma (1), Hepatitis E (7), Hepatic encephalopathy (2), Hepatitis D (5), Wilson's disease (1), Cirrhosis

(97), Ascites (48)). Additionally, depending on how an individual study was annotated, some samples in the original dataset have no disease label or are instead labeled as "control". We chose to be conservative in our construction and only included as "healthy" samples where the disease column in the curatedMetagenomicData database was "healthy". Thus, we define 'healthy' in our study as those samples coming from patients who were explicitly considered to be healthy in the original study that the sample came from.

3.2. *Graph Convolutional Neural Networks*

Convolutional neural networks (CNNs) have been extremely successful in the field of machine-vision.[34] Intuitively, their effectiveness derives from their ability to encode geometric properties of images such as translation-invariance in order to learn a better representation of the data.[35] In recent years, there has been interest in developing neural architectures that can capture analogous symmetries on non-Euclidean domains such as graphs and manifolds in order to extend the success of CNNs to those domains.[36]

There are many architectures for doing this, and we chose to use the method outlined by Kipf and Welling,[37] which is a computationally simple method to generalize the convolution operation to graphs (Figure 1). The method falls into the category of spectral methods, which model convolutions as multiplication by a filter operator g_θ in the Fourier domain against input $x \in \mathbb{R}^n$, where g_θ is a diagonal matrix with parameters $\theta \in \mathbb{R}^n$. The multiplication takes place with respect to the Fourier basis of eigenvectors U of the graph Laplacian $L = I_N - D^{-\frac{1}{2}}AD^{\frac{1}{2}}$, where D is the diagonal matrix of vertex degrees, and A is the adjacency matrix of the graph (in our case, A is the adjacency matrix of the phylogenetic tree used in our model). In Euclidean domains, U is the Eigenbasis of the Laplacian operator in $\mathbb{R}^n$, which is the standard Fourier basis $\{e^{2\pi ik\cdot x} : k \in \mathbb{N}^n, x \in \mathbb{R}^n\}$.

$$g_\theta \star x = U g_\theta U^\top x$$

Instead of considering all Fourier modes, we choose a cutoff K and use the first K Laplacian Eigenvectors when constructing U in order to simplify computation. However, this operation is still very expensive on graphs because there is no general analogue of Fast Fourier Transforms outside of a Euclidean domain (meaning we would have to perform direct matrix multiplications). The method of Kipf and Welling overcomes this problem by setting $K = 1$ and reformulating the above operation to get

$$g_\theta \star x \approx \theta(I_N + D^{-\frac{1}{2}}AD^{\frac{1}{2}})x$$

The choice of cutoff specifies that all nodes up to K edges away from a given node contribute to the output of the convolution operator at that node. Thus, $K = 1$ implies that the output at each node is a function of that node and its immediate neighbors. By stacking layers of this form together (after applying a non-linearity at each level), we can integrate information from increasingly farther nodes. Each individual convolution layer can thus be written as

$$x' = \sigma(g_\theta \star x)$$

for some non-linear activation function σ. Lastly, this framework can be extended to incorporate more than one input / output channel per layer by adding an additional channel dimension to each of these parameters (see original paper for details).

3.3. *Models*

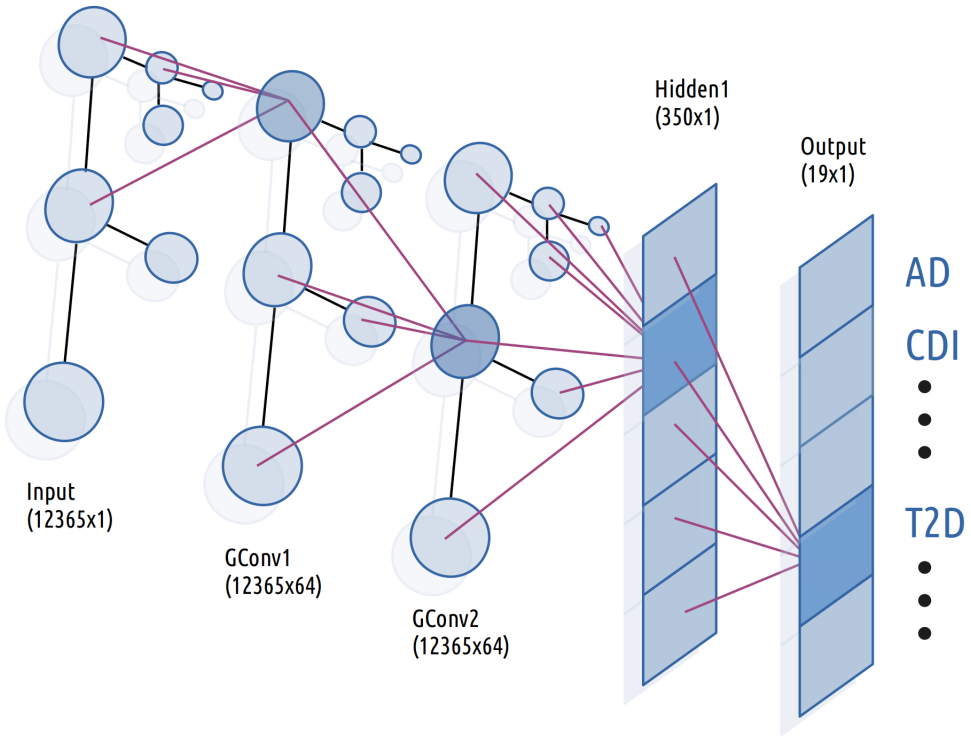

Fig. 1. Architecture of Graph Convolutional Classifier. Purple lines indicate the flow of information from the previous to the highlighted neuron in each layer. In the convolutional layers, each neuron receives information only from its immediate neighbors in the preceding layer.

We implemented three types of classifiers for comparison: a feed-forward deep neural network (DNN), a graph convolutional neural network (GCN), and a random forest (RF). We constructed the GCN model by stacking two convolutional layers with 64 channels each followed by a fully-connected linear layer with 350 nodes. The DNN model consisted of two fully connected linear layers with 1000 and 350 neurons respectively. We used exponential linear units (eLU) as our activations between each layer in either model and a sigmoid activation at the top level for classifications.[38] Additionally, the GCN model was initialized with the adjacency matrix of the phylogenetic tree corresponding to the taxa present in our 12365-dimensional input vector. All three classifiers took such a 12365-dimensional abundance vector as input and produced a 19-dimensional output (number of classes). We implemented our neural networks using PyTorch along with the PyTorch Geometric library for the GCN components.[39,40] Lastly, our random forest model was constructed using the default settings from the random forest module in the Python package Scikit-learn[41] except for the number of trees, which was set to 1000. We settled on these configurations after manual experimentation. We found that increasing the number of trees improved generalization performance of the random forest up to a certain point, and the size of hidden layers in our networks did not make a dramatic difference in performance.

3.4. *Training*

The biggest challenge in this study was dealing with the extreme class imbalance caused by more than half of the samples in our dataset coming from healthy patients and many diseases having only a few dozen samples. Resampling the dataset to artificially balance it is one way to deal with this problem. Many resampling methods have been shown to perform similarly, so we used a simple oversampling protocol.[42] For each experiment, we oversampled the training set by computing the size of the largest class (healthy) and randomly resampling from every class until each class had the same number of samples.[43]

We augmented this by assigning each class a weight of $1 - \frac{1}{n}$, where n is the number of classes, and training our neural networks in a one-vs-all manner for each class (using a binary cross-entropy loss function).[44] To compute the top prediction of the classifier, we ran the outputs of the network through a softmax function and returned the index of the highest class. This is a commonly used technique in multiclass classification, because it reduces the difficult problem of discriminating between n classes to n easier problems of discriminating between two classes.[45] We used a $1 - \frac{1}{n}$ weight in order to address a new class imbalance problem that arises in this setting; if there are m examples of each class in the dataset after oversampling, then each binary classifier will see m positive examples and mn negative examples. A $1 - \frac{1}{n}$ weight magnifies the importance of the positive samples for each class to compensate.

Next, we trained our classifiers. We used a 70-30 split between training and test set, generated a new split for each training run, and ran 30 iterations of the GCN model, 30 of the DNN, and 20 of the random forest. Data preprocessing, ingest, and analysis for our neural networks was performed using Fireworks, a PyTorch-based library that we previously developed to facilitate common machine learning tasks.[46] The GCN and DNN were trained using the Adam optimizer with $2 * 10^{-5}$ and $1 * 10^{-4}$ respective learning rates, 40 and 100 respective batch sizes, weight decay parameter set to 1, binary cross-entropy loss, and an early stopping condition when the loss dropped to 2. We performed an analogous procedure with our random forest classifier by assigning the same class weights and using the same over-sampled training set as with the neural network.

4. Results

We evaluated our classifiers by measuring their accuracy, AUC (area-under-receiver-operating-characteristic (ROC)), and AUPR (area-under-precision-recall-curve (PRC)). The average accuracy varied widely by disease and classifier, indicating the difficulty of this problem given the number of positive samples for each category (Table 2). In general, the convolutional net performed on par with or better than the deep net in terms of accuracy, AUC, and AUPR, implying that the graph structure of the data is useful for making classifications. Both of these models had higher accuracy than the RF for most diseases. The RF excelled in the healthy vs. disease task, producing an impressive 99% accuracy. Because healthy was also the largest component of the test set (66.9% of samples on average), it was responsible for the bulk of the RF's weighted accuracy measurement. However, overall accuracy can be biased by the distribution of the test set (ie. if there were more samples for the classes that a given classifier excelled at, then that classifier would benefit). When we weighted each class equally, then the

	Accuracy (%)		
Disease	**GCN** (n=30)	**DNN** (n=30)	**RF** (n=20)
Healthy	91 ± 1	87 ± 1	$\mathbf{99 \pm 0}$
Colorectal Cancer (CRC)	$\mathbf{36 \pm 2}$	$\mathbf{35 \pm 2}$	19 ± 2
Type 2 Diabetes (T2D)	$\mathbf{39 \pm 3}$	$\mathbf{40 \pm 3}$	08 ± 1
Rheumatoid Arthritis (RA)	73 ± 4	$\mathbf{79 \pm 2}$	$\mathbf{80 \pm 3}$
Hypertension	$\mathbf{56 \pm 3}$	$\mathbf{55 \pm 3}$	$\mathbf{56 \pm 3}$
Inflammatory Bowel Disease (IBD)	$\mathbf{40 \pm 3}$	$\mathbf{40 \pm 3}$	07 ± 1
Adenoma	$\mathbf{15 \pm 2}$	13 ± 2	04 ± 1
Otitis	$\mathbf{12 \pm 3}$	$\mathbf{13 \pm 2}$	00 ± 0
Hepatitis B Virus (HBV)	62 ± 3	59 ± 3	$\mathbf{80 \pm 4}$
Fatty Liver	$\mathbf{32 \pm 4}$	22 ± 2	02 ± 1
Psoriasis	$\mathbf{67 \pm 4}$	$\mathbf{63 \pm 4}$	51 ± 5
Type 1 Diabetes (T1D)	$\mathbf{37 \pm 7}$	$\mathbf{41 \pm 7}$	25 ± 6
Metabolic Syndrome	$\mathbf{52 \pm 6}$	$\mathbf{50 \pm 4}$	08 ± 3
Impaired Glucose Tolerance (IGT)	12 ± 3	$\mathbf{18 \pm 3}$	00 ± 0
Periodontitis	$\mathbf{93 \pm 2}$	$\mathbf{93 \pm 3}$	$\mathbf{93 \pm 3}$
Atopic Dermatitis (AD)	$\mathbf{80 \pm 6}$	$\mathbf{78 \pm 6}$	$\mathbf{82 \pm 6}$
C dificile infection (CDI)	$\mathbf{67 \pm 7}$	53 ± 5	54 ± 9
Infectious Gastroenteritis	$\mathbf{08 \pm 4}$	$\mathbf{09 \pm 5}$	01 ± 2
Bronchitis	$\mathbf{02 \pm 3}$	$\mathbf{05 \pm 4}$	$\mathbf{0 \pm 0}$
Average	46 ± 27	45 ± 12	35 ± 14

Table 2: Percent accuracy by disease for each model.
Boldface indicates the model(s) with the highest score in the category.

GCN had the highest average accuracy (46% vs 34.6% for random forest and 44.9% for DNN), indicating that it was more accurate across a broader range of diseases than the RF model.

Next, we computed ROCs, which plot the tradeoff between true-positive-rates and false-positive-rates for a model[47] (Table 3). This statistic is important because it is invariant to class distribution, and is thus useful for models that try to rule-in a diagnosis, because a high AUC (area-under the ROC plot) score implies that the model can achieve a high true-positive rate while generating few false-positives. We generated ROCs for our neural nets by varying the bias threshold of the final layer and evaluating their true positive and false positive rate on the test set, and for the random forest by using the `sklearn.metrics.roc_curve` function on the test set predictions. We found that the GCN model had higher or statistically equivalent AUCs across all labels than both the random forest and the deep net. In particular, we achieved an average AUC of 89.5% for T2D, which previous studies have found to be a particularly challenging task[8] (AUCs typically range in the 60s).

In a clinical context, class-imbalance is the norm because because most patients do not have a given disease. Thus, for screening purposes, we want a classifier with a low false-negative rate in order to avoid under-diagnosing patients. To evaluate this, we computed

	AUC (%)			AUPR (%)		
Disease	**GCN**	**DNN**	**RF**	**GCN**	**DNN**	**RF**
Healthy	$\mathbf{84 \pm 1}$	82 ± 0	83 ± 0	73 ± 8	88 ± 1	$\mathbf{93 \pm 0}$
CRC	$\mathbf{84 \pm 1}$	81 ± 2	50 ± 0	41 ± 4	35 ± 3	$\mathbf{53 \pm 0}$
T2D	$\mathbf{90 \pm 1}$	86 ± 1	50 ± 0	40 ± 5	31 ± 2	$\mathbf{52 \pm 0}$
RA	$\mathbf{98 \pm 1}$	$\mathbf{99 \pm 0}$	$\mathbf{85 \pm 2}$	73 ± 4	74 ± 3	$\mathbf{85 \pm 2}$
Hypertension	$\mathbf{94 \pm 1}$	92 ± 1	66 ± 1	51 ± 5	44 ± 3	$\mathbf{60 \pm 2}$
IBD	$\mathbf{94 \pm 2}$	90 ± 1	50 ± 0	41 ± 4	36 ± 2	$\mathbf{52 \pm 0}$
Adenoma	$\mathbf{81 \pm 2}$	75 ± 2	50 ± 0	15 ± 3	12 ± 1	$\mathbf{51 \pm 0}$
Otitis	$\mathbf{86 \pm 2}$	74 ± 2	50 ± 0	18 ± 3	10 ± 1	$\mathbf{51 \pm 0}$
HBV	$\mathbf{97 \pm 1}$	$\mathbf{96 \pm 1}$	69 ± 2	$\mathbf{64 \pm 5}$	58 ± 3	$\mathbf{70 \pm 2}$
Fatty Liver	$\mathbf{89 \pm 2}$	80 ± 2	5 ± 0	35 ± 7	20 ± 3	$\mathbf{51 \pm 0}$
Psoriasis	$\mathbf{98 \pm 1}$	93 ± 2	73 ± 3	$\mathbf{74 \pm 3}$	57 ± 4	62 ± 3
T1D	$\mathbf{98 \pm 0}$	$\mathbf{97 \pm 1}$	53 ± 3	$\mathbf{47 \pm 7}$	$\mathbf{48 \pm 7}$	$\mathbf{53 \pm 2}$
Metabolic Syndrome	$\mathbf{98 \pm 1}$	$\mathbf{97 \pm 1}$	51 ± 1	$\mathbf{63 \pm 8}$	$\mathbf{55 \pm 3}$	51 ± 1
IGT	$\mathbf{95 \pm 1}$	88 ± 2	50 ± 0	22 ± 3	16 ± 1	$\mathbf{50 \pm 0}$
Periodontitis	$\mathbf{99 \pm 0}$	$\mathbf{99 \pm 0}$	96 ± 2	$\mathbf{95 \pm 2}$	93 ± 2	$\mathbf{96 \pm 2}$
AD	$\mathbf{99 \pm 0}$	$\mathbf{99 \pm 0}$	88 ± 3	$\mathbf{89 \pm 4}$	78 ± 5	$\mathbf{85 \pm 3}$
CDI	$\mathbf{99 \pm 0}$	91 ± 4	64 ± 3	$\mathbf{75 \pm 4}$	64 ± 8	62 ± 3
Infectious Gastroenteritis	$\mathbf{88 \pm 3}$	70 ± 1	50 ± 0	21 ± 4	14 ± 1	$\mathbf{50 \pm 0}$
Bronchitis	$\mathbf{78 \pm 4}$	$\mathbf{74 \pm 6}$	50 ± 0	13 ± 3	13 ± 2	$\mathbf{50 \pm 0}$
Average	$\mathbf{92 \pm 3}$	$\mathbf{88 \pm 12}$	62 ± 7	50 ± 11	45 ± 12	$\mathbf{62 \pm 7}$

Table 3: Percent area-under-precision-recall (AUPR) and area-under-ROC (AUC) by disease for each model. Boldface indicates the model(s) with the highest score in the category.

AUPR values, which summarize the relationship between positive-predictive-value and true-positive rate.[48] A high AUPR implies that a model can accurately identify positive labels while avoiding false-negatives. For most diseases, the random forest performed much better than the neural network models with respect to AUPR. This result, along with its superior accuracy on healthy patients, implies that the random forest would be more useful as a screening tool for identifying dysbiosis in general, while the neural networks are more useful as a diagnostic tool for ruling-in a specific disease in patients that have already been screened.

Next, we analyzed the ranking of predictions by our models. A ranked ordering of probabilities may be more useful to a physician than a single output, because that information can be integrated with the entire patient examination to generate a differential diagnosis. We examined the accuracy of the top-3 and top-5 predictions for each classifier (Figure 2). Most diseases were correctly identified within the top-3 classifications, and almost almost every disease was correctly present with at least 90% of the time in the top-5 classifications for each model. We measured which classes were most often predicted for a given label when an incorrect prediction was made for a subset of the labels (ie. how often the model confused a given disease for another disease). Healthy was often misclassified as T2D or CRC, and every disease

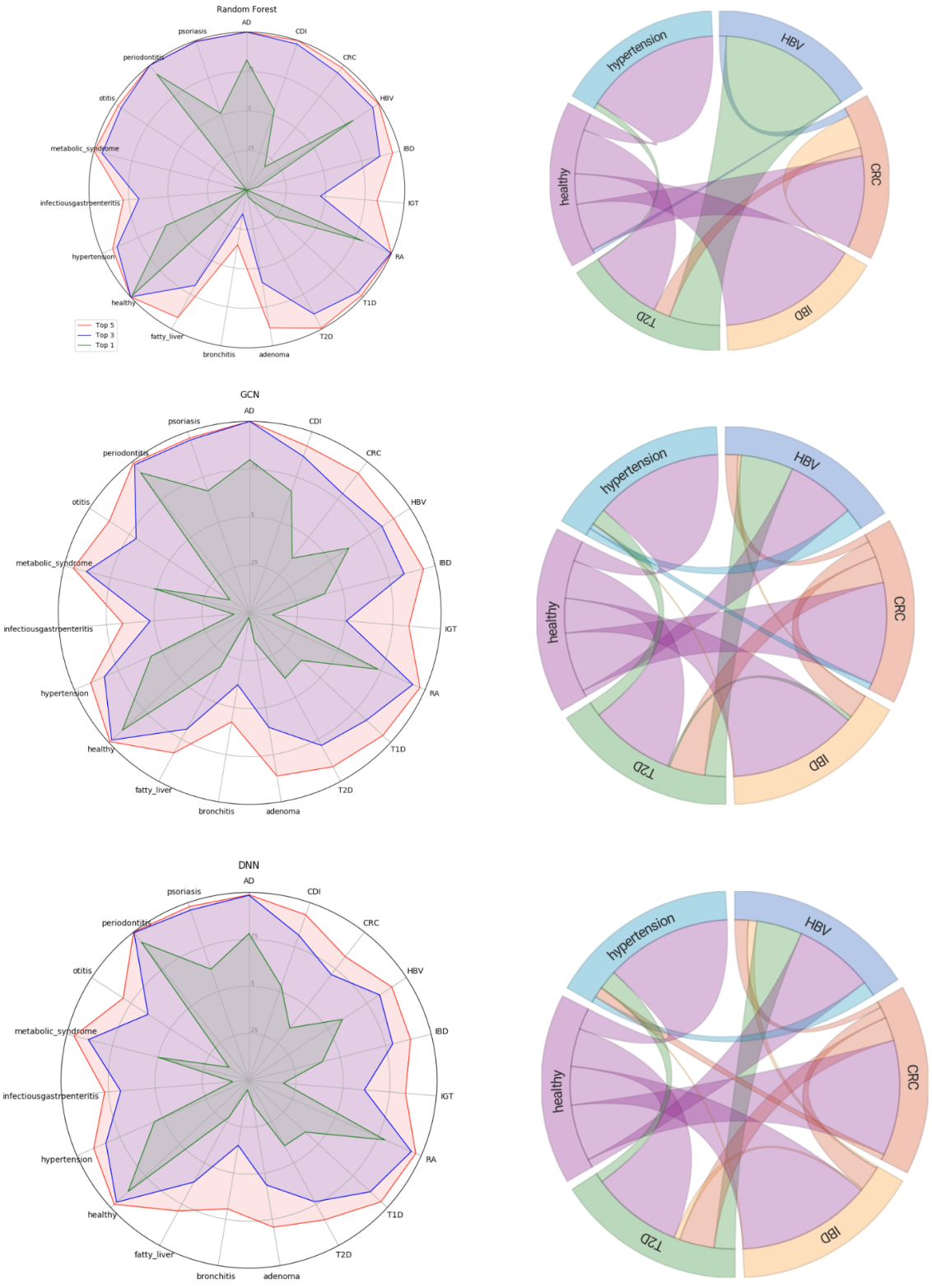

Fig. 2. (left) Accuracy at top-1,3, and 5 levels for (top to bottom) Random Forest, GCN, DNN. (right) Chord diagram showing (for a subset of labels) the most common classification made when an incorrect classification was made for a given class.

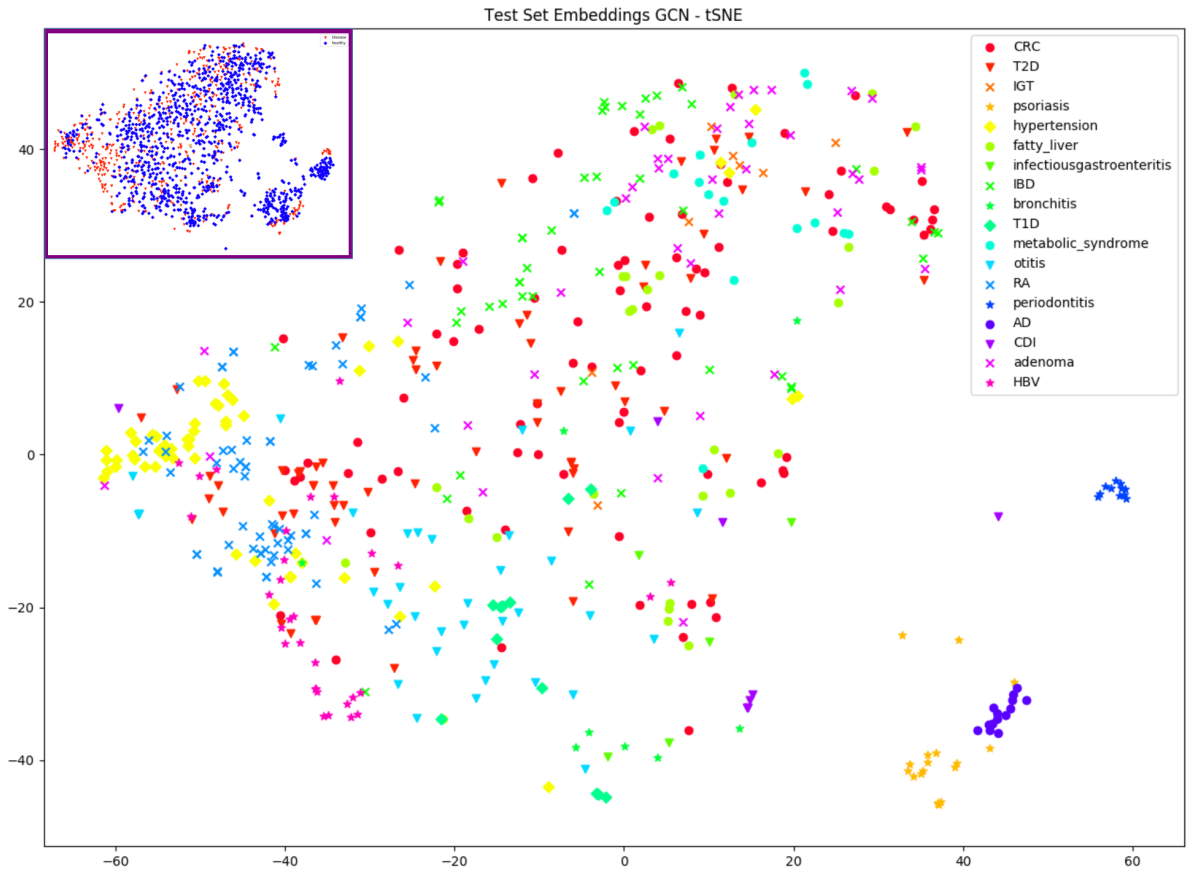

Fig. 3. tSNE visualization of the final layer activations of the GCN model (excluding healthy). Inset shows the same tSNE labeled by healthy (blue) and disease (red).

was often misclassified as healthy. These patterns were also consistent across the three models. These errors may simply be due to the limited size of the dataset or noise in the system. But it could also indicate that the concept of a 'healthy microbiome' is vacuous due to the broad of range of microbiomes that a healthy individual can have. Additionally, diseases can manifest with a broad range of severities, which may result in different metagenomic signatures which would all get grouped under a single label.

Lastly, we visualized the hidden layer activations of our GCN model on the test set using tSNE to see how different diseases cluster (Figure 3). We found that the skin conditions AD and psoriasis clustered together, separate from the other diseases which were evaluated on stool samples. Periodontitis, which was represented via oral microbiome samples, also clustered independently. Some other conditions, such as hypertension, adenoma, and otitis weakly clustered in some regions of the graph. We also visualized healthy and disease using the same tSNE and found that healthy microbiomes scattered throughout the plot alongside disease, consistent with the idea that there is a broad range of possible healthy microbiomes.

5. Conclusion

We have extended the results of previous work on microbiome-phenotype prediction here by demonstrating that multiclass disease prediction from whole community metagenomes, a clinically relevant task for machine learning, is a tractable problem and is improved by using the taxonomic structure of bacterial communities. We implemented multiple classifiers that were able to discriminate between 18 different diseases and healthy with greater than 70% accuracy on a dataset of over 7000 samples. Moreover, while the GCN model generally outperformed the DNN, the RF excelled on a completely different set of metrics, indicating that these two models could potentially complement one another. RF based models, achieving 99% accuracy on healthy vs. not-healthy, could be used as a screening test to identify dysbiosis in general, while GCNs could potentially then be used to discriminate between individual diseases. Additionally, the success of the GCN model implies that the geometric structure entailed in microbial phylogeny contains meaningful information for disease classification. This is a particularly exciting result, because graph and tree structures are ubiquitous in systems biology, so GCN architectures may be applicable to many other biological problems.[49]

While there are obvious clinical applications of a successful phenotype classifier, there are many questions that still need to be answered before these techniques can be deployed to the clinic. For example, we need to understand why the model makes certain predictions in order to give physicians more confidence in its diagnoses. Recently, a class of algorithms called attribution methods have emerged which can identify predictive features in the input on a per-sample basis.[50,51] Attribution methods could be useful from a personalized genomics standpoint by helping explain which bacteria are contributing to dysbiosis in an individual patient and potentially suggesting probiotic interventions to alleviate the dysbiosis. We will perform attribution analysis for our models to consider such questions in a future study.

There are also shortcomings stemming from the available data. For example, we are not aware of any method to infer causality between the microbiome and disease (ie. if the disease signatures the model detects are a cause or effect of the disease). Large-scale causal inference is typically done using time-series or interventional data,[52] which unfortunately there is very little of in this field.[53] Additionally, there is a great deal of patient metadata, such as gender, age, and body site, along with study-specific metadata that may be useful for making these predictions and could be explored in a future work. In general, there is no 'standardized-dataset' for this problem, making it difficult to compare different models. To make this exercise easier for other researchers, we have made our code available on Github, which contains not only our models and training scripts, but also our scripts for downloading and pre-processing the data. We believe that the results shown here will be improved upon in the future as more studies are added to curatedMetagenomicData and better models and training procedures emerge. We hope that the code we provide will help other researchers attack this problem and similar problems involving machine learning with metagenomics.

6. Acknowledgments

Saad Khan was supported by the Einstein Medical Scientist Training Program (2T32GM007288-45) and an NIH T32 fellowship on Geographic Medicine and Emerging In-

fectious Diseases (2T32AI070117-13). Libusha Kelly is supported in part by a Peer Reviewed Cancer Research Program Career Development Award from the United States Department of Defense (CA171019).

7. External Links

Source: github.com/kellylab/Metagenomic-Disease-Classification-With-Deep-Learning

References

1. Zhou, Y.-h. & Gallins, P. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction **10**, 1–14 (2019).
2. Vogtmann, E. *et al.* Colorectal cancer and the human gut microbiome: Reproducibility with whole-genome shotgun sequencing. *PLoS ONE* **11**, 1–13 (2016).
3. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications* **8**, 1784 (2017). URL http://www.nature.com/articles/s41467-017-01973-8.
4. Lin, H. Y. Efficient classifiers for multi-class classification problems. *Decision Support Systems* **53**, 473–481 (2012).
5. Reiman, D., Metwally, A. A. & Dai, Y. PopPhy-CNN : A Phylogenetic Tree Embedded Architecture for Convolution Neural Networks for Metagenomic Data 1–9 (2018).
6. Fioravanti, D. *et al.* Phylogenetic Convolutional Neural Networks in Metagenomics 1–12 (2017). URL http://arxiv.org/abs/1709.02268. 1709.02268.
7. Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine* **25**, 667–678 (2019). URL http://dx.doi.org/10.1038/s41591-019-0405-7.
8. LaPierre, N., Ju, C. J., Zhou, G. & Wang, W. MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods* 0–1 (2019). URL https://doi.org/10.1016/j.ymeth.2019.03.003.
9. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* **35**, 833–844 (2017).
10. McIver, L. J. *et al.* BioBakery: A meta'omic analysis environment. *Bioinformatics* **34**, 1235–1237 (2018).
11. Davidson, R. M. & Epperson, L. E. Microbiome sequencing methods for studying human diseases. In *Methods in Molecular Biology*, **77**–90 (Springer New York, 2018). URL https://doi.org/10.1007/978-1-4939-7471-9_5.
12. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *bioRxiv* (2017). URL http://biorxiv.org/content/early/2017/01/27/103085.abstract.
13. Chng, K. R. *et al.* Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare. *Nature Microbiology* **1**, 1–10 (2016). URL http://dx.doi.org/10.1038/nmicrobiol.2016.106.
14. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nature Communications* **6**, 6528 (2015). URL http://www.nature.com/doifinder/10.1038/ncomms7528.
15. Hannigan, G. D., Duhaime, M. B., Ruffin, M. T., Koumpouras, C. C. & Schloss, P. D. Diagnostic potential and interactive dynamics of the colorectal cancer virome. *mBio* **9**, 1–13 (2018).
16. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology* **10**, 766–766 (2014).
17. Yassour, M. *et al.* Natural history of the infant gut microbiome and impact of antibiotic treatments on strain-level diversity and stability. *Sci Trans Med* **8**, 1173–1178 (2016). 15334406.
18. Vincent, C. *et al.* Bloom and bust: Intestinal microbiota dynamics in response to hospital exposures and Clostridium difficile colonization or infection. *Microbiome* **4**, 1–11 (2016). URL http://dx.doi.org/10.1186/s40168-016-0156-3.
19. Yu, J. *et al.* Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70–78 (2017).
20. Loomba, R. *et al.* Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. *Cell Metabolism* **25**, 1054–1062.e5 (2017). URL https://linkinghub.elsevier.com/retrieve/pii/S1550413117302061.
21. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014). URL http://dx.doi.org/10.1038/nature13568.
22. Li, Y., Yu, R., Shahabi, C. & Liu, Y. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting 1–16 (2017). URL http://arxiv.org/abs/1707.01926. 1707.01926.
23. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology* **32**, 822–828 (2014).
24. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).
25. David, L. A. *et al.* Diet rapidly and reproducibly alters the humen gut micobiome. *NIH Public Access* **505**, 559–563 (2014).
26. Vrieze, A. *et al.* Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. *Gastroenterology* **143**, 913–916.e7 (2012). URL http://dx.doi.org/10.1053/j.gastro.2012.06.031.
27. Shi, B. *et al.* Dynamic changes in the subgingival microbiome and their potential for diagnosis and prognosis of periodontitis. *mBio* **6**, 1–11 (2015).
28. Tett, A. *et al.* Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis. *npj Biofilms and Microbiomes* **3**, 1–11 (2017). URL http://dx.doi.org/10.1038/s41522-017-0022-5.
29. Wen, C. *et al.* Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biology* **18**, 1–13 (2017).
30. Heintz-Buschart, A. *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature Microbiology* **2**, 16180 (2016). URL http://www.nature.com/articles/nmicrobiol2016180.
31. Kostic, A. D. *et al.* The dynamics of the human infant gut microbiome in development and in progression towards type 1 diabetes. *Cell Host and Microbe* **17**, 260–73 (2015).
32. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012). URL http://www.nature.com/doifinder/10.1038/nature11450. arXiv:1011.1669v3.
33. Asnicar, F. *et al.* Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2**, e00164–16 (2017). URL http://msystems.asm.org/lookup/doi/10.1128/mSystems.00164-16.
34. Aloysius, N. & Geetha, M. A review on deep convolutional neural networks. In *2017 International Conference on Communication and Signal Processing (ICCSP)* (IEEE, 2017). URL https://doi.org/10.1109/iccsp.2017.8286426.
35. Kauderer-Abrams, E. Quantifying Translation-Invariance in Convolutional Neural Networks (2017). URL http://arxiv.org/abs/1801.01450. 1801.01450.
36. Bronstein, M. M., Bruna, J., Lecun, Y., Szlam, A. & Vandergheynst, P. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine* **34**, 18–42 (2017). 1611.08097.
37. Kipf, T. N. & Welling, M. SEMI-SUPERVISED CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS 1–14 (2017). 1609.02907.
38. Ng, A. *Machine Learning Yearning* (deeplearning.ai).
39. Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric 1–9 (2019). URL http://arxiv.org/abs/1903.02428. 1903.02428.
40. Paszke, A. *et al.* Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop* (2017).
41. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
42. Batista, G. E. A. P. A., Prati, R. C. & Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* **6**, 20 (2004).
43. Malla, M. A. *et al.* Exploring the human microbiome: The potential future role of next-generation sequencing in disease diagnosis and treatment. *Frontiers in Immunology* **10**, 1–23 (2019).
44. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *Journal of Big Data* **6** (2019). URL https://doi.org/10.1186/s40537-019-0192-5.
45. Publication, K. S. & E, R. In Defense of One-Vs-All Classification. *Journal Of Machine Learning Research* **5**, 2–6 (2004).
46. Khan, S. & Kelly, L. Fireworks: Reproducible Machine Learning and Preprocessing with PyTorch. *Journal of Open Source Software* **4**, 1478 (2019). URL http://joss.theoj.org/papers/10.21105/joss.01478.
47. Tharwat, A. Classification assessment methods. *Applied Computing and Informatics* (2018). URL https://doi.org/10.1016/j.aci.2018.08.003.
48. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, 1–21 (2015).
49. Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical Science* **10**, 370–377 (2019).
50. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences (2017). URL http://arxiv.org/abs/1704.02685. 1704.02685.
51. Lundberg, S. & Lee, S.-I. An unexpected unity among methods for interpreting model predictions 1–6 (2016). URL http://arxiv.org/abs/1611.07478. 1611.07478.
52. Pearl, J. An introduction to causal inference. *The International Journal of Biostatistics* **6** (2010). URL https://doi.org/10.2202/1557-4679.1203.
53. Oh, J., Byrd, A. L., Park, M., Kong, H. H. & Segre, J. A. Temporal Stability of the Human Skin Microbiome. *Cell* **165**, 854–866 (2016). 15334406.