Title

MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction

Authors

LaPierre, Nathan
Ju, Chelsea J-T
Zhou, Guangyu
et al.

Publication Date

2019-08-01

DOI

10.1016/j.ymeth.2019.03.003

Peer reviewed

HHS Public Access

Author manuscript

Methods. Author manuscript; available in PMC 2020 August 15.

Published in final edited form as:

Methods. 2019 August 15; 166: 74–82. doi:10.1016/j.ymeth.2019.03.003.

MetaPheno: A Critical Evaluation of Deep Learning and Machine Learning in Metagenome-Based Disease Prediction

Nathan LaPierre, **Chelsea J.-T. Ju**, **Guangyu Zhou**, **Wei Wang***

Department of Computer Science, University of California at Los Angeles, Los Angeles CA 90095, USA

Abstract

The human microbiome plays a number of critical roles, impacting almost every aspect of human health and well-being. Conditions in the microbiome have been linked to a number of significant diseases. Additionally, revolutions in sequencing technology have led to a rapid increase in publicly-available sequencing data. Consequently, there have been growing efforts to predict disease status from metagenomic sequencing data, with a proliferation of new approaches in the last few years. Some of these efforts have explored utilizing a powerful form of machine learning called deep learning, which has been applied successfully in several biological domains. Here, we review some of these methods and the algorithms that they are based on, with a particular focus on deep learning methods. We also perform a deeper analysis of Type 2 Diabetes and obesity datasets that have eluded improved results, using a variety of machine learning and feature extraction methods. We conclude by offering perspectives on study design considerations that may impact results and future directions the field can take to improve results and offer more valuable conclusions. The scripts and extracted features for the analyses conducted in this paper are available via GitHub: https://github.com/nlapier2/metapheno

Keywords

Deep Learning; Machine Learning; Metagenomics; Phenotype Prediction

1. Introduction

The human body is home to a highly complex and densely populated microbial ecosystem, the so-called “human microbiome” [1, 2]. The microbes in the human body outnumber human cells and play a critical role in almost every aspect of human health and functioning [2]. The advent of High Throughput Sequencing (HTS) has enabled the direct study of microbial environments, forming the rich field of *metagenomics*. As sequencing becomes cheaper, vastly increased amounts of metagenomic sequencing data are becoming publicly available, including large-scale efforts such as the Human Microbiome Project, which aims

*To whom correspondence should be addressed. weiwang@cs.ucla.edu.

to understand the microbial environments in different human body sites [3]. We can interrogate this data to answer two key questions about the microorganisms in a community: who is there, and what are they doing [1]? By studying the taxonomic composition and metabolic activities of the microbes, we can begin to decipher how these properties contribute to human health and disease.

One recent development is the availability of a large amount of metagenomic shotgun sequence data matched to patients with labeled disease phenotypes, sometimes called "metagenome-wide association studies" or MGWAS [4]. This has in turn motivated computational researchers to develop machine learning methods to predict patient phenotype from their metagenomic sequence data. These models rely on extracting "features" from the sequence data. These features can represent different aspects of the microbiome, for instance, taxonomic composition or functional profiles. Ideally, the most informative features can provide insights into how the microbiome relates to the disease. The methods for extracting the features from the raw sequence data and the methods for predicting the disease based on the features are both important to the performance of the model.

An important step forward in this effort was perhaps the first machine learning meta-analysis of publicly-available MGWAS data, performed by Pasolli *et al.* [5]. In this study, the authors used a method called MetaPhlAn2 [6] to predict the composition of the patient's microbiome based on the sequence data. Using the predicted microorganisms and their abundances as features, they applied several well-known classical machine learning algorithms such as Support Vector Machines (SVMs) and Random Forests (RFs) to predict the patient's disease status. These approaches performed well at predicting some patient diseases such as liver cirrhosis, colorectal cancer, and inflammatory bowel disease, but poorly on the others, such as type 2 diabetes and obesity [5].

Since the publication of the meta-analysis, several other papers have emerged, attempting to use different machine learning methods to improve upon the original results [7, 8, 9] or apply machine learning to different types of data such as 16S rRNA [10]. Many of these methods have involved the use of deep learning [11], a powerful class of machine learning methods that have achieved record results in a number of domains, and has recently seen great success in biological prediction problems [12, 13, 14]. Briefly, deep learning uses a network of so-called "neurons" (inspired by real neural networks in the brain) to learn complex functions mapping input data, such as sequencing data, to an output value, such as a prediction of the disease status.

Here, we review recent methodological advancements in the prediction of disease from metagenomic data, with a particular focus on deep learning methods. These are discussed in Section 3. Readers who are unfamiliar with machine learning or deep learning may want to first read our review of these subjects in Section 2. In Section 4, we present the reported results on the data from the Pasolli *et al.* meta-analysis, as it serves as a common basis for comparison among recent methods. In Section 5, we present an in-depth analysis of a type 2 diabetes dataset from the meta-analysis that has eluded improved results. We apply a number of machine learning methods, including an autoencoder-pretrained neural network that we developed, to the data, and also explore an alternate *k*-mer-based feature extraction method.

In Section 6, we offer perspectives gained from the review, including considerations for study design and interpretability, and possible avenues to improve results in the future. Section 7 briefly summarizes the conclusions.

2. Overview of Machine Learning and Deep Learning Methods

2.1. Primer on Machine Learning

Machine learning, broadly defined, involves the use of computer algorithms to find the structure in data. In this study we focus on so-called "supervised learning", in which a mapping is learned from input data to an output label. Here, the "structure" of the data is represented as a set of features, extracted from the input data. In the context of this paper, the input data is metagenomic sequence reads, the extracted features are taxonomic or functional annotations, and the output label is the binary disease status prediction.

A crucial aspect of machine learning is ensuring that the learned model can work well not only on the available dataset but also on examples not included in the dataset. This is often called the "generalizability" of a model. To achieve this, the model is learned on a subset of the data, called "training data", and then evaluated on the rest of the data, called "testing data", which is held out from the training process. The testing data serves as a proxy for data outside the study.

A common way to split the data into training and testing is k-fold cross-validation (k-fold CV) [15]. The data is partitioned into k equally-sized subsets, called "folds". Each fold is used once as the testing data, and the model is trained on the rest of the data. The performance of the model is the average of the results for all k folds. In some cases, there may be a large imbalance of labels between two classes of data, for instance, many more case examples than control examples. For such cases, a slight variation in the k-fold CV technique can be used, in which each fold has the same case-to-control ratio as the entire dataset. This is called "stratified k-fold cross-validation". Finally, there is a specific case of k-fold cross-validation called "Leave One Out Cross-Validation (LOOCV)", in which k equals the number of individuals in the dataset. Thus, each individual is used as the test set once and the model is trained on the rest of the individuals.

Each machine learning model has a set of properties that can be set by the user prior to the training process, called "hyperparameters". For instance, before learning a decision tree, a hyperparameter can be set to limit the depth of the tree. The performance of the machine learning model can vary significantly based on the settings of hyperparameters [16, 17], so it is important to set them optimally. The traditional and most comprehensive way to do this is to perform a "grid search". Based on a user-specified set of hyperparameters and possible settings of them, the grid search exhaustively enumerates each combination of these settings. The grid search selects the best settings as measured by cross-validation [18].

2.2. Classical Machine Learning Algorithms

Two classical machine learning methods are commonly used in metagenome-based disease prediction: Support Vector Machines (SVMs) [19] and Random Forests (RFs) [20]. SVMs can be thought of as representing the input data as points in space, and their objective is to

learn a decision boundary to maximally separate different classes. To do this, SVMs search for the points in each class that are the closest to the decision boundary. Those points are called “support vectors”. RFs are an example of ensemble learning, in which a complex model is made by combining many simple models. In this case, the simple models are decision trees [20]. RFs take many random subsamples of the complete dataset. For each of these subsamples, a decision tree is learned. The final output of a RF is the most common prediction of the individual decision trees. As these are well-studied methods, they are used as baselines for comparison in many studies. Additionally, both SVMs and RFs can output the most informative features to the predictive model. In the context of metagenome-based disease prediction, these features are the microbes or the functional elements that contribute most to the disease prediction, enhancing the interpretability of the model [5, 9, 7, 21].

Several new methods have been proposed to improve upon these classical methods. eXtreme Gradient Boosting (XGBoost) [22] is similar to RFs, in that it builds an ensemble of decision trees. The main difference is that trees are sequentially built to reduce the errors of the previous trees. Another variant of the forest approach, called multi-Grained Cascade Forest (gcForest) or “deep forest” [23], performs an ensemble of forests, i.e. an ensemble of ensembles.

2.3. Deep Learning Algorithms

Deep learning is a powerful class of machine learning algorithms consisting of artificial neural networks (ANN) with many layers. These neural networks are inspired by biological neural networks in the human brain. They are composed of one or more inter-connected “layers”, each of which consists of separate simple computational units called “neurons”. The input information flows through the network as follows: each layer receives input data for each of its neurons, each neuron then executes a simple user-defined function, and then the output of the neuron is transmitted as input to neurons in the next layer. Two neurons are said to be connected if a neuron in one layer sends output to the other neuron in the next layer. The connections are weighted, reflecting the contribution to the prediction. The learning process of a neural network is the updating of these connection weights, based on prediction errors made with training data. By composing the numerous simple functions executed by each neuron in a network structure, complex relationships between inputs and their relevance to the output can be learned [11]. Networks with more layers can learn more complex functions, thus explaining the power of deep learning [11]. However, since the input features are sent through a complex network of functions, it is difficult to pinpoint the most informative features. This confounds the interpretability of the model [24]. Nevertheless, deep learning models have achieved record-breaking results in the fields of natural language processing [11], image classification [25, 26], and speech recognition [27]. The successes of these applications have encouraged the exploration of this approach in the field of bioinformatics [12, 28], including the analysis of metagenomic data [8, 9].

We review three main types of deep learning architectures in this paper: fully-connected feedforward deep neural networks (which we simply refer to as DNNs) [29, 30, 31], convolutional neural networks (CNNs) [32, 25], and auto-encoders (AEs) [33]. DNNs are general-purpose architectures, CNNs are specialized for image-based tasks, and AEs are

used for dimensionality reduction (see below). Another common architecture, Recurrent Neural Networks (RNNs) have thus far not often been used in metagenome-based disease prediction, so we do not review them in this paper.

As the name of DNNs suggests, every neuron in one layer is connected to every neuron in the next layer without backward connections. DNNs are sometimes also referred to as multilayer perceptrons (MLP) [29]. However, MLP can have a more general definition that includes other types of architectures, so we use DNNs to avoid ambiguities here.

CNNs are designed specifically to process images with spatial information. CNNs focus on summarizing local information with a mathematical function, called “convolution”, which greatly reduces the computational burden. For example, when analyzing a pixel in an image, the nearby pixels are the most relevant and there is no need to incorporate distant pixels. Because CNNs are very powerful for image processing, researchers have developed methods for encoding different types of information as images for a variety of applications, including metagenome-based disease predictions. Methods that leverage the CNNs architecture are discussed in Section 3.

AEs represent a different type of deep learning. In this case, the goal is not to predict an output value, but rather to find a more compressed representation of the input data [33]. This is also referred to as “dimensionality reduction” of the feature space. Dimensionality reduction addresses a common issue of deep learning, called overfitting. Overfitting refers to learning a model that is very specific to the training data but will not generalize well to the testing data. This is a concern when there are more features than samples, as is often the case in metagenome-based disease prediction [34]. AEs take a set of input features and learn a smaller set of latent features that capture the same amount of information. This is done by ensuring that the original set of features can be recovered from the smaller set with minimal loss [33]. By first applying AEs to obtain a reduced set of features, which are then used as input to DNNs, the model can avoid overfitting and generalize better [11, 33].

3. Current Methods in Metagenome-Based Disease Prediction

3.1. Feature Extraction

In disease phenotype prediction, there are three types of commonly used features extracted from metagenomic sequence reads: the abundances of different microbes, functional annotation of the metagenomic samples, or the *k*-mer abundances from raw reads.

Given the metagenomic sequence data, one of the key questions is to identify and quantify the presence of different microorganisms. Under the assumption that microbiome composition is different between healthy and diseased individuals, the profiles of microbial abundances are widely used as a type of feature in disease prediction. MetaPhlAn2 [6] is a popular tool to estimate the relative abundance of microbial taxa. It uses a set of clade-specific marker genes to assign reads to microbial clades. It then estimates the relative abundance of each taxon based on the read coverage. The majority of the metagenome-based disease predictive models in this paper leverage MetaPhlAn2 profiles for the underlying features [5, 8, 9]. Met2Img [8] uses the species abundances as the raw features; PopPhy-

CNN [9] aggregates the abundances reported by MetaPhlAn2 up to the genus level; MetaML [5] investigates the performance of using species abundances or the presence of strain-specific markers as the microbiome features. Alternatives to MetaPhlAn2 include Quikr [35], Bracken [36], and CLARK [37]. The platform UGENE can combine the results of several of these tools into a single “ensemble” prediction [38].

The other aspect of understanding a microbial community is addressing the question of “what are they doing?” through functional annotation. One example of this approach was demonstrated by Yazdani *et al.* [39], who detected protein family shifts between healthy and diseased gut microbiomes by using KEGG [40] annotations and a random forest classifier. Other methods attempt to infer the functional and metabolic properties of microbiomes from either shotgun [41] or 16S rRNA [42] sequence data. Predicted functional and metabolic profiles have been used to predict ecological roles in the rhizosphere [43] and general human gut dysbiosis [44].

The major drawback of the aforementioned feature extraction approaches is that they are limited by the reference database. In microbial abundance profiling, we can only estimate the abundance of known microbes, or the microbes present in the database. In functional profiling, we rely on the annotated genes and pathways that can be recognized in the sequencing data. Consequently, these two approaches toss away unmapped reads with valuable information [7]. In order to fully utilize all of the reads, several frameworks have proposed using the *k*-mer abundances directly acquired from the raw reads [45, 46, 47, 48]. These frameworks first count the *k*-mer frequencies of the metagenomic reads in each individual. Common *k*-mer counters include Jellyfish [49] and KMC [50]. The next step is to identify the significantly differentially abundant *k*-mers between the cases and controls through a statistical test, such as Student’s t-test, Wilcoxon rank-sum test, or likelihood ratio test. The false discovery rate is then controlled for multiple hypothesis testings. The statistically significant *k*-mers are sometimes used directly as the features. In other cases, the raw *k*-mer counts are used without statistical testing in pipelines alongside other steps, such as assembly and clustering [7].

3.2. Meta-Analysis of Classical Machine Learning Approach

Recently, the work of Pasolli *et al.* [5], MetAML (Metagenomic prediction Analysis based on Machine Learning) comprehensively assesses different machine learning approaches to metagenome-based disease prediction tasks. In MetAML, six available disease-associated metagenomic datasets spanning five diseases are discussed. They are: liver cirrhosis [51], colorectal cancer [52], inflammatory bowel diseases (IBD) [53], obesity [54], and type 2 diabetes (T2D) (two distinct studies [4] and [55]). Each dataset is evaluated independently by cross-validation. MetaPhlAn2 [6] taxa abundances are used as features. Several classical machine learning and statistical methods are evaluated. RFs performs the best followed by SVMs. However, deep learning methods (neural networks) are not evaluated.

Overall, the proposed MetAML method works well for some phenotypes such as liver cirrhosis, IBD, and colorectal cancer. However, it performs relatively poorly on T2D and obesity [5]. There are two main limitations of using MetaPhlAn2 for feature extraction. First, MetaPhlAn2 is limited to detecting only species in its reference database. A

metagenomic benchmark study by the CAMI consortium found that MetaPhlAn2 has a high false negative rate, meaning that it fails to identify many taxa present in the sample [56]. Due to the false negatives, the relative abundances are mis-estimated, leading to noise in the extracted features. Second, MetaPhlAn2 does not consider functional elements of the microbiome, limiting potentially valuable information that can be used to predict diseases.

3.3. Deep Learning Approaches

As deep neural networks (DNNs) have achieved excellent classification results, researchers have recently attempted to apply them to the problem of metagenome-based disease prediction. However, several challenges remain, with various methods attempting to address them.

Reiman *et al.* [9, 57] argue that the DNN architecture may not be suitable to predict diseases using metagenomic data. Learning through a deep architecture often requires an excessive amount of data, which is currently impractical with the limited number of sampled patients [7, 24]. In addition, as previously discussed, extracting the significant features from the learned models is not trivial. To mitigate these issues, Reiman *et al.* propose a framework that leverages the architecture of CNNs to predict diseases from microbial abundance profiles [9, 57]. Reiman *et al.*'s method PopPhy-CNN [9] uses phylogenetic trees to describe the relatedness of different features, i.e. microbes. The tree is further embedded in a 2D matrix to include the observed relative abundance of microbial taxa, allowing the CNNs to fully exploit the spatial relationship of the microbes and their quantitative characteristics in metagenomic data. A comprehensive evaluation has demonstrated that the framework can efficiently train models without an excessive amount of data. The significant microbes contributing to different diseases can also be extracted and visualized on the phylogenetic tree.

Another common issue in this domain is overfitting. To alleviate this issue when conducting disease predictions, Nguyen *et al.* [8] propose the Met2Img approach, which relies on embedding taxonomic abundances as color pixels in an image, called "synthetic images". Each image corresponds to an individual and each pixel corresponds to a taxon, with a color representing the abundance of that taxon. Pixels are arranged by phylogenetic sorting such that pixels near each other represent taxa that are phylogenetically similar. Nguyen *et al.* explore a variety of ways to set the colors and arrange the pixels. Finally, a CNN is used to predict the disease based on the image created. Evaluating on twelve benchmark datasets shows that Met2Img outperforms classical machine learning algorithms (RFs and SVMs) [8]. Nguyen *et al.* claim that the integration of phylogenetic information alongside abundance data improves classification [8].

Several other related approaches have been developed that use deep learning to predict both host and environmental phenotypes. MicroPheno [10] uses extracted *k*-mer counts to predict various host and environmental phenotypes, reporting that deep learning outperforms random forests for predicting environmental phenotypes but not disease phenotypes. MetaNN [58] uses microbe abundance profiles, augments them with simulated samples generated from a negative binomial distribution, and predicts host and body site phenotypes using either a DNN or a CNN. They report that their method improves on classical machine

learning approaches, and that the DNN outperformed the CNN [58]. Ditzler *et al.* apply a DNN and a recurrent neural network (RNN) to host and environmental phenotype prediction. They find that the DNN outperforms the RNN and a RF at predicting sample pH and body site, while the RF is the best at predicting host phenotype [59].

3.4. Other Machine Learning Approaches

Other methods have attempted to model the learning problem in a different way. RegMIL [7] is one such method. RegMIL takes the approach of Multiple Instance Learning (MIL), which considers a set of samples called "bags" that have known labels, and which contain a number of "instances", which have unknown labels. In this case, the bags are the individuals in a study, the known labels are the disease phenotypes, and the instances are the metagenomic sequence reads. Because individual sequence reads provide limited information, RegMIL begins by assembling reads into contigs and then binning and clustering contigs. Normalized *k*-mer counts are then obtained for the sequences in each cluster. Based on the association between *k*-mers and disease status in the training set, a neural network is used to predict which *k*-mers in the test set are associated with disease status. A RF classifier uses these predictions as features to predict the disease status of the individual. The authors claim that this approach leads to improved results over MetaML in both accuracy and AUC on the liver cirrhosis and IBD datasets [7]. RegMIL thus illustrates both a different way to model the classification problem and an alternate way to employ neural networks beyond the final disease prediction step.

Several other approaches have focused not directly on classification, but on feature selection. Ditzler *et al.* introduce a feature selection method called Fizzy that attempts to select important microbes or functional elements for downstream classification algorithms to analyze [60]. A competing taxonomy-aware feature selection method was recently released by Oudah and Henschel [34]. The authors claim that applying it prior to classification improves colorectal cancer prediction from 16S rRNA metagenomic data [34]. Feature selection for metagenome-based disease prediction seems to be a less-explored area, but may be just as important as the classification method used and may enhance interpretability, motivating further research in this direction.

4. Results from Previous Works on MetAML Datasets

Since several methods, including PopPhy-CNN [9], Met2Img [8], and RegMIL [7], have been developed in comparison to the results of MetAML [5], we review their results here in order to provide a comparison of several recent and related papers in the field. The datasets we cover in this review are profiled in Table 1. As previously stated, MetAML's best results are obtained with a RF, PopPhy-CNN and Met2Img are CNN based methods, and RegMIL models the problem with Multiple Instance Learning (MIL), while using both a nerual network and RF as part of their pipeline. Below, we compare and contrast the experimental procedures used in each study and then review the results

4.1. Cross-Validation Settings

MetAML performs grid search using stratified 5-fold cross-validation to select the hyperparameters for each classifier, and then runs 10-fold cross-validation 20 times using the selected hyperparameters to determine the disease classification results [5]. PopPhy-CNN performs hyperparameter grid search for SVMs using 5-fold cross-validation, while the CNN is manually tuned and the settings of RFs are mostly left to the default [9]. Met2Img performs 10 runs of stratified 10-fold cross-validation to gather results and hyperparameter tuning is not mentioned for any of the methods in the paper [8]. RegMIL [7] performs Leave One Out Cross-Validation (LOOCV), and sets the hyperparameters manually.

4.2. Evaluation Protocols

Commonly used evaluation metrics for binary classification include accuracy, precision, recall, F1-Score and area under the receiver operating characteristic curve (AUC). Accuracy simply refers to the percentage of correctly predicted individuals. Precision is the percentage of *predicted* cases that are actual cases. Recall is the percentage of *actual* cases that are correctly identified by the classifier. In other words, precision measures the rate of falsely predicting disease, while recall measures the rate of falsely predicting healthy. The F1-Score is the harmonic mean of precision and recall, defined as follows:

$$F1\text{-}Score = \frac{2 \times precision \times recall}{precision + recall}$$

Most classifiers can report the probability of their prediction, which can be considered as the confidence in the prediction. The AUC uses this information to summarize the false prediction rate at different confidence levels. While accuracy is the most straightforward representation of performance, the F1-Score and AUC are better metrics when there is an imbalance of cases and controls. PopPhy-CNN [9] reports AUC, Met2Img [8] reports accuracy, RegMIL [7] reports both accuracy and AUC, and MetAML [5] reports all of the above metrics.

4.3. Summary of Results

Because of the inconsistencies in cross-validation and hyperparameter tuning, we only report the results of the baseline RF model and the proposed model for each study without making cross-study comparisons. In Table 2, we show the comparison between PopPhy-CNN and its RF baseline, denoted by PopPhy-RF. They report increased AUC in liver cirrhosis, T2D, and obesity of between 0.8% and 3.4% [9]. Met2Img-CNN is reported to outperform their RF baseline (denoted as Met2Img-RF) in liver cirrhosis, obesity and IBD, and the differences are statistically significant based on the one-tailed t-test (p-value < 0.05) [8]. RegMIL compares their proposed model with the MetAML package (denoted as RegMIL baseline) and reports that it outperforms the baseline in terms of accuracy and AUC for both liver cirrhosis and IBD by 0.5–2%.

5. In-depth Analysis of Type 2 Diabetes and Obesity Datasets

As summarized in Section 4, machine learning methods present promising power (high accuracy and AUC) in predicting liver cirrhosis and IBD using only the information from metagenomic reads. However, these methods still struggle to predict T2D and obesity. Here we analyze the performance of many different classification algorithms on the T2D and obesity datasets. We also explore an alternate feature extraction method to see if the results can be improved.

5.1. k-mer-based Feature Extraction

Many existing approaches rely on MetaPhlAn2 to estimate the relative abundances of microbes in each individual based on the metagenomic data. To address the drawback of this approach as discussed in Section 3, we examine the potential to improve T2D and obesity prediction using *k*-mer abundance profiles.

We count the *k*-mer frequencies of the metagenomic reads in each individual using Jelly-fish [49]. To avoid the bias of different sequencing depths, the *k*-mer counts are normalized by the total number of possible *k*-mers in each sample. Each *k*-mer is represented by its canonical form (i.e., the lexicographical minimum of itself and its reverse complementary sequence). In our study, we set k to 12, resulting in 8,390,656 unique *k*-mers. Shorter *k*mers have a higher chance to randomly appear in the genome; longer *k*-mers generate more candidates in exponential order for the statistical analysis in the next step, which can be computationally intractable. We empirically find that setting $k = 12$ leads to sufficiently significant *k*-mers while still being computationally feasible to process.

To identify the significant *k*-mers, we conduct a statistical test based on the abundance of each *k*-mer between the cases and controls. We first pool the *k*-mer counts from all diseased samples into a case group and other samples into a control group. For each *k*-mer, we calculate the *p*-value with the Student's t-test, followed by the Benjamini-Hochberg procedure [64] to control the false discovery rate from multiple hypothesis testings. We sort the *k*-mers based on their adjusted *p*-values, and retain the top 1,000 *k*-mers as our significant features. This criterion is used because retaining all *k*-mers with *p*-values smaller than 0.05 increases the computational time and does not improve the performance. It is important to note that the significant *k*-mers are extracted from the training data for our machine learning analysis, and the same set of *k*-mers is then used for the testing data.

5.2. Evaluation Protocols

We compare the performance of *k*-mer-based features against the microbial abundance profile estimated by MetaPhlAn2 down to the strain level. These features are used as input to five different machine learning algorithms: SVM, RF, XGBoost, gcForest, and an AE-pretrained DNN (henceforth referred to as AutoNN). Hyperparameter grid search is performed for all five algorithms using 5-fold cross-validation to select the best settings. With these settings, each model is evaluated over five independent runs of 5-fold cross-validation. We report the accuracy, precision, recall, F1-Score, and AUC for each model (defined in Section 4). We also conduct a pairwise statistical test to determine if the result of

the best *k*-mer-based approach is significantly better than the result of the best MetaPhlAn-based approach.

5.3. Summary of Classification Results

Tables 3 and 4 show that different models yield different performances when learning from these two types of features. When learning from the microbial abundance profiles, there is no single model that outperforms the others in all metrics. SVM achieves the best recall and F1-Score in both the T2D and obesity analyses. However, the SVM simply leverages the class imbalance in the obesity data (see Table 1) to achieve perfect recall, and reasonable precision and accuracy, by predicting positive for every sample. This highlights the importance of careful interpretation for metrics on imbalanced data such as the obesity and IBD datasets. The best accuracy using the microbial features is achieved by AutoNN and RF for T2D and obesity, respectively; the best precision is demonstrated by RF for T2D and XGBoost for obesity.

On the other hand, gcForest is particularly effective at learning from the *k*-mer abundance profiles. It consistently outperforms the others in all metrics in the T2D analysis. A similar observation is shown in the obesity analysis, except that RF achieves the best recall. We further evaluate whether the best accuracy results are significantly different between the *k*-mer and microbial abundance features. The pairwise Student's t-test reveals that gcForest with *k*-mer features is not significantly different from AutoNN with microbial abundance features in the T2D dataset (p-value of 0.096 after Benjamini-Hochberg correction). Similarly, in the obesity dataset, gcForest with *k*-mer features is not significantly different from RF with microbial abundance features (p-value 0.558). These analyses further the evidence that T2D and obesity will continue to be challenging traits to predict using only metagenomic reads.

5.4. Hyperparameter Grid Search Details

Here we discuss the details of the grid search that was performed to select the best hyperparameters for classification. Grid search was performed for all five algorithms using 5-fold cross-validation to select the best settings, which were then used in the subsequent classification steps. We attempted to identify a limited number of critical hyperparameters for each algorithm that significantly modified performance, as comprehensively evaluating all combinations of all possible hyperparameters is computationally infeasible. Similarly, we ran some small tests to evaluate choices for these settings that were computationally feasible and positively affected results. These hyperparameters (and the settings evaluated) were: the type of kernel (linear/polynomial) and the error term penalty (0.25/0.5/0.75/1.0/1.25/1.5/1.75/2.0) for the SVM; the maximum tree depth (2/6/10), number of estimators (10/50/100), and the splitting criterion (entropy/gini) for the Random Forest; the maximum tree depth (2/6/10), "alpha" L1 regularization term (0/0.25/0.5), and "lambda" L2 regularization term (0.5/1.0/1.5) for XGBoost; the number of training rounds (3/5) and the maximum forest depth (unlimited/50/100) for gcForest; the number of autoencoder layers (none/1/2/3), number of feedforward layers (3/5/10), dropout rate (0/0.25/0.5), optimizer (stochastic gradient descent[65]/adagrad[66]/adam[67]), and learning rate (0.01/0.001) for AutoNN. SVM and RandomForest were implemented via the scikit-

learn library [68] and the AutoNN was implemented in Keras [69]. For more information on the XGBoost [22] and gcForest [23] hyperparameters, see their respective papers and software packages.

For the taxonomic features, the SVM's best hyperparameter settings were a linear kernel and an error term penalty parameter of 1.75. For the Random Forest, the best hyperparameter settings were a maximum tree depth of 6, 100 estimators, and the entropy splitting criterion. For XGBoost, the best settings were a maximum tree depth of 2, an alpha of 0.0, and a lambda of 1.0. For gcForest, the best settings were 5 rounds of training and unlimited maximum forest layers. For AutoNN, the best settings were a single autoencoder layer, five feedforward layers, a dropout rate of 0.5, the adagrad optimizer, and a learning rate of 0.001.

For the *k*-mer-based features, the SVM's best hyperparameter settings were a linear kernel and an error term penalty parameter of 0.25. For the Random Forest, the best hyperparameter settings were a maximum tree depth of 6, 50 estimators, and the gini splitting criterion. For XGBoost, the best settings were a maximum tree depth of 2, an alpha of 0.25, and a lambda of 1.5. For gcForest, the best settings were 3 rounds of training and unlimited maximum forest layers. For AutoNN, the best settings were a single autoencoder layer, three feedforward layers, a dropout rate of 0.25, the adam optimizer, and a learning rate of 0.001.

6. Discussion

We have reviewed several methods that claim to improve disease prediction on several datasets from a popular meta-analysis by Pasolli *et al.* [5]. There are several inconsistencies that make a comparative analysis of these methods difficult, namely different cross-validation and hyperparameter searching methods used both between and within studies, and different classification metrics being reported between studies. Any valid cross-validation analysis is reasonable to report in a given study, whether 5-fold, 10-fold, or LOOCV, but *within* the same study, each method should be run with the same cross-validation and comprehensive hyperparameter search settings. As for which cross-validation method is ideal for this setting, there is no obvious best choice, but LOOCV has been shown to have low bias and strong generalization to new data [15, 70, 71], with the main drawback being computational cost [15]. It is often recommended for small datasets and has the additional benefit of avoiding questions surrounding stratification and different numbers of independent *k*-fold runs. Performance metric inconsistency is also an issue. With case-control class imbalances, different metrics may vary in usefulness, but reporting all of the ones mentioned in Section 4 makes it clear why an algorithm is outperforming others, whether due to fewer false case predictions or fewer false control predictions. Some papers also report the Matthews Correlation Coefficient (MCC) which is robust to case-control class imbalances [72]. Overall, greater clarity and robustness of results can be achieved by keeping study methodology and performance metrics consistent across all tested algorithms.

There are several other ways that interpretability can be enhanced. PopPhy-CNN, RegMIL, and MetAML all discuss the most significant microbes for their classification models. This facilitates comparisons between the biological implications suggested by each model.

Met2Img provides results for many different variants of their method, and also used a t-test to highlight significant results [8]. All of these methods provide confidence bounds for their predictions. Each of these factors help to determine the robustness and the relevance of results. Another aid to replicable results and consistent experiments is public, centralized resources for metagenomic data analysis. One example of this is ExperimentHub [73], which compiles many phenotyped metagenomic datasets, including those used in the Pasolli *et al.* meta-analysis. ExperimentHub provides both microbiome taxonomic and functional annotations [73].

Feature extraction plays an important role in the performance of the classification model. We have reviewed the benefits and limitations of MetaPhlAn2-based feature extraction and also discussed an alternative *k*-mer-based approach in this paper. One difficulty of the *k*-mer-based approach is the computational burden of analyzing *k*-mers with a large *k* because of the exponential increase of the numbers of possible *k*-mers. With short *k*-mers, the interpretability is challenging, as it is unclear what the *k*-mers represent. One less explored feature extraction approach is attempting to explicitly infer functional characteristics of the microbiome, using methods such as HUMAnN [41] or PICRUSt [42]. Finally, integration between different types of extracted features can be explored and further research in this direction is critical.

Ultimately, however, there has been extensive effort put into these studies with increasingly powerful machine learning algorithms, but with only minor performance improvements and modest changes in feature importance rankings. This suggests that there are upper limits on predictive accuracy that can be achieved from only metagenomic sequence read data. Thus, perhaps the greatest way to improve results is to include genetic data from the human subjects from whom metagenomic samples are taken. While this increases the cost of studies, it is likely critical to understanding the microbiome's role in complex phenotypes such as obesity and T2D. For instance, it has been demonstrated recently that combining micro-biome and genetic data can significantly improve the prediction accuracy of several human traits, including obesity [74]. Additonally, microbiome and genetic data are largely complementary in contributing to this predictive performance, and the microbiome is largely shaped by the environment [74]. Critically, these results indicate that using microbiome data alongside host genetic data can help disentangle the intricate web of genetic and environmental factors that lead to complex traits. Additional multi-omic data sources, such as metatranscriptomics, are just now seeing increased availability and hold significant potential for elucidating the function of the microbiome [71]. Finally, deep learning has been suggested as a promising method for successfully integrating multiple data types [75], and existing methods such as Similarity Network Fusion [76] can also be employed.

We note that, while disease prediction has been challenging in some cases, deep learning methods in particular seem to perform extremely well at classifying the body-site origin of microbial samples from the HMP [3] and other datasets as reported by MicroPheno [10] and others [58, 59]. Other works have performed strongly at predicting phenotypes of the micro-biome itself (as opposed to host phenotype) [59, 77], predicting disease with deep learning on non-metagenomic data [78], or identifying protein family shifts in microbiomes of diseased patients [39]. While these directions are outside the scope of this review, they

highlight other interesting applications of machine learning and deep learning in metagenome-based phenotype prediction.

7. Conclusion

Disease prediction using metagenomic sequence data has shown some potential, with a particularly large amount of effort having been put into deep learning methods, but remains challenging. Study methodology must remain consistent to compare different classification methods, especially when margins of difference in performance are so small. Feature extraction is as crucial to predictive performance as the classification methods themselves, and deserves increased attention. Supplementing metagenomic data with human genetic data may be the best way to improve both classification performance and biological understanding, especially with hard-to-classify complex traits such as obesity and type 2 diabetes. This is because genetic and metagenomic data provide complementary information about the host and environment, respectively [74].

Acknowledgements

The authors would like to thank the NSF the NIH for their funding and support via NSF grants DGE-1829071, DBI-1565137 and NIH grants T32 EB016640, R01 GM115833.

References

[1]. Handelsman J, Metagenomics: application of genomics to uncultured microorganisms, Microbiology and molecular biology reviews 68 (4) (2004) 669–685. [PubMed: 15590779]

[2]. Wooley JC, Godzik A, Friedberg I, A primer on metagenomics, PLoS computational biology 6 (2) (2010) e1000667. [PubMed: 20195499]

[3]. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI, The human microbiome project, Nature 449 (7164) (2007) 804. [PubMed: 17943116]

[4]. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, et al., A metagenome-wide association study of gut microbiota in type 2 diabetes, Nature 490 (7418) (2012) 55. [PubMed: 23023125]

[5]. Pasolli E, Truong DT, Malik F, Waldron L, Segata N, Machine learning meta-analysis of large metagenomic datasets: tools and biological insights, PLoS computational biology 12 (7) (2016) e1004977. [PubMed: 27400279]

[6]. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C,Segata N, Metaphlan2 for enhanced metagenomic taxonomic profiling, Nature methods 12 (10) (2015) 902. [PubMed: 26418763]

[7]. Rahman MA, Rangwala H, Regmil: Phenotype classification from metagenomic data, in: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM, 2018, pp. 145–154.

[8]. Nguyen TH, Prifti E, Chevaleyre Y, Sokolovska N, Zucker J-D, Disease classification in metagenomics with 2d embeddings and deep learning, arXiv preprint arXiv:1806.09046.

[9]. Reiman D, Metwally AA, Dai Y, Popphy-cnn: A phylogenetic tree embedded architecture for convolution neural networks for metagenomic data, bioRxiv (2018) 257931.

[10]. Asgari E, Garakani K, McHardy AC, Mofrad MR, Micropheno: Predicting environments and host phenotypes from 16s rrna gene sequencing using a k-mer based representation of shallow sub-samples, bioRxiv (2018) 255018.

[11]. LeCun Y, Bengio Y, Hinton G, Deep learning, nature 521 (7553) (2015) 436. [PubMed: 26017442]

[12]. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J,Nguyen N, Afshar PT, et al., A universal snp and small-indel variant caller using deep neural networks, Nature biotechnology.

[13]. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115. [PubMed: 28117445]

[14]. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, arXiv preprint arXiv:1711.05225.

[15]. Arlot S, Celisse A, A survey of cross-validation procedures for model selection, Statistical Surveys 4 (2010) 40–79.

[16]. Claesen M, De Moor B, Hyperparameter search in machine learning, arXiv preprint arXiv: 1502.02127.

[17]. Hoos H, Leyton-Brown K, An efficient approach for assessing hyperparameter importance, in: International Conference on Machine Learning, 2014, pp. 754–762.

[18]. Hsu C-W, Chang C-C, Lin C-J, et al., A practical guide to support vector classification.

[19]. Cortes C, Vapnik V, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[20]. Breiman L, Random forests, Machine learning 45 (1) (2001) 5–32.

[21]. Duvallet C, et al., Meta-analysis of gut microbiome studies identifies disease-specific and shared responses, Nature Communications 8 (2017) e1784.

[22]. Chen T, Guestrin C, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016, pp. 785–794.

[23]. Zhou Z-H, Feng J, Deep forest: Towards an alternative to deep neural networks, arXiv preprint arXiv:1702.08835.

[24]. Ching T, et al., Opportunities and obstacles for deep learning in biology and medicine, Journal of the Royal Society Interface 15 (141) (2018) e20170387.

[25]. Krizhevsky A, Sutskever I, Hinton GE, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[26]. LeCun Y, Bottou L, Bengio Y, Haffner P, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[27]. Deng L, Yu D, Deep convex net: A scalable architecture for speech pattern classification, in: Twelfth Annual Conference of the International Speech Communication Association, 2011.

[28]. Min S, Lee B, Yoon S, Deep learning in bioinformatics, Briefings in bioinformatics 18 (5) (2017) 851–869. [PubMed: 27473064]

[29]. Svozil D, Kvasnicka V, Pospichal J, Introduction to multi-layer feed-forward neural networks, Chemo-metrics and intelligent laboratory systems 39 (1) (1997) 43–62.

[30]. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, Journal of machine learning research 11 (Dec) (2010) 3371–3408.

[31]. Hinton GE, Salakhutdinov RR, Reducing the dimensionality of data with neural networks, science 313 (5786) (2006) 504–507. [PubMed: 16873662]

[32]. LeCun Y, Boser BE, Denker JS, Henderson D, Howard RE, Hubbard WE, Jackel LD, Handwritten digit recognition with a back-propagation network, in: Advances in neural information processing systems, 1990, pp. 396–404.

[33]. Hinton G, Salakhutdinov R, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507. [PubMed: 16873662]

[34]. Oudah M, Henschel A, Taxonomy-aware feature engineering for microbiome classification, BMC Bioinformatics 19 (1) (2018) e227.

[35]. Koslicki D, Foucart S, Rosen G, Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing, Bioinformatics 29 (17) (2013) 2096–2102. [PubMed: 23786768]

[36]. Lu J, Breitwieser FP, Thielen P, Salzberg SL, Bracken: estimating species abundance in metagenomics data, PeerJ Computer Science 3 (2017) e104.

[37]. Ounit R, Wanamaker S, Close TJ, Lonardi S, Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers, BMC genomics 16 (1) (2015) 236. [PubMed: 25879410]

[38]. Rose R, Golosova O, Sukhomlinov D, Tiunov A, Prosperi M, Flexible design of multiple metagenomics classification pipelines with ugene, Bioinformatics.

[39]. Yazdani M, et al., Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease, in: International Conference on Big Data, Vol. 28, Association for Computing Machinery, 2016, pp. 1272–1280.

[40]. Kanehisa M, Goto S, Kegg: kyoto encyclopedia of genes and genomes, Nucleic Acids Research 28 (1) (2000) 27–30. [PubMed: 10592173]

[41]. Abubucker S, et al., Metabolic reconstruction for metagenomic data and its application to the human microbiome, PLoS Computational Biology 8 (6) (2012) e1002358. [PubMed: 22719234]

[42]. Langille MG, et al., Predictive functional profiling of microbial communities using 16s rrna marker gene sequences, Nature Biotechnology 31 (9) (2013) 814–821.

[43]. Larsen PE, Collart FR, Dai Y, Predicting ecological roles in the rhizosphere using metabolome and transportome modeling, PloS one 10 (9) (2015) e0132837. [PubMed: 26332409]

[44]. Larsen PE, Dai Y, Metabolome of human gut microbiome is predictive of host dysbiosis, Gigascience 4 (1) (2015) 42. [PubMed: 26380076]

[45]. Han W, Wang M, Ye Y, A concurrent subtractive assembly approach for identification of disease associated sub-metagenomes, in: International Conference on Research in Computational Molecular Biology, Springer, 2017, pp. 18–33.

[46]. Wang M, Doak TG, Ye Y, Subtractive assembly for comparative metagenomics, and its application to type 2 diabetes metagenomes, Genome biology 16 (1) (2015) 243. [PubMed: 26527161]

[47]. Dubinkina VB, Ischenko DS, Ulyantsev VI, Tyakht AV, Alexeev DG, Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis, BMC bioinformatics 17 (1) (2016) 38. [PubMed: 26774270]

[48]. Rahman A, Hallgrímsdóttir I, Eisen M, Pachter L, Association mapping from sequencing reads using k-mers, eLife 7 (2018) e32920. [PubMed: 29897334]

[49]. Marçais G, Kingsford C, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, Bioinformatics 27 (6) (2011) 764–770. [PubMed: 21217122]

[50]. Kokot M, Długosz M, Deorowicz S, Kmc 3: counting and manipulating k-mer statistics, Bioinformatics 33 (17) (2017) 2759–2761. [PubMed: 28472236]

[51]. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, et al., Alterations of the human gut microbiome in liver cirrhosis, Nature 513 (7516) (2014) 59. [PubMed: 25079328]

[52]. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F,Habermann N, et al., Potential of fecal microbiota for early-stage detection of colorectal cancer, Molecular systems biology 10 (11) (2014) 766. [PubMed: 25432777]

[53]. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al., A human gut microbial gene catalogue established by metagenomic sequencing, nature 464 (7285) (2010) 59. [PubMed: 20203603]

[54]. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto J-M, Kennedy S, et al., Richness of human gut microbiome correlates with metabolic markers, Nature 500 (7464) (2013) 541. [PubMed: 23985870]

[55]. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, Nielsen J, Bäckhed F, Gut metagenome in european women with normal, impaired and diabetic glucose control, Nature 498 (7452) (2013) 99. [PubMed: 23719380]

[56]. Sczyrba A, et al., Critical assessment of metagenome interpretationa benchmark of metagenomics software, Nature Methods 14 (2017) 1063–1071. [PubMed: 28967888]

[57]. Reiman D, Metwally A, Dai Y, Using convolutional neural networks to explore the microbiome, in: Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE, IEEE, 2017, pp. 4269–4272.

[58]. Lo C, Marculescu R, Metann: Accurate classification of host phenotypes from metagenomic data using neural networks, in: International Conference on Bioinformatics, Computational Biology, and Health Informatics, Association for Computing Machinery, 2018, pp. 608–609.

[59]. Ditzler G, Polikar R, Rosen G, Multi-layer and recursive neural networks for metagenomic classification, IEEE Transactions on NanoBioscience 14 (6) (2015) 608–616. [PubMed: 26316190]

[60]. Ditzler G, et al., Fizzy: feature subset selection for metagenomics, BMC Bioinformatics 16 (1) (2015) e358.

[61]. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, et al., Alterations of the human gut microbiome in liver cirrhosis, Nature 513 (7516) (2014) 59. [PubMed: 25079328]

[62]. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto J-M, Kennedy S, et al., Richness of human gut microbiome correlates with metabolic markers, Nature 500 (7464) (2013) 541. [PubMed: 23985870]

[63]. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al., A human gut microbial gene catalogue established by metagenomic sequencing, nature 464 (7285) (2010) 59. [PubMed: 20203603]

[64]. Benjamini Y, Hochberg Y, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the royal statistical society. Series B (Methodological) (1995) 289–300.

[65]. Bottou L, Large-scale machine learning with stochastic gradient descent, in: Proceedings of COMPSTAT'2010, Springer, 2010, pp. 177–186.

[66]. Duchi J, Hazan E, Singer Y, Adaptive subgradient methods for online learning and stochastic optimization, Journal of Machine Learning Research 12 (Jul) (2011) 2121–2159.

[67]. Kingma DP, Ba J, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[68]. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[69]. Chollet F, keras, https://github.com/fchollet/keras (2015).

[70]. Varma S, Simon R, Bias in error estimation when using cross-validation for model selection, BMC Bioinformatics 7 (1) (2006) e91.

[71]. Waldron L, Data and statistical methods to analyze the human microbiome, mSystems 3 (2) (2018) e00194–17. [PubMed: 29556541]

[72]. Boughorbel S, Jarray F, El-Anbari M, Optimal classifier for imbalanced data using matthews correlation Coefficient metric, PloS one 12 (6) (2017) e0177678. [PubMed: 28574989]

[73]. Pasolli E, et al., Accessible, curated metagenomic data through experimenthub, Nature Methods 14 (2017) 1023–1024. [PubMed: 29088129]

[74]. Rothschild D, et al., Environment dominates over host genetics in shaping human gut microbiota, Nature 555 (2018) 210–215. [PubMed: 29489753]

[75]. Camacho DM, et al., Next-generation machine learning for biological networks, Cell 173 (7) (2018) 1581–1592. [PubMed: 29887378]

[76]. Wang B, et al., Similarity network fusion for aggregating data types on a genomic scale, Nature Methods 11 (2014) 333–337. [PubMed: 24464287]

[77]. Feldbauer R, et al., Prediction of microbial phenotypes based on comparative genomics, BMC Bioinformatics 16 (14) (2015) S1.

[78]. Fakoor R, et al., Using deep learning to enhance cancer diagnosis and classification, in: International Conference on Machine Learning, Vol. 28, Association for Computing Machinery, 2013.

Table 1:

Summary of the datasets covered in this study. More information is available in the MetAML paper [5].

	Number of Case samples	Number of Control samples	Citation
Liver Cirrhosis	118	114	[61]
T2D	170	174	[4]
Obesity	164	89	[62]
IBD	25	85	[63]

Table 2:

Comparison of machine learning approaches in predicting different diseases.

	Liver Cirrhosis		T2D		Obesity		IBD	
	Accuracy	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
MetAML-SVM	0.834 (0.052)	0.922 (0.041)	0.613 (0.057)	0.663 (0.066)	0.636 (0.042)	0.648 (0.071)	0.809 (0.066)	0.862 (0.083)
MetAML-RF	0.877 (0.043)	0.945 (0.036)	0.664 (0.052)	0.744 (0.056)	0.644 (0.052)	0.744 (0.056)	0.809 (0.050)	0.890 (0.078)
PopPhy-RF	NA	0.932	NA	0.727	NA	0.642	NA	NA
PopPhy-CNN	NA	0.94	NA	0.753	NA	0.676	NA	NA
Met2Img-RF	0.877 (0.060)	NA	0.672 (0.080)	NA	0.645 (0.042)	NA	0.808 (0.068)	NA
Met2Img-CNN	0.905 (0.071)	NA	0.651 (0.094)	NA	0.680 (0.066)	NA	0.868 (0.081)	NA
RegMIL baseline	0.923 (0.041)	0.922 (0.040)	NA	NA	NA	NA	0.8387 (0.028)	0.8242 (0.0374)
RegMIL-RF	0.928 (0.036)	0.927 (0.035)	NA	NA	NA	NA	0.847 (0.035)	0.844 (0.026)

Table 3:

Comparison of different types of features used to train the models for T2D. The mean and standard deviation are recorded for different evaluation metrics after five runs of 5-fold cross-validation. The best performances are highlighted in bold.

	Microbial Abundances					k-mer Abundances				
	Accuracy	**Precision**	**Recall**	**F1-Score**	**AUC**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
SVM	0.6429 (0.0072)	0.6259 (0.0058)	**0.7204 (0.0148)**	**0.6644 (0.0083)**	0.7250 (0.0048)	0.6375 (0.0204)	0.6405 (0.0278)	0.6168 (0.0215)	0.6246 (0.0156)	0.6945 (0.0215)
RF	0.6567 (0.0177)	**0.6805 (0.0162)**	0.6022 (0.0250)	0.6324 (0.0219)	0.7285 (0.0127)	0.6796 (0.0164)	0.6935 (0.0159)	0.6418 (0.0165)	0.6630 (0.0175)	0.7461 (0.0082)
XGBoost	0.6398 (0.0181)	0.6449 (0.0198)	0.6146 (0.0253)	0.6259 (0.0215)	0.6911 (0.0112)	0.6764 (0.0250)	0.6957 (0.0292)	0.6317 (0.0296)	0.6573 (0.0267)	0.7310 (0.0154)
gcForest	0.6550 (0.0181)	0.6524 (0.0198)	0.6669 (0.0253)	0.6547 (0.0215)	0.7341 (0.0112)	**0.6942 (0.0059)**	**0.6979 (0.0101)**	**0.6845 (0.0152)**	**0.6874 (0.0068)**	**0.7616 (0.0106)**
AutoNN	**0.6626 (0.0184)**	0.6644 (0.0222)	0.6598 (0.0192)	0.6574 (0.0183)	**0.7343 (0.0160)**	0.6517 (0.0085)	0.6444 (0.0116)	0.6762 (0.0194)	0.65 26 (0.0102)	0.7134 (0.0039)

Table 4:

Comparison of different types of features used to train the models for obesity. The mean and standard deviation are recorded for different evaluation metrics after five runs of 5-fold cross-validation. The best performances are highlighted in bold.

	Microbial Abundances					k-mer Abundances				
	Accuracy	**Precision**	**Recall**	**F1-Score**	**AUC**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
SVM	0.6374 (0.0008)	0.6374 (0.0008)	**1.0000 (0.0000)**	**0.7768 (0.0012)**	0.5133 (0.0361)	0.6154 (0.0271)	0.6923 (0.0199)	0.7229 (0.0296)	0.7031 (0.0202)	0.5993 (0.0174)
RF	**0.6480 (0.0112)**	0.6512 (0.0034)	0.9675 (0.0182)	0.7764 (0.0087)	0.6416 (0.0062)	0.6139 (0.0161)	0.6733 (0.0054)	**0.7786 (0.0272)**	0.7170 (0.0155)	0.5937 (0.0268)
XGBoost	0.6352 (0.0241)	**0.6749 (0.0112)**	0.8277 (0.0366)	0.7407 (0.0205)	0.6055 (0.0241)	0.6169 (0.0261)	0.6818 (0.0116)	0.7614 (0.0427)	0.7145 (0.0253)	0.5979 (0.0196)
gcForest	0.6404 (0.0125)	0.6553 (0.0094)	0.9247 (0.0163)	0.7644 (0.0082)	**0.6495 (0.0148)**	**0.6365 (0.0242)**	**0.7042 (0.0194)**	0.7470 (0.0282)	**0.7211 (0.0184)**	**0.6186 (0.0337)**
AutoNN	0.6238 (0.0072)	0.6432 (0.0024)	0.9299 (0.0247)	0.7572 (0.0077)	0.6031 (0.0127)	0.5972 (0.0124)	0.6665 (0.0106)	0.7525 (0.0178)	0.70 01 (0.0108)	0.5666 (0.0135)