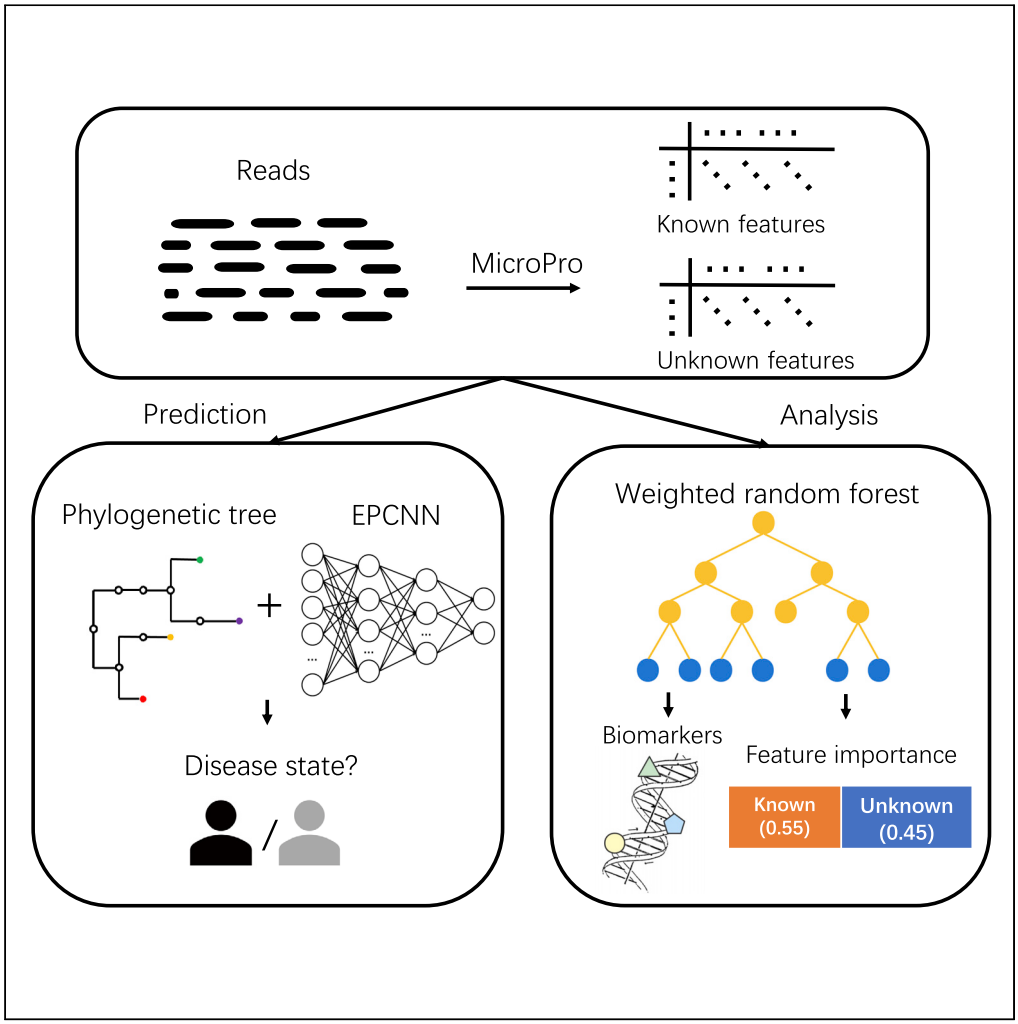
iScience

CellPress
OPEN ACCESS

Article

Human disease prediction from microbiome data by multiple feature fusion and deep learning



Xingjian Chen, Zifan Zhu, Weitong Zhang, Yuchen Wang, Fuzhou Wang, Jianyi Yang, Ka-Chun Wong

kc.w@cityu.edu.hk

Highlights

Both known and unknown abundance profiles are utilized for human disease prediction

Two taxonomic representation methods combined with deep learning are introduced

The weighted random forest complements the poor interpretability of deep learning

MetaDR achieves competitive prediction performance with less running time

Chen et al., iScience 25, 104081
April 15, 2022
https://doi.org/10.1016/j.isci.2022.104081

iScience

CellPress
OPEN ACCESS

Article

Human disease prediction from microbiome data by multiple feature fusion and deep learning

Xingjian Chen,[1] Zifan Zhu,[2] Weitong Zhang,[1] Yuchen Wang,[1] Fuzhou Wang,[1] Jianyi Yang,[3] and Ka-Chun Wong[1,4,5,*]

SUMMARY

Human disease prediction from microbiome data has broad implications in metagenomics. It is rare for the existing methods to consider abundance profiles from both known and unknown microbial organisms, or capture the taxonomic relationships among microbial taxa, leading to significant information loss. On the other hand, deep learning has shown unprecedented advantages in classification tasks for its feature-learning ability. However, it encounters the opposite situation in metagenome-based disease prediction since high-dimensional low-sample-size metagenomic datasets can lead to severe overfitting; and black-box model fails in providing biological explanations. To circumvent the related problems, we developed MetaDR, a comprehensive machine learning-based framework that integrates various information and deep learning to predict human diseases. Experimental results indicate that MetaDR achieves competitive prediction performance with a reduction in running time, and effectively discovers the informative features with biological insights.

INTRODUCTION

The rapid development of sequencing technologies has yielded myraids of microbial data; and mounting evidence correlates microbes with various diseases (e.g., diabetes, allergy, and cancer) (Livanos et al., 2016; Huang et al., 2017; Zhou and Gallins, 2019). Compared with 16S rRNA sequencing which only targets 16S rRNA genes, recent shotgun metagenomic sequencing can provide sample information by sequencing all given genomic DNA from a sample, thus it is becoming increasingly popular and adopted in microbiome research (Zeller et al., 2014; Zhu et al., 2019). In the past several years, some studies indicate that the metagenomic predictive study can be formulated as a supervised learning task based on a species-by-sample matrix (i.e., microbial abundance profiles), which makes it possible to predict human diseases from metagenome-based data (Vangay et al., 2019; Manzoor et al., 2020).

Briefly, microbial abundance profiles can be obtained by two different strategies (Zhu et al., 2019). The first is reference-based methods, while the second is *de novo* assembly-based methods (Chikhi and Rizk, 2013; Xing et al., 2017; Knight et al., 2018). The reference-based methods estimate the microbial abundances by mapping sequencing reads to existing databases (e.g., NCBI RefSeq (Pruitt et al., 2014)). However, some reads may not be mapped and thus the relevant information is lost (Knight et al., 2018). In contrast, the *de novo* assembly-based methods consider all reads to create metagenome-assembled genomes (MAGs) and estimate their abundances (Chikhi and Rizk, 2013). Nevertheless, they are computationally time-consuming as all the reads are considered (Xing et al., 2017). The abundance profiles obtained by these two strategies can be defined as known and unknown microbial features (Zhu et al., 2019). Theoretically, both known and unknown microbial features can be utilized as the input to train a supervised learning model for disease classification. However, most existing methods are based on the abundance profiles of known microbial organisms obtained by reference-based methods since ,compared to *de novo* assembly-based methods, the reference-based methods can provide more biological insights from the existing taxonomic annotations (Kim et al., 2016; Knight et al., 2018; Vangay et al., 2019). On the other hand, the reference-based methods are computationally efficient, although the incomplete mappings can lead to the loss of valuable information (Zhu et al., 2019).

After obtaining the microbial features from sequencing reads, traditional machine learning-based methods combine the features with the corresponding labels to train a classifier. For example,

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong SAR

[2]Quantitative and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA

[3]School of Mathematical Sciences, Nankai University, Tianjin, China

[4]Hong Kong Institute for Data Science, City University of Hong Kong, Kowloon Tong, Hong Kong SAR

[5]Lead contact

*Correspondence: kc.w@cityu.edu.hk

https://doi.org/10.1016/j.isci.2022.104081

Table 1. Performance comparison of average AUCs and the standard deviations with the state-of-the-art methods (Bold means the best performance)

	Karlsson_T2D	Qin_T2D	Qin_LC	Zeller_CRC
Method	AUC (Std)	AUC (Std)	AUC (Std)	AUC (Std)
Micro-Pro (Zhu et al., 2019)	0.7581 (0.0753)	0.7252 (0.0642)	0.9386 (0.0165)	0.8780 (0.0447)
MetaML (Pasolli et al., 2016)	0.5184 (0.1257)	0.5290 (0.1202)	0.8755 (0.0342)	0.6874 (0.0578)
DeepMicro (Oh and Zhang, 2020)	0.6251 (0.0867)	0.6284 (0.0734)	0.9001 (0.0264)	0.7208 (0.0502)
MetaNN (Lo and Marculescu, 2019)	0.4896 (0.0482)	0.5105 (**0.0180**)	0.7576 (0.0162)	0.7839 (0.0172)
DeepForest (Zhu et al., 2018)	0.7560 (0.1160)	0.7650 (0.0809)	0.9309 (0.0389)	0.8657 (0.0361)
WRF	0.7118 (**0.0236**)	0.6738 (0.0205)	0.9341 (**0.0043**)	**0.9416** (**0.0074**)
EPCNN	**0.7890** (0.0559)	**0.8131** (0.0238)	**0.9535** (0.0074)	0.9063 (0.0127)

Qin et al. (2012) predicted liver cirrhosis based on a support vector machine (SVM) using gene markers. Zeller et al. (2014) introduced a LASSO-based model to identify whether the subject has colorectal cancer. Harris et al. (2019) trained a random forest to predict patient phenotypes based on their metagenomic profiles. Although these methods achieved satisfactory results in their respective applications, the input of them is essentially tabular data (i.e., a species-by-sample matrix) that assumes feature independence and neglects the hierarchical relationships among the neighboring taxa (e.g., phylogenetic arrangement and order on a reference tree) (Harris et al., 2019). Therefore, it has been proposed to integrate the phylogenetic relationship with deep learning to improve the prediction performance. For example, Reiman et al. (2020) introduced PopPhy-CNN to transform the phylogenetic tree into an abundance matrix, combining with a convolutional neural network for disease classification. Oudah and Henschel (2018) proposed a similar taxonomy-aware algorithm for feature engineering to exploit the phylogenetic hierarchy for metagenome-based phenotype classification. The same approach was also leveraged by Bogart et al. (2019) for human host status inference. Although the embedded architecture can successfully integrate the phylogenetic structure into microbial features, however, it also introduces extra noises when constructing the phylogenetic matrix, since some species in the phylogenetic tree may not originally exist in microbial features. Recently, Fioravanti et al. (2018) combined the patristic distance and multidimensional scaling to embed the phylogenetic tree into the Euclidean space. However, the feature transformation results in high time complexity and the model design is difficult to optimize. In addition, both Lo and Marculescu (2019) and Nguyen and Zucker (2019) utilized the alphabet-sorting approach to sort the microbial features based on their taxonomic ranks (i.e., phylum, class, order, family, and genus). This method extracts the phylogenetic information by congregating the neighboring taxa with similar species names; however, it still neglects the relationship between taxa that have the same ancestor but the names are farther in alphabetical order.

On the other hand, throughout the related studies in the past several years, most existing methods are tailored for focused studies; when applied to other datasets, they can suffer from underfitting for different reasons (e.g., small data size adaptation issue, noisy features, lack of data processing, etc.) (Pasolli et al., 2016). Besides, the training and prediction time for the existing methods cannot fulfill the clinic requirements anymore (Zhou and Gallins, 2019), necessitating a new framework that can be scalable with the rapid increase of microbiome data.

Recently, deep learning has shown unprecedented advantages for its feature-learning ability and accelerated training technique. One of the most popular deep learning algorithms is convolutional neural networks (CNNs), which have been utilized in a variety of medical imaging applications including tumor classification, mutation prediction, and cancer subtype detection (O'Shea and Nash, 2015). Compared with traditional hand-craft feature-based machine learning methods, CNN can capture the spatial locality in the image by extracting the spatial correlation among neighboring pixels (LeCun et al., 2015). Inspired by its structure, we propose a taxonomic representation approach for microbial features, which can keep the

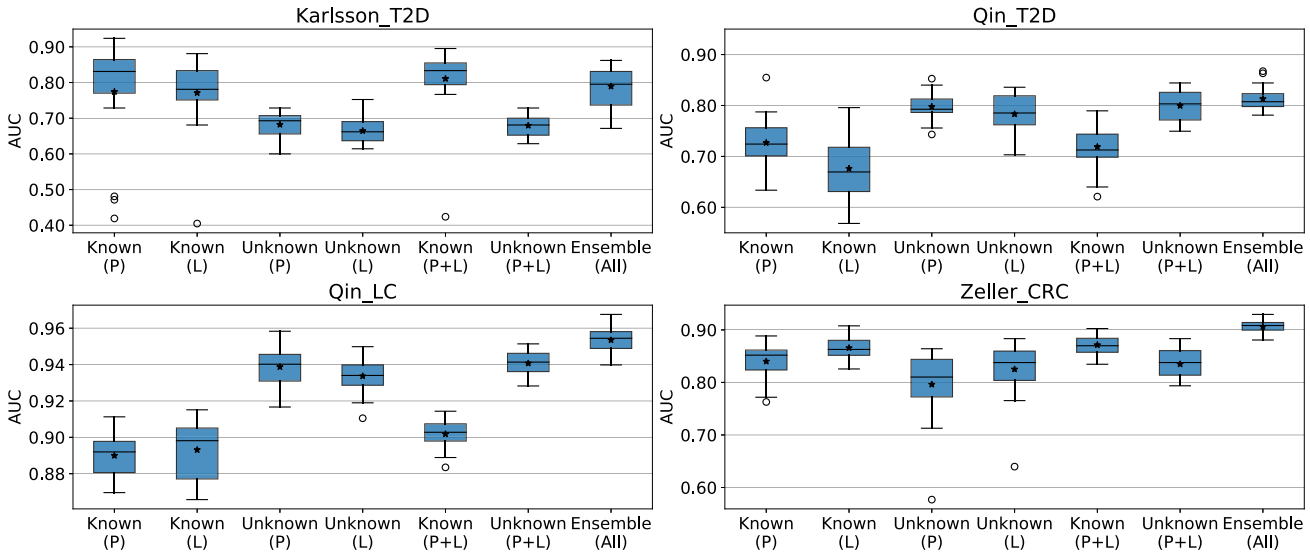


Figure 1. Boxplot of the comparison result of single and ensemble-based models

Known (P), Known (L), Unknown (P), and Unknown (L) are the results obtained by single models which are known-profile-based model with postorder representation, known-profile based-model with level-traversal representation, unknown-profile-based model with postorder-traversal representation, and unknown-profile-based model with level-traversal representation. Known (P + L) is the averaged result obtained by ensembling the first two single models, and Unknown (P + L) is the averaged result obtained by ensembling the last two single models. Ensemble (All) is the final result obtained by ensembling all four models (EPCNN). Asterisk means the average value.

phylogenetic relationship and enable the utilization of CNN for disease classification. Besides, an ensemble model is constructed to enrich the microbial feature representation and reduce the risk of overfitting. In view of the black-box property of deep learning, we introduce a weighted random forest to extract biological insights. The study is conducted in the context of the shotgun-sequenced metagenomic datasets.

RESULTS

Comparison with state-of-the-art predictors

In order to assess the effectiveness of our framework, we compare both weighted random forest (WRF) and ensemble phylogenetic convolutional neural network (EPCNN) with five state-of-the-art predictors, which are MetaML (Pasolli et al., 2016), DeepForest (Zhu et al., 2018), MetaNN (Lo and Marculescu, 2019), Micro-Pro (Zhu et al., 2019), and DeepMicro (Oh and Zhang, 2020). The five comparison tools all have been published recently and declared to achieve state-of-the-art performance. Since the utilized datasets in original publications are all based on different analysis pipelines and evaluation standards, to keep a fair comparison, we ensure the input for all the comparison predictors are consistent which includes abundance profiles of both known and unknown microbial organisms, and the obtained results for different predictors are all based on the same evaluation standard. In addition, we strictly followed the suggested parameters and model architectures with the best performance in original experiments (see STAR Methods). The evaluation metric is the receiver operating characteristic (ROC) curve (AUC). As we can see in Table 1, our WRF and EPCNN achieve superior performance compared with the state-of-the-art predictors. Especially, the performance of our EPCNN is the best on three out of four datasets. The improvements for Karlsson_T2D and Qin_T2D datasets are remarkable as our EPCNN increases the average AUCs from 0.4896 to 0.7890 and 0.5105 to 0.8183, respectively, compared to MetaML and MetaNN. For Qin_LC dataset, even if the AUCs of DeepForest, DeepMicro, and Micro-Pro are all higher than 0.9 which are close to the perfect prediction, the AUC of our EPCNN still increases to 0.9535, achieving the improvements of 2%–5%. Compared with MetaNN based on the same deep learning method, the AUCs of our EPCNN increase by about 20% across the datasets by utilizing ensemble strategy to reduce the overfitting.

Besides, it can be noticed that the performance of Micro-Pro and DeepForest is better than that of the benchmarks. We attribute this to the mapping weights utilized in Micro-Pro and phylogenetic information fused in

Table 2. Average mapping ratios and average feature weights for four datasets

Dataset	Mean mapping rate (reads)	Known mapping ratio (profiles)	Weight for known profiles	Weight for unknown profiles
Karlsson_T2D	0.3410	0.6055	0.4699	0.5301
Qin_T2D	0.4225	0.6910	0.4614	0.5386
Qin_LC	0.4356	0.6789	0.5072	0.4928
Zeller_CRC	0.3578	0.5824	0.4957	0.5043

DeepForest. On the other hand, DeepMicro utilizes a convolutional autoencoder (CAE) to reduce the feature dimensions but it also deleted some important features. Therefore, the performance is unsatisfying, especially on Karlsson_T2D and Qin_T2D datasets with the originally high-dimensional features. Besides, the performance of MetaML is bad; we infer that SVM may not be applicable for classifying metagenomic data.

It is notable that the performance of EPCNN is no better than that of WRF on Zeller_CRC dataset, the reason may be that the overall feature dimension of this dataset is the highest among all datasets. Therefore, the introduced noise amount may also be the highest during the taxonomic representation since some unrelated profiles of microbial organisms may be convoluted in the training process of CNN.

Comparison with different microbial features and taxonomic representation

In addition, we compare the performance of single models and ensemble models based on different microbial features and taxonomic representation. As seen in Figure 1, the ensemble strategy obviously improves the prediction performance compared with the single feature-based models on three out of four datasets. The average AUCs are increased by 5%–20% on different datasets. However, for Karlsson_T2D dataset, we can see that the performance of EPCNN was slightly worse than Known (P + L), and no matter for the known-profile-based model Known (P) or Known (L) with different taxonomic representation, the average AUC is almost 10% higher than the performance of unknown-profile-based model Unknown (P) and Unknown (L). We infer that the abundance profiles of unknown microbial organisms on this dataset may have a negative influence on the abundance profiles of known microbial organisms, thereby causing the terrible performance of ensemble results. Besides, we can observe that for Karlsson_T2D dataset, the mapping ratio of its MAGs are the least among all the datasets (Table 2), which indicates that the taxonomic information of the abundance profiles of unknown microbial organisms on this dataset may not be accurate to support the model performance. For the performance of different features in different taxonomic representations, it can be observed that the abundance profiles of unknown microbial organisms with postorder-traversal representation can contribute more in CNN models because compared with the performance of abundance profiles of known microbial organisms, the average AUC score was increased by 4%–12% on two of the four datasets.

Influence of taxonomic information

Regarding the benefits of the proposed taxonomic information, we have additionally conducted extra experiments on four datasets to compare the prediction performance with/without taxonomic tree spatial information. The results are shown in Table S1. Herein, Known (w/P), Known (w/L), Unknown (w/P), and Unknown (w/L) are the prediction results obtained by single models which are known-profile-based model with postorder representation, known-profile-based model with level-traversal representation, unknown-profile-based model with postorder-traversal representation, and unknown-profile-based model with level-traversal representation. Known (w/o) represents the prediction result of known profiles without taxonomic tree spatial information, and Unknown (w/o) represents the prediction results of unknown profiles without taxonomic tree spatial information. It can be observed that the taxonomic tree spatial information does increase the final prediction performance by 2%–5% across different datasets, although there is a performance discrepancy between different profiles with different taxonomic tree spatial information.

Prediction based on different architectures

Regarding the model architecture of EPCNN, we validate and compare a series of deep neural network (DNN)-based models which are multilayer perceptron (MLP), one-dimensional (1D)-CNN, and two-dimensional (2D)-CNN with different numbers of layers and nodes (Table S2 and STAR Methods). As can be seen in Table S3, comparing the performance of MLP_2layer and MLP_3layer, we can see that

Table 3. Number of selected informative features across four datasets (among top 30 features)

	Raw		Micro-Pro		WRF	
Dataset	Known	Unknown	Known	Unknown	Known	Unknown
Karlsson_T2D	0	30	5	25	3	27
Qin_T2D	0	30	0	30	4	26
Qin_LC	4	26	10	20	7	23
Zeller_CRC	4	26	3	27	6	24

the performance is slightly better as the model complexity increases. Although the performance is not good without embedding the phylogenetic information, however, the standard deviations of MLP are the least and the model is the most stable among all the comparison architectures. Comparing the performance of CNN1D_conv3_pool2 and CNN1D_conv5_pool3, they suffer from severe overfitting problems, which achieve similar results as MetaNN in comparison with the state-of-the-art predictors. However, we can see that the performance increases as the filter and pooling sizes increase. We believe that the complicated model structure and large filter size can improve the prediction performance since larger sizes mean larger receptive fields and more information. The comparison results between CNN2D_conv3_pool3 and EPCNN can also support that. However, the improvement is not absolute because the performance of Qin_LC dataset slightly drops. The same tendency also appears on CNN2D_ResNet. Herein, ResNet is the most complicated architecture among all the comparison models, but it also suffers from severe overfitting, resulting in performance degradation, which indicates that complex architectures may not be applicable to all the datasets. The reason for the bad performance of ResNet may also be from its pre-trained ImageNet weights since ImageNet has a totally different feature distribution from our metagenomic data. It is notable that there is a huge difference in the results between 1D-CNN and 2D-CNN models; we attribute that to the matrix conversion which can simulate the image structure to extract both local and global features between the neighboring taxa, making it more suitable for CNN classification.

Prediction based on OTU-level data

Besides, our MetaDR is also scalable and can be generalized to the task based on operational taxonomic units (OTUs). The difference is, because the OTU features are obtained by the 16S amplicon analysis pipeline which is different from the previous shotgun metagenomic analysis pipeline, we cannot generate the abundance profiles of unknown microbial organisms from 16S amplicon analysis. However, because the taxonomic tree information of OTUs can be still obtained from 16S amplicon analysis, our MetaDR is able to be utilized in the OTU-based prediction. Herein, we evaluated our MetaDR on one public available OTU-based 16S rRNA sequencing dataset to predict T2D and compared it with PopPhy-CNN (Karlsson et al., 2013; Qin et al., 2012; Reiman et al., 2020), which is one of the state-of-the-art tools for OTU-based prediction. Surprisingly, our MetaDR still achieved an average AUC of 0.72210 while PopPhy-CNN is simply 0.6810, which reflects that our framework can also have a promising generalization and better performance on OTU-level data.

Analyses of informative features

To evaluate our WRF, we calculated the mean mapping rate of reads (to reference database), the ratio of the features with known taxonomic annotations, and the obtained weights for each feature (Table 2). The predicted weights for abundance profiles of both known and unknown microbial organisms represent the importance of the different microbial profiles. We can observe that for both Karlsson_T2D and Qin_T2D datasets, the weights of abundance profiles of unknown microbial organisms are all higher than the weights of abundance profiles of known microbial organisms, while for Qin_LC and Zeller_CRC datasets, the abundance profiles of known and unknown microbial organisms have approximately equal weights. The tendency indicates that for T2D, the abundance profiles of unknown microbial organisms are more important than the abundance profiles of known microbial organisms.

It is noticeable that for Karsson_T2D dataset, the abundance profiles of unknown microbial organisms have higher weights based on WRF than those for the abundance profiles of known microbial organisms, yet in Figure 1, the models using abundance profiles of unknown microbial organisms performed worse

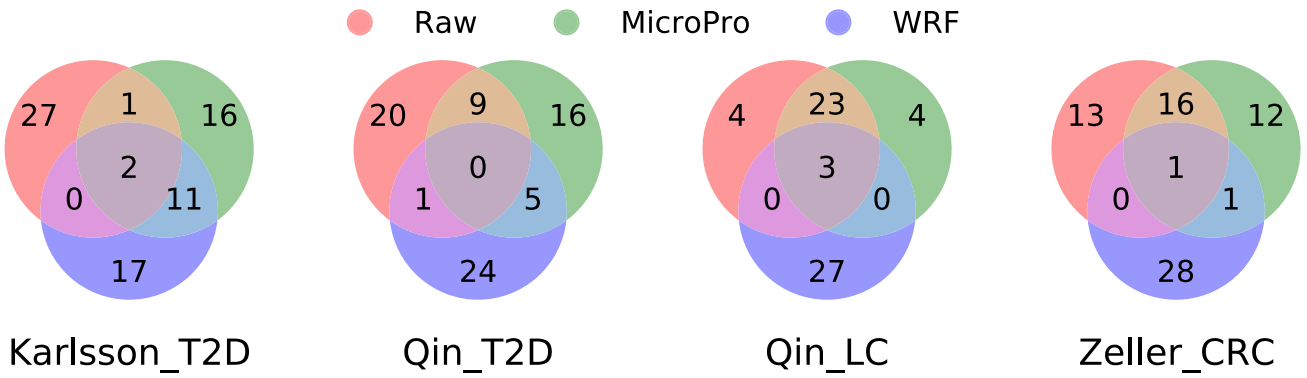


Figure 2. Venn diagrams for the selected top 30 informative features by three methods

than the models using abundance profiles of known microbial organisms. We infer the discrepancy is because WRF and CNN have totally different input and training strategies. Although the two models utilized the abundance profiles of both known and unknown microbial organisms, the input of WRF is substantially a one-dimensional vector assuming feature independence. In contrast, the input for CNN is a matrix that can extract the spatial correlation among neighboring taxa. Besides, WRF is an ensemble-based method composed of many decision trees, while CNN is a deep learning-based method that automatically learns the features. For the results of Qin_LC dataset, the mapping ratio for the abundance profiles of known microbial organisms to the reference database is the largest. Analogously, the weight for the abundance profiles of known microbial organisms is also larger than the weight of the abundance profiles of unknown microbial organisms, which can further support the rationality and accuracy of our results.

Additionally, we explored the microbial features that were significantly associated with a certain disease in this study. The top 30 features associated with highly ranked importance coefficients are selected for further analyses. In order to assess the effectiveness of the proposed WRF for feature selection, we compared it with two other feature selection methods (Raw and Micro-Pro) based on different feature fusion strategies (see STAR Methods). Besides, we also list the numbers of selected features of known and unknown microbial organisms among the top 30 informative features for comparison (Table 3). As we can see in Venn diagrams (Figure 2), the three methods have completely different results for the extracted top 30 informative features. The overlaps of three methods on four datasets are all small, with less than five same features. However, the overlap informative features between Raw and Micro-Pro are large, with more than five informative features in common on three out of four datasets. Contrarily, the overlaps between WRF and Raw or Micro-Pro are small. Only for Karlsson_T2D dataset, they share more than 10 informative features.

As can be seen in Table 3, our WRF is the only one that discovered both known and unknown informative features across all the datasets. For the selected unknown informative features, because the taxonomic assignments of MAGs are not accurate, they need to be evaluated by further analyses. However, for the selected features of known microbial organisms, compared with Raw and Micro-Pro, our WRF successfully identified *Gordonibacter urolithinfaciens* and *Lactobacillus salivarius* in both two T2D datasets (Sáez-Lara et al., 2016; Corrêa et al., 2019); Erysipelotrichaceae *bacterium* in LC dataset (Kaakoush, 2015; Lu et al., 2015); and *Anaerostipes hadrus, Dialister pneumosintes, Akkermansia muciniphila,* and *Streptococcus* in CRC dataset (Boleij et al., 2011; Dingemanse et al., 2015; Louis and Flint, 2017; Li et al., 2019). These biomarkers all have been proved to have associations with the related diseases. However, the other two methods (Raw and Micro-Pro) did not successfully identify them in the top 30 features, which indicate that our proposed WRF is more effective than the other two methods. In addition, the proposed WRF can also be utilized to predict diseases. A detailed comparison of the average AUCs can be found in Table S4.

Execution time

We report the execution time of different baseline predictors to compare with our methods (Table 4 and STAR Methods). For the running time of our EPCNN, we beat all the other predictors except MetaML. For MetaML,

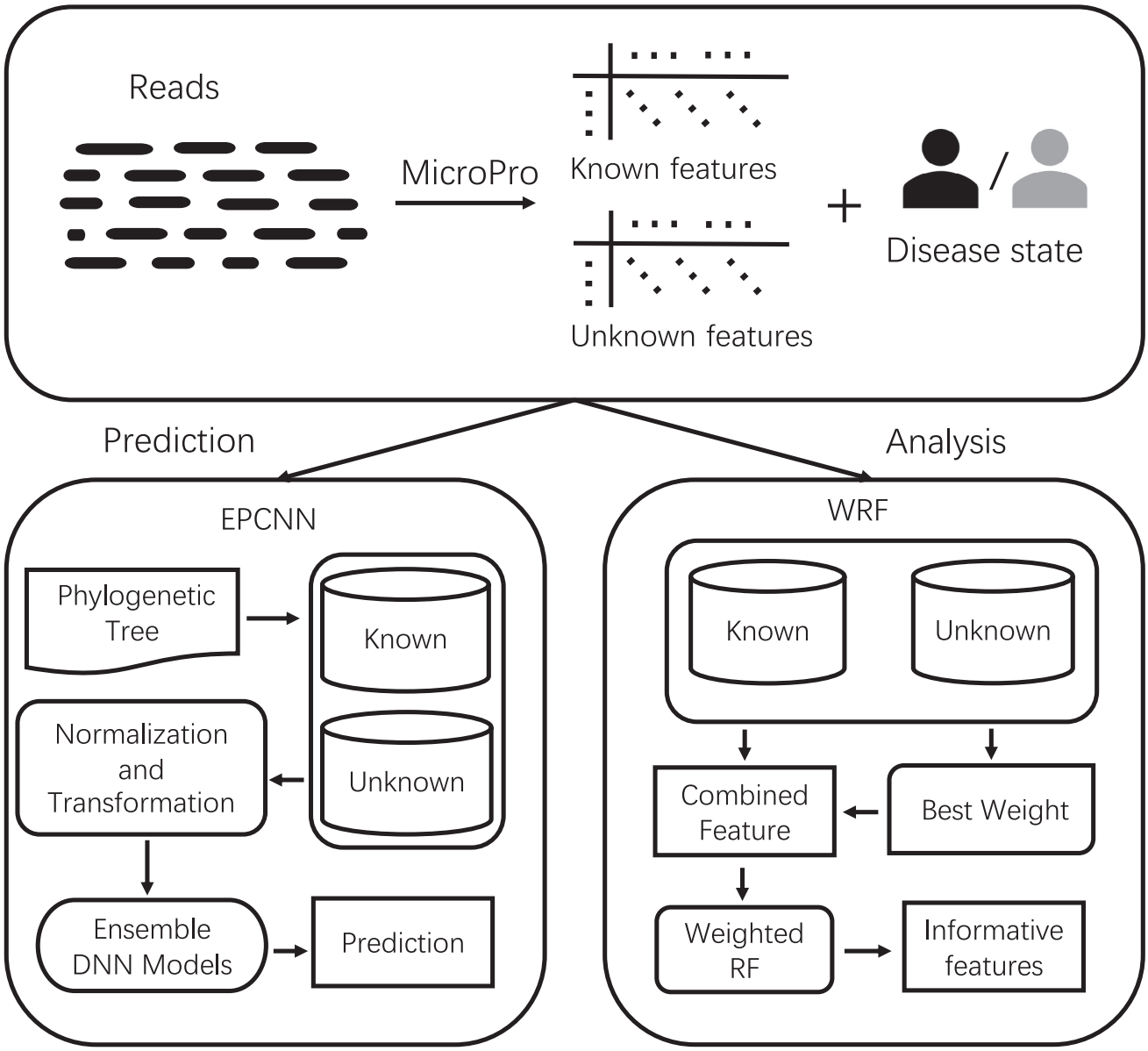


Figure 3. Overview of the proposed MetaDR

[Top] At the beginning, the raw sequencing reads are sent to Micro-Pro to extract the abundance profiles of known and unknown microbial organisms. [Left] The taxonomic information is extracted from a phylogenetic tree, combined with the ensemble convolutional neural network to train the models. [Right] The weighted random forest obtains the weight for each feature and outputs the informative features for biological explanations.

although it is the fastest one, it achieves the worst prediction results among all the comparison tools (as evidenced in Table 1). In summary, although the proposed EPCNN is not the fastest, however, considering the prediction performance, our EPCNN may be the only one that can make a perfect trade-off between the prediction performance and the running time. Besides, it is more scalable as it can be trained on GPUs.

DISCUSSION

Deep learning has been proved to be promising on large and complex classification tasks such as natural language processing and image recognition. However, it is rarely applied in metagenomic research since high-dimensional low-sample-size metagenomic datasets can lead to severe overfitting, and the black-box model fails to provide biological explanations. To circumvent the related problems, we propose MetaDR, a

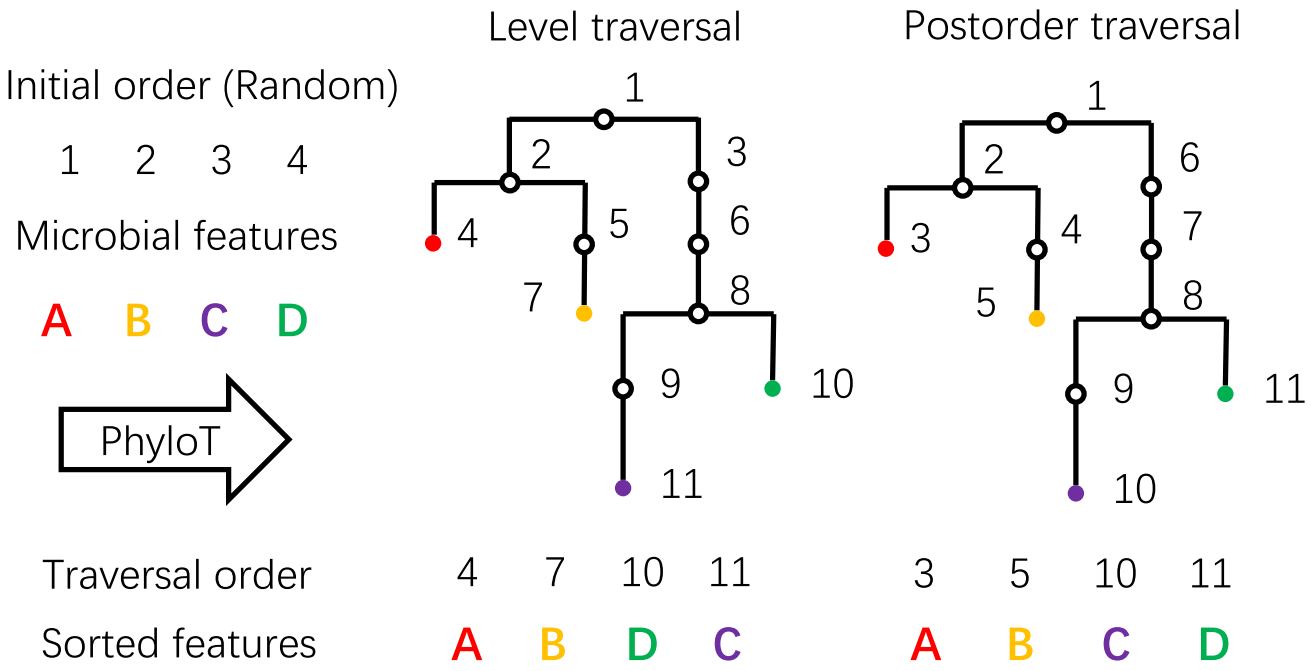


Figure 4. Traversal order of the taxonomic representation

Given the initial random order of the microbial features as 1, 2, 3, and 4, and the corresponding taxonomic name "A", "B", "C", and "D", PhyloT constructs a phylogenetic tree with the taxonomic name as input. After traversing the phylogenetic tree, new sorting orders are kept to replace the initial orders.

comprehensive machine learning-based framework that integrates various information and deep learning to predict human diseases.

In the current work, the experiments are conducted in the context of shotgun-sequenced metagenomic datasets. However, given that the 16S amplicon analysis pipeline is also popular in microbiome research, we also evaluated the generalization of our MetaDR on OTU-level data. The result indicates that the proposed MetaDR is also scalable and can perform well on OTU-level data. Regarding the benefits of the proposed taxonomic information, it can be observed that some features (the abundance profiles of known/unknown microbial organisms) with specific taxonomic tree spatial information may have a negative influence on prediction (Figure 1). However, the additional ablation experiments (see STAR Methods) demonstrate that the taxonomic tree spatial information does increase the final prediction performance by 2%–5% across different datasets, although there is a performance discrepancy between different profiles with different taxonomic tree spatial information.

On the other hand, the proposed weight random forest successfully explains different features (i.e., the abundance profiles of known and unknown microbial organisms), and further proves that unknown profiles are necessary for human disease prediction. In addition, the obtained weight for each kind of profile also supports the performance of our EPCNN. In this study, the proposed weight random forest complements the poor interpretability of the black-box deep learning model (i.e., our EPCNN), since CNN cannot explicitly identify the feature importance when conducting the prediction. Additionally, the proposed weight random forest also helps to explore the microbial features that were significantly associated with a certain disease and successfully identifies some potential biomarkers for further analyses. Besides, the performance of our weight random forest is promising, which can be an alternative to the black-box CNN model when users prefer a more strongly explanatory predictor.

In the future, we consider applying more advanced deep learning models, such as recurrent neural network (RNN) and graph neural network (GNN) to further improve the disease prediction in metagenomics, combining with post hoc interpretable algorithms to explore more representative features to lay the groundwork for future development of specific disease biomarkers. We believe that MetaDR can contribute to microbiome research and help explore the patterns among the microbial features and its

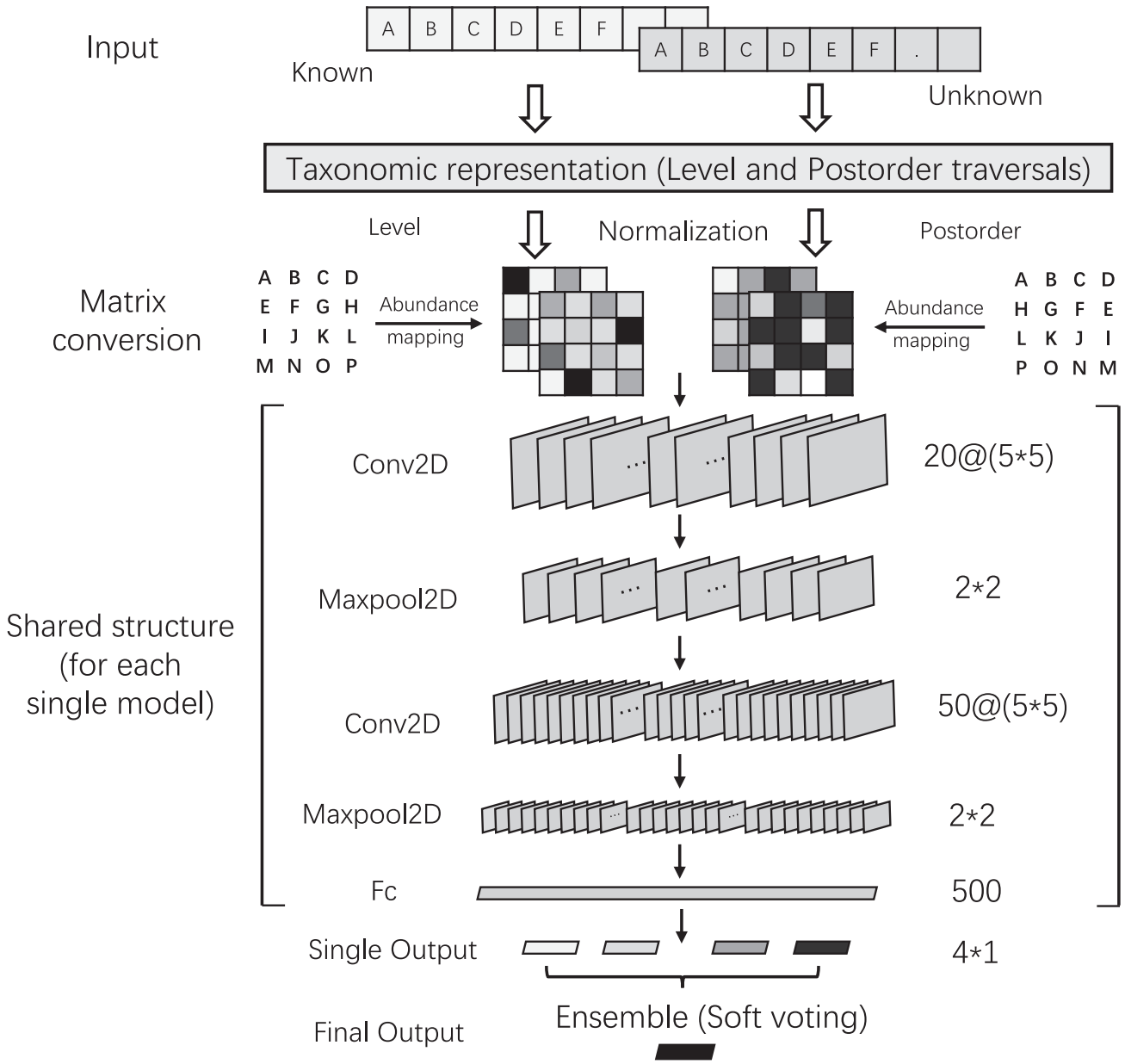


Figure 5. The proposed neural network architecture of EPCNN

From top to bottom, the abundance profiles of known and unknown microbial organisms are transformed into matrices according to different taxonomic representations. Then, the species abundances are mapped to microbial features with specific taxonomic annotations, and the abundance matrices are converted to grayscale images as the input of CNN. Four single models are then trained with the same CNN architecture. Finally, the outputs of the four models are averaged to obtain the ensemble prediction result.

host disease, paving the metagenomic diagnosis foundation for personalized medicine in the coming future.

Limitation of the study

The current study is limited by the absence of a truly independent validation set. In addition, while the ensemble model can achieve state-of-the-art performance, the running time complexity can be high.

Table 4. Running time (s) for different predictors

Method	Karlsson_T2D	Qin_T2D	Qin_LC	Zeller_CRC
Micro-Pro (Zhu et al., 2019)	153	277	236	267
MetaML (Pasolli et al., 2016)	7	14	37	40
DeepMicro (Oh and Zhang, 2020)	7517	9201	13,327	7576
MetaNN (Lo and Marculescu, 2019)	175	221	309	232
DeepForest (Zhu et al., 2018)	588	556	546	516
WRF	610	853	2025	2270
EPCNN	192	215	225	229

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Datasets
  - Overview of MetaDR
  - Ensemble phylogenetic convolutional neural network
  - Weighted random forest
  - Description of baseline approaches
  - Ablation settings
  - Development environment
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Evaluation metrics

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104081.

ACKNOWLEDGMENTS

The three anonymous reviewers are thanked for their time and efforts, improving numerous aspects of the study. This research was substantially sponsored by the research projects (Grant No. 32000464 and Grant No. 32170654) supported by the National Natural Science Foundation of China and was substantially supported by the Shenzhen Research Institute, City University of Hong Kong. The work described in this paper was substantially supported by the grant from the Research Grants Council of the Hong Kong Special Administrative Region [CityU 11200218], one grant from the Health and Medical Research Fund, the Food and Health Bureau, The Government of the Hong Kong Special Administrative Region [07181426], and the funding from Hong Kong Institute for Data Science (HKIDS) at City University of Hong Kong. The work described in this paper was partially supported by the grants from City University of Hong Kong (CityU 11202219, CityU 11203520, CityU 11203221).

Table 5. Disease types, data sizes, and feature dimensions of four datasets

Dataset	Disease	Known features	Unknown features	Number of controls	Number of cases
Karlsson_T2D (Karlsson et al., 2013)	T2D	785	594	43	53
Qin_T2D (Qin et al., 2012)	T2D	925	486	74	71
Qin_LC (Qin et al., 2014)	LC	936	506	114	123
Zeller_CRC (Zeller et al., 2014)	CRC	1287	1026	93	91

AUTHOR CONTRIBUTIONS

K.-C.W. conceived the study; X.C. designed and implemented the algorithms; X.C., Z.Z., and W.Z. carried out the result analysis; X.C., Z.Z., W.Z., Y.W., F.W., and J.Y. developed the software; X.C. and K.-C.W. wrote the manuscript; K.-C.W. supervised the study. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interest.

Received: April 11, 2021
Revised: September 16, 2021
Accepted: March 13, 2022
Published: April 15, 2022

REFERENCES

Bogart, E., Creswell, R., and Gerber, G.K. (2019). Mitre: inferring features from microbiota time-series data linked to host status. Genome Biol. *20*, 186. https://doi.org/10.1186/s13059-019-1788-y.

Boleij, A., van Gelder, M.M., Swinkels, D.W., and Tjalsma, H. (2011). Clinical importance of streptococcus gallolyticus infection among colorectal cancer patients: systematic review and meta-analysis. Clin. Infect. Dis. *53*, 870–878. https://doi.org/10.1093/cid/cir609.

Breiman, L. (2001). Random forests. Machine Learn. *45*, 5–32. https://doi.org/10.1023/a:1010933404324.

Chikhi, R., and Rizk, G. (2013). Space-efficient and exact de bruijn graph representation based on a bloom filter. Algorithms Mol. Biol. *8*, 22. https://doi.org/10.1186/1748-7188-8-22.

Corrêa, T.A.F., Rogero, M.M., Hassimotto, N.M.A., and Lajolo, F.M. (2019). The two-way polyphenols-microbiota interactions and their effects on obesity and related metabolic diseases. Front. Nutr. *6*, 188. https://doi.org/10.3389/fnut.2019.00188.

Dimitriadis, S.I., Liparas, D., and Tsolaki, M.N.; Alzheimer's Disease Neuroimaging Initiative (2018). Random forest feature selection, fusion and ensemble strategy: combining multiple morphological MRI measures to discriminate among healhy elderly, MCI, cMCI and alzheimer's disease patients: from the alzheimer's disease neuroimaging initiative (ADNI) database. J. Neurosci. Methods *302*, 14–23. https://doi.org/10.1016/j.jneumeth.2017.12.010.

Dingemanse, C., Belzer, C., van Hijum, S.A., Günthel, M., Salvatori, D., den Dunnen, J.T., Kuijper, E.J., Devilee, P., de Vos, W.M., van Ommen, G.B., and Robanus-Maandag, E.C. (2015). Akkermansia muciniphila and Helicobacter typhlonius modulate intestinal tumor development in mice. Carcinogenesis *36*, 1388–1396. https://doi.org/10.1093/carcin/bgv120.

Fioravanti, D., Giarratano, Y., Maggio, V., Agostinelli, C., Chierici, M., Jurman, G., and Furlanello, C. (2018). Phylogenetic convolutional neural networks in metagenomics. BMC Bioinf. *19*, 49. https://doi.org/10.1186/s12859-018-2033-5.

Harris, Z.N., Dhungel, E., Mosior, M., and Ahn, T.-H. (2019). Massive metagenomic data analysis using abundance-based machine learning. Biol. Direct *14*, 12. https://doi.org/10.1186/s13062-019-0242-0.

Huang, Y.J., Marsland, B.J., Bunyavanich, S., O'Mahony, L., Leung, D.Y., Muraro, A., and Fleisher, T.A. (2017). The microbiome in allergic disease: current understanding and future opportunities—2017 PRACTALL document of the American Academy of Allergy, Asthma & Immunology and the European Academy of Allergy and Clinical Immunology. J. Allergy Clin. Immunol. *139*, 1099–1110. https://doi.org/10.1016/j.jaci.2017.02.007.

Kaakoush, N.O. (2015). Insights into the role of erysipelotrichaceae in the human host. Front. Cell. Infect. Microbiol. *5*, 84. https://doi.org/10.3389/fcimb.2015.00084.

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359. https://doi.org/10.7717/peerj.7359.

Kang, Q., Meng, J., Cui, J., Luan, Y., and Chen, M. (2020). Pmlipred: a method based on hybrid model and fuzzy decision for plant mirna–lncrna interaction prediction. Bioinformatics *36*, 2986–2992. https://doi.org/10.1093/bioinformatics/btaa074.

Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C.J., Fagerberg, B., Nielsen, J., and Bäckhed, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature *498*, 99–103. https://doi.org/10.1038/nature12198.

Kim, D., Song, L., Breitwieser, F.P., and Salzberg, S.L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. *26*, 1721–1729. https://doi.org/10.1101/gr.210641.116.

Knight, R., Vrbanac, A., Taylor, B.C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.I., McDonald, D., et al. (2018). Best practices for analysing microbiomes. Nat. Rev. Microbiol. *16*, 410–422. https://doi.org/10.1038/s41579-018-0029-9.

Kursa, M.B., and Rudnicki, W.R. (2010). Feature selection with the boruta package. J. Stat. Softw. *36*, 1–13. https://doi.org/10.18637/jss.v036.i11.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444. https://doi.org/10.1038/nature14539.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In Proceedings of the IEEE, *86*Proceedings of the IEEE (IEEE), pp. 2278–2324. https://doi.org/10.1109/5.726791.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. Bioinformatics *31*, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.

Li, Y.-D., He, K.-X., and Zhu, W.-F. (2019). Correlation between invasive microbiota in margin-surrounding mucosa and anastomotic healing in patients with colorectal cancer. World J. Gastrointest. Oncol. *11*, 717. https://doi.org/10.4251/wjgo.v11.i9.717.

Livanos, A.E., Greiner, T.U., Vangay, P., Pathmasiri, W., Stewart, D., McRitchie, S., Li, H., Chung, J., Sohn, J., Kim, S., et al. (2016). Antibioticmediated gut microbiome perturbation accelerates development of type 1 diabetes in mice. Nat. Microbiol. *1*, 16140. https://doi.org/10.1038/nmicrobiol.2016.140.

Lo, C., and Marculescu, R. (2019). MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks. BMC Bioinf. *20*, 1–14. https://doi.org/10.1186/s12859-019-2833-2.

Louis, P., and Flint, H.J. (2017). Formation of propionate and butyrate by the human colonic microbiota. Environ. Microbiol. *19*, 29–41. https://doi.org/10.1111/1462-2920.13589.

Lu, H., Qian, G., Ren, Z., Zhang, C., Zhang, H., Xu, W., Ye, P., Yang, Y., and Li, L. (2015). Alterations of Bacteroides sp., Neisseria sp., Actinomyces sp., and Streptococcus sp. populations in the oropharyngeal microbiome are associated with liver cirrhosis and pneumonia. BMC Infect. Dis. *15*, 1–11. https://doi.org/10.1186/s12879-015-0977-x.

Manzoor, S.S., Doedens, A., and Burns, M.B. (2020). The promise and challenge of cancer microbiome research. Genome Biol. *21*, 131. https://doi.org/10.1186/s13059-020-02037-9.

Nguyen, T.H., and Zucker, J.D. (2019). Enhancing metagenome-based disease prediction by unsupervised binning approaches. In 2019 11th International Conference on Knowledge and Systems Engineering (KSE) (IEEE), pp. 1–5. https://doi.org/10.1109/KSE.2019.8919295.

Oh, M., and Zhang, L. (2020). Deepmicro: deep representation learning for disease prediction based on microbiome data. Sci. Rep. *10*, 1–9. https://doi.org/10.1038/s41598-020-63159-5.

Ondov, B., Treangen, T., Melsted, P., Mallonee, A., Bergman, N., Koren, S., and Phillippy, A. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 1–14. https://doi.org/10.1186/s13059-016-0997-x.

O'Shea, K., and Nash, R. (2015). An introduction to convolutional neural networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1511.08458.

Oudah, M., and Henschel, A. (2018). Taxonomy-aware feature engineering for microbiome classification. BMC Bioinf. *19*, 1–13. https://doi.org/10.1186/s12859-018-2205-3.

Pasolli, E., Truong, D.T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput. Biol. *12*, e1004977. https://doi.org/10.1371/journal.pcbi.1004977.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., et al. (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. *42*, D756–D763. https://doi.org/10.1093/nar/gkt1114.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60. https://doi.org/10.1038/nature11450.

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., Wu, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. Nature *513*, 59–64. https://doi.org/10.1038/nature13568.

Reiman, D., Metwally, A.A., Sun, J., and Dai, Y. (2020). PopPhy-CNN: a phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. IEEE J. Biomed. Health Inform. *24*, 2993–3001. https://doi.org/10.1109/JBHI.2020.2993761.

Sáez-Lara, M.J., Robles-Sanchez, C., Ruiz-Ojeda, F.J., Plaza-Diaz, J., and Gil, A. (2016). Effects of probiotics and synbiotics on obesity, insulin resistance syndrome, type 2 diabetes and non-alcoholic fatty liver disease: a review of human clinical trials. Int. J. Mol. Sci. *17*, 928. https://doi.org/10.3390/ijms17060928.

Vangay, P., Hillmann, B.M., and Knights, D. (2019). Microbiome Learning Repo (ML Repo): a public repository of microbiome regression and classification tasks. Gigascience *8*, giz042. https://doi.org/10.1093/gigascience/giz042.

Xing, X., Liu, J.S., and Zhong, W. (2017). Metagen: reference-free learning with multiple metagenomic samples. Genome Biol. *18*, 1–15. https://doi.org/10.1186/s13059-017-1323-y.

Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol. Syst. Biol. *10*, 766. https://doi.org/10.15252/msb.20145645.

Zhou, Y.H., and Gallins, P. (2019). A review and tutorial of machine learning methods for microbiome host trait prediction. Front. Genet. *10*, 579. https://doi.org/10.3389/fgene.2019.00579.

Zhu, Q., Zhu, Q., Pan, M., Jiang, X., Hu, X., and He, T. (2018). The phylogenetic tree based deep forest for metagenomic data classification. In 2018 IEEE international conference on bioinformatics and biomedicine (BIBM) (IEEE), pp. 279–282. https://doi.org/10.1109/BIBM.2018.8621463.

Zhu, Z., Ren, J., Michail, S., and Sun, F. (2019). Micropro: using metagenomic unmapped reads to provide insights into human microbiota and disease associations. Genome Biol. *20*, 1–13. https://doi.org/10.1186/s13059-019-1773-5.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Karlsson_T2D	European Nucleotide Archive (ENA)	ERP002469
Qin_T2D	European Nucleotide Archive (ENA)	SRA045646
Qin_LC	European Nucleotide Archive (ENA)	ERP005860
Zeller_CRC	European Nucleotide Archive (ENA)	ERP005534
Software and algorithms		
MetaDR	This study (https://github.com/Microbiods/MetaDR)	MetaDR (v3.0.0)
MicroPro	https://github.com/zifanzhu/MicroPro	MicroPro (v1.0.1)
Megahit	https://github.com/voutcn/megahit	Megahit (v1.1.3)
MetaBAT2	https://bitbucket.org/berkeleylab/metabat/src/master/	MetaBAT2 (v2.12.1)
Centrifuge	https://ccb.jhu.edu/software/centrifuge/	Centrifuge (v1.0.3)
PhyloT	https://phylot.biobyte.de/	PhyloT (v2)

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr. Ka-Chun Wong (kc.w@cityu.edu.hk).

Materials availability

This study did not generate new biological data.

Data and code availability

The datasets utilized in this study are publicly available. The accession numbers for the datasets are listed in the key resources table.

The current version of MetaDR is implemented in python and can be found at https://github.com/Microbiods/MetaDR.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This paper analyzes existing, publicly available data. The study does not use experimental models typical in life sciences.

METHOD DETAILS

Datasets

We retrieve the datasets from four available shotgun-sequenced metagenomic studies which are related to three different diseases: Type 2 Diabetes (T2D) (Qin et al., 2012; Karlsson et al., 2013), Liver Cirrhosis (LC) (Qin et al., 2014) and Colorectal Cancer (CRC) (Zeller et al., 2014). The raw sequencing reads can be downloaded from European Nucleotide Archive (ENA) database (https://www.ebi.ac.uk/ena), and the related accession numbers can be found in Table S5.

In this study, since we focus on developing a disease-related predictive framework rather than the sequence feature extraction pipeline, we directly utilized our recently published pipeline MicroPro (Zhu et al., 2019) to obtain the abundance profiles of known and unknown microbial organisms from the raw sequencing reads. MicroPro is a software to perform profiling of both known and unknown microbial

organisms for metagenomic datasets. For abundance profiles of known microbial organisms, MicroPro applied Centrifuge (Kim et al., 2016) to map the reads to NCBI RefSeq database and estimate their abundances. In terms of Centrifuge command, we set flag "-q" which indicated the input was in FASTQ format, and the other arguments were set as default. Regarding the reads that cannot be mapped to NCBI RefSeq, Megahit (Li et al., 2015) and MetaBAT2 (Kang et al., 2019) are executed to generate MAGs and extract the abundance profiles of unknown microbial organisms. Herein, Megahit was executed cross-assembly on the unmapped reads from all samples to generate the assembled contigs which can be considered as originated from an unknown organism. MetaBAT2 is then utilized to perform binning on the assembled contig set. In this work, Megahit and MetaBAT2 were used with default parameters according to the previous study (Zhu et al., 2019). Before assembly, reads with lengths less than 1000bp were filtered out. In addition, low-quality MAGs were kept for the follow-up analyses. Once we finished cross-assembly and metagenomic binning, we treated each contig bin as an unknown organism, and the binned reads as a part of its genome. In terms of defining the feature of the unknown organisms, we still used the relative abundance, just as we did for known species.

After obtaining both known and unknown microbial features, the states or the existence of a certain disease are utilized as the labels for each sample. Detailed descriptions for each dataset are listed in Table 5.

Overview of MetaDR

MetaDR is composed of two separate modules, which are a deep learning module and a random forest module. The deep learning module (EPCNN) utilizes the taxonomic information and microbial features to construct an ensemble convolutional neural network for classification. The random forest module (WRF) extracts insights from different microbial features and identifies the informative features on disease pathways. The architecture of the proposed MetaDR is described in Figure 3.

Ensemble phylogenetic convolutional neural network

Construction of phylogenetic tree

A phylogenetic tree can be constructed by comparing the microbial genomes based on multiple sequence alignment, and the similar taxa are organized into clades close to each other (Fioravanti et al., 2018). Since the construction of the phylogenetic tree is not the point in our work, considering the easy accessibility, we directly utilize PhyloT (https://phylot.biobyte.de/) to generate least-pruned phylogenetic trees based on the taxonomic annotations of known and unknown microbial features. For abundance profiles of known microbial organisms, since we utilized MicroPro to map the raw reads to NCBI RefSeq database, the annotation of each taxon can be directly obtained by Centrifuge (Kim et al., 2016). For abundance profiles of unknown microbial organisms, in order to obtain the corresponding phylogenetic tree, we followed MicroPro utilizing Mash v.2.0 (Ondov et al., 2016) to calculate the pairwise distance between each MAG and the reference genomes in Centrifuge, and identified the taxonomic assignment of each MAG. However, as none of the pairwise Mash distance was below 0.05 for all the utilized datasets, we adopted the Mash distance of 0.34 suggested by (Zhu et al., 2019) to classify the microbes into the genus level. For MAGs that cannot be classified to the genus level, we directly arranged them in an initial random order. A similar approach can be leveraged to distribute all the MAGs into specific levels, and the obtained taxonomy annotations can be used to generate the phylogenetic tree of unknown microbial organisms.

Taxonomic representation

In most metagenomic predictive studies, the obtained microbial features for each sample are a one-dimensional vector that ignores the correlations among taxa (Qin et al., 2012; Zeller et al., 2014; Harris et al., 2019). Recent studies suggest that the phylogenetic tree which integrates the taxonomic relationships can be utilized to enrich the microbial feature representation (Oudah and Henschel, 2018; Reiman et al., 2020). Therefore, in this study, we consider the taxa on a phylogenetic tree as pixels in an image to extract the taxonomic spatial information from the tree-based topological structure. After obtaining the phylogenetic trees from the previous steps, we extract the taxonomic information by traversing the taxon nodes on the tree. Specifically, two popular algorithms in the data structure are leveraged to traverse the phylogenetic tree. The first is postorder traversal:

- Traverse the left subtree,
- Traverse the right subtree,

- Visit the root;

and the second one is level traversal:

- Traverse every node on each level of the tree from left to right and top to bottom.

The traversing order of the taxonomic annotations of microbial features is preserved when traversing the phylogenetic tree, which can be treated as the hierarchical information among taxa. For the postorder traversal, neighboring taxa on the phylogenetic tree are rearranged as a sequential order from left to right, and the obtained sorting order congregates the taxa with the same ancestors. In contrast, for the level traversal, different levels on the phylogenetic tree can be treated as diverse evolution stages. Theoretically, the farther the distance to the root node is, the longer time needed for evolution and the fewer similarities with the ancestor. In this case, the level traversing order implies genetic information. Herein, the proposed taxonomic representation is similar to the alphabet-sorting representation approach which assembles the microbial features with similar taxonomic names (Oudah and Henschel, 2018; Reiman et al., 2020). However, the difference is, instead of solely considering the alphabetically taxonomic annotations, we match the related taxa with the kinship on a phylogenetic tree by traverse. The instance of taxonomic representations is depicted in Figure 4.

The obtained sorted features can preserve the hierarchical information on the phylogenetic tree and enable us to adapt CNN for feature extraction. Herein, since the two proposed taxonomic representations are implemented in different perspectives, to further simulate the image structure, we conduct different matrix conversions for the obtained sequential vectors. For postorder traversal-based representation, to preserve the topological structure and congregate neighboring taxa as much as possible, we reshape the vector following a 'Z' shape path to construct the matrix. In contrast, for level traversal-based representation, the matrix is obtained by directly reshaping the vector from top to bottom and left to right, to simulate the evolution stage on the phylogenetic tree. After that, the species abundances are mapped to microbial features with specific taxonomic annotations. Herein, both matrices with different taxonomic representations can be regarded as images, while the normalized species abundances in the matrices can be regarded as the pixels. The image size is defined to be $n * n$, where n is the minimum integer to make $n * n \geq N$, and N is the dimension of the input features. Specifically, the ceiling function is applied to solve the non-integer case. If the square of the obtained image size is more than the original dimension, the missing part will be complemented by zeros. Finally, for each taxonomic representation, we obtain an abundance matrix with N microbial features. The features are then converted to a grayscale image as the input of CNN. Transformation details can be seen in Figure 5.

Convolutional neural network

After utilizing the taxonomic representation to embed the biological knowledge into microbial features, we adopt CNN to extract the taxonomic relationship and train the classifier. A standard CNN usually consists of a series of convolutional layers, pooling layers, and fully connected layers. In this study, the convolutional layer is utilized to extract deep spatial patterns in microbial features. A 1D convolution can be formulated as:

$$s(n) = (x * w)[n] = \sum_{m=0}^{N-1} x(m)w(n-m) \quad \text{(Equation 1)}$$

where $*$ is the symbol for convolutional calculation. $x(n)$ is the input sequence, $w(n)$ is the filter, and $s(n)$ is the output sequence. m represents each element when traversing the sequence $x(n)$, and N defines the length of the input sequence. Similarly, the equation can be extended to 2D convolution:

$$s(n,m) = x * w[n,m] = \sum_{a=0}^{N-1}\sum_{b=0}^{M-1} x(a,b)w(n-a,m-b) \quad \text{(Equation 2)}$$

where $x(n,m)$, $w(n,m)$, and $s(n,m)$ respectively represent the input matrix, filter, and output matrix over two different dimensions. a and b represent each element in row and column when traversing the matrix $x(n,m)$. N and M define the length and width of the input matrix. Besides, we utilize pooling layers to remove the redundancy and simplify the model complexity, and the fully connected layers are applied to accept the extracted features for final classification. Softmax is adopted as the activation function for the output layer to calculate the class probability:

$$p(y = c|x) = \frac{exp\left(w_c^T x\right)}{exp\left(w_1^T x\right) + \cdots + exp\left(w_K^T x\right)} \quad \text{(Equation 3)}$$

where $p(y = c|x)$ defines the probability of the c-th class, y is the class label and x is the input feature. w is the weight coefficient and K represents the number of total classes. The class with the largest response of w_c^T has the highest predicted probability. A binary cross-entropy is utilized as the loss function:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(p(y_i)\right) + (1 - y_i)\log\left(1 - p(y_i)\right) \quad \text{(Equation 4)}$$

where the notation $H_p(q)$ represents the cross-entropy of the distribution q relative to a distribution p. y_i is the actual label and $p(y_i)$ is the predicted label. N is the number of training samples in a given set. The utilized optimizer is Adam, and Rectified Linear Unit (ReLU) is applied as the activation function for all the previous layers. Specifically, to avoid overfitting, Dropout and L2 regularization are executed on all previous layers. Besides, we conduct early stopping to prevent the network from being over-trained. In this study, LeNet (LeCun et al., 1998) architecture is adopted for our CNN as it obtained the best overall performance among all the datasets.

Ensemble strategy

The ensemble model has been proved to relieve the overfitting risk and achieve better performance than a single model (Kang et al., 2020). In our study, an ensemble model is constructed by fusing the abundance profiles of known and unknown microbial organisms and two kinds of taxonomic representations to comprehensively assess the decisions from multiple perspectives. Specifically, EPCNN is composed of four single models: a known-profile-based (abundance profiles of known microbial organisms) model with postorder representation, a known-profile-based model with level representation, an unknown-profile-based (abundance profiles of unknown microbial organisms) model with postorder representation, and an unknown-profile-based model with level representation. Finally, the class probability of each model is output and the soft-voting strategy is used to calculate an average result. The illustration of our EPCNN can be seen in Figure 5.

Weighted random forest

Since the proposed deep learning module is a black-box model that cannot explicitly indicate the informative features for biological insights, we additionally introduce a random forest model to complement its poor interpretability. Herein, a weighted strategy is utilized based on the random forest to evaluate the abundance profiles of known and unknown microbial organisms, providing the reference weight to explain each feature. Specifically, assume that the abundance profiles of both known and unknown microbial organisms have the same weights (i.e., 0.5), for each feature set, we utilize cross-validation to test the different combinations of the hyperparameters in a random forest and choose the model with the best average performance. The best average metric (i.e., AUC) is considered as the weight for each kind of feature since it represents the classification ability of the isolated datasets to some extent.

The idea is leveraged from the Out-of-Bag (OOB) strategy (Dimitriadis et al., 2018) to compute the respective feature weights for both abundance profiles of known and unknown microbial organisms. Since the best hyperparameters are chosen in the same parameter space and the evaluation metrics are computed separately for each considered type of feature, the obtained weights are representative and relatively fair. After that, the performance metrics are normalized by dividing by their sum and serving as the weights for different feature sets. Compared to the simple concatenation, the proposed strategy can significantly reduce the influence of feature redundancy as the models are separately trained. The basic classifier is set to be a random forest since it can output the weights to measure the importance of each feature. Finally, the informative features in random forests associated with highly ranked importance coefficients are picked out for further analyses.

Description of baseline approaches

We compare both WRF and EPCNN with five state-of-the-art predictors, which are MetaML (Pasolli et al., 2016), DeepForest (Zhu et al., 2018), MetaNN (Lo and Marculescu, 2019), MicroPro (Zhu et al., 2019), and DeepMicro (Oh and Zhang, 2020). The input for all the comparison predictors includes abundance profiles of both known and unknown microbial organisms. Besides, the results are obtained by averaging

20 times running to avoid the bias caused by overfitting and random weight initialization. MetaML is a computational tool for metagenomics-based prediction tasks including automatic model and feature selection steps. The basic classifier is SVM with a radial basis function (RBF) kernel, combined with 5-fold cross-validation and grid search to choose the best parameters. DeepForest introduced a cascade deep forest (CDF) to keep the spatial structure between nodes through embedding phylogenetic tree information for metagenomic data classification. MetaNN utilized the alphabet-sorting for feature representation combining with 1D-CNN as the final classifier. The architecture for 1D-CNN includes 2 convolutional layers and 2 max-pooling layers with the kernel size of 1*3 and pooling size of 1*2 respectively. DeepMicro is a deep representation learning framework for microbiome profiles that uses deep autoencoders to transform the high-dimensional microbiome data into a robust low-dimensional representation. Herein, we select the suggested CAE and MLP classifier for comparison. For MicroPro, we fused the abundance profiles of both known and unknown microbial organisms and train a random forest for classification. The grid search and cross-validation are used to choose the best parameters, the number of max features, max depth, and estimators in the random forest are chosen in the range of [0,1.0], [1,10], and [100, 2000] respectively.

The implementations of all the methods are under the same evaluation standard and hardware equipment. The extracted abundance profiles of known and unknown microbial organisms are used as the standard starting point for comparison, with five methods conducting their feature representation and classification respectively. For SVM in MetaML, random forest in MicroPro, and ensemble classifiers in DeepForest, since they all utilized cross-validation and grid research functions in scikit-learn (https://scikit-learn.org/stable/) to choose the best parameters, in order to keep a relatively fair comparison, we set "n jobs" to be −1 to use all the cores to run the program in parallel. For DNN-based classifiers (MetaNN and MicroPro), we used the GPU acceleration computation for training. The average running time of 20 times experiments is recorded for comparison.

Regarding the baseline methods of WRF, we implement different feature selection methods with respect to considering the features of both known and unknown microbial organisms (Raw and MicroPro). Raw is based on the random forest of the original version (Breiman, 2001). It simply concatenates the known and unknown as a one-dimensional vector without any normalization as the input. MicroPro is proposed by Zhu et al. (2019), which utilizes Boruta feature selection method to select the important features (Kursa and Rudnicki, 2010). MicroPro calculates the mapping rate for each sample and multiplies the known and unknown abundances to satisfy the sum of them to be one constraint as the input. For both Raw and MicroPro, the model is trained under the 5-fold cross-validation with grid search to choose the best parameters. For our WRF, the corresponding best mean cross-validated scores of the best estimators chosen by grid search are normalized and served as the weights for abundance profiles of both known and unknown microbial organisms. After that, the combination of the weighted features is used to train the random forest as well as sort out the important features. Notice that for feature selection, we trained WRF based on all the data which makes the selected features more representative. In addition, the proposed WRF can be utilized as a classifier. When being applied for the prediction tasks, we train a WRF on the training data to obtain the weights for abundance profiles of both known and unknown microbial organisms. After that, the same weights are applied to the test data to evaluate the performance of the new classifier.

Ablation settings

Regarding the ablation experiments of the proposed taxonomic information, the comparisons of the prediction performance with/without taxonomic tree spatial information are based on the same parameter settings and DNN architectures. For models without taxonomic tree spatial information, we simply removed the feature extraction layers of CNN.

For DNN-based comparisons, the input of MLP is the raw microbial features. For 1D-CNN model, the input is the raw microbial features with phylogenetic representation. For 2D-CNN model, the input is the microbial features with phylogenetic representation and matrix conversion. For all the layers in DNN models, Dropout is utilized with the ratio of 0.2. We also implemented L2 regularization and the ratio is set to be 0.0001. Besides, we conduct early stopping to avoid the network being over-trained. The patience is set to be 10 and the tolerance is set to be 0.00001. The number of hidden units is selected among 512, 256, 128, and the number of layers is selected among 1, 2, 3. The batch size is set to be 1 for all the datasets.

Development environment

In this study, Keras is adopted as the deep learning library support for its accessibility and flexibility. The experimental environment is Intel Core i79700K CPU, GTX 2080Ti graphics card, 64G memory, and 1T hard disk. The operating system is Ubuntu 18.04, and Anaconda platform is used for Python development. Besides, in order to speed up the training process of our deep learning models, we utilized the GPU acceleration technology CUDA developed by Nvidia.

QUANTIFICATION AND STATISTICAL ANALYSIS

Evaluation metrics

In order to facilitate the comparisons with other state-of-the-art predictors, we follow the same evaluation standard as MicroPro (Zhu et al., 2019) to examine the effectiveness of our models. Herein, we randomly separate the dataset into the training set and the test set with a ratio of 7:3. For all the datasets utilized in this study, the classification task is to predict the states or existence of a certain disease (T2D, LC, or CRC) given the information of the microbial composition for a specific subject (i.e., a binary classification problem is defined for each considered experiment where true denotes disease and false denotes control). AUC is selected as the evaluation metric in our work according to the previous studies (Pasolli et al., 2016; Zhu et al., 2019; Lo and Marculescu, 2019; Nguyen and Zucker, 2019). AUC can show both sensitivity and specificity for prediction. Generally, the higher the AUC score, the better a classifier performs for the given task. A 0.5 AUC means a random guess while a 1 AUC indicates perfect classification. To prevent the optimistic bias caused by the random split and weight initialization in our models, we repeatedly run the experiments 20 times and the final results are obtained by an average of 20 times running. The above procedure was executed for all the experiments including the comparison methods used in this study. Besides, the boxplots are provided for robust and fair comparisons.