RESEARCH

# Kernel principal components based cascade forest towards disease identification with human microbiota

Jiayu Zhou[1†], Xuwen Wang[2], Yanqing Ye[1,3] and Jiang Jiang[1,2*]

[*]Correspondence: jiangjiangnudt@nudt.edu.cn
[1]College of Systems Engineering, National University of Defense Technology, Changsha, China
[2]Channing Division of Network Medicine, Harvard Medical School, Boston, American
Full list of author information is available at the end of the article
[†]Equal contributor

## Abstract

**Background:** Numerous pieces of clinical evidence have shown that many phenotypic traits of human disease are related to their gut microbiome, i.e., inflammation, obesity, HIV and diabetes. Through supervised classification, it is feasible to determine the human disease states by revealing the intestinal microbiota compositional information. However, the abundance matrix of microbiome data is so sparse, an interpretable deep model is crucial to further represent and mine the data for expansion, such as the deep forest. What's more, overfitting can still exist in the original deep forest model when dealing with such "large p, small n" biology data. Feature reduction is considered to improve the ensemble forest model especially towards the disease identification in the human microbiota.

**Results:** In this work, we propose the kernel principal components based cascade forest method, so-called KPCCF, to classify the disease states of patients by using taxonomic profiles of the microbiome at the family level. In detail, the kernel principal components analysis method is first used to reduce the original dimension of human microbiota datasets. Besides, the processed data is fed into the cascade forest to preliminarily discriminate the disease state of the samples. Thus, the proposed KPCCF algorithm can represent the small-scale and high-dimension human microbiota datasets with the sparse feature matrix. Systematic comparison experiments demonstrate that our method consistently outperforms the state-of-the-art methods with the comparative study on 4 datasets.

**Conclusion:** Despite sharing some common characteristics, a one-size-fits-all solution does not exist in any space. The traditional depth model has limitations in the biological application of the unbalanced scale between small samples and high dimensions. KPCCF distinguishes from the standard deep forest model for its excellent performance in the microbiota field. Additionally, compared to other dimensionality reduction methods, kernel principal components analysis method is more suitable for microbiota datasets.

**Keywords:** Human microbiota; Supervised classification; Kernel principal components; Cascade forest; Disease identification

## Background

The human microbiota is made up of about 100 trillion microbial cells. Compared to 10 trillion humanoid cells in our body, microbiota provides many missing features of human biology [1]. The content and number of gut microbes keep a dynamic balance during their hosts' evolution, and microbes also assist their host to maintain normal physiological functions [2, 3]. There are numerous clinical studies exploring the association between microbiome and phenotype which is used to identify differentially abundant taxa between health and disease [4], including inflammation [5], obesity [6, 7, 8, 9], autism [10, 11], immune system diseases [12], neurological diseases [13] and cancer [14, 15, 16]. Recent advances in sequencing technologies have made it feasible to profile the microbiome via metagenomic sequencing, which is a technique to extract DNA from environmental samples [17]. Human microbiota genomics cooperative research programs have been launched internationally in recent years, such as the European Metagenomics of the Human Intestinal Tract [18] and the Human Microbiome Project [19]. These programs aim to understand the gut microbiota of healthy individuals through large-scale sequencing and use this as a reference to study the intestinal tract under disease conditions.

Biology classifies and names various taxa of organisms according to different levels, normally including Domain (d), Kingdom (k), Phylum (p), Class (c), Order (o), Family (f), Genus (g), and Species (s). At present, the classification of diseases by intestinal microbes is mainly based on the genus level [20]. However, there are many microbiomes cannot be completely and accurately classified to the genus level. Besides, due to microorganisms themselves are very rich in the genus level, the established "sample-feature" matrix tends to be so sparse leading to unnecessary biological detection and calculation. If we can get good identification results from a higher level in meta-genome data, it will be more beneficial to be applied in the real application. By comparing prediction accuracy in the genus level and the family level, we found that datasets generated by family level perform more stable than that of the genus level. As a result, our work applies the microbiome data at the family level as the diagnosis basis.

Using machine learning algorithms to identify highly complex and unknown patterns in datasets (such as human microbiota) is of great value [1]. It has been demonstrated that several existing supervised classifiers, such as Random Forests (RFs) and Support Vector Machine (SVMs) [21], can be effectively used to classify and predict the disease based on microbiota population. However, because of inconsistent individual studies and lack of standardized data analysis methods, the accuracy of classifying and predicting the diseases through the human intestinal microbiome is still unsatisfied. Enhancing the complexity of an algorithm by deepening the network, increases not only the number of computing function but also the degree of its embedding. [22] published an article, and the concept of "Deep Learning" (DL) was officially proposed. DL is a high-level abstraction algorithm that uses multiple complex structures to represent multiple nonlinear changes [23]. Deep Neural Networks (DNNs) have been widely exploited recently for meta-genomic association studies [24, 25], meta-genomic classification [26, 27], and disease diagnose [28, 29]. Large training data is necessary for DNNs to realize good performance, which may not be possible in small-scale datasets like biology and medical science. For example, almost all CNN faces over-fitting problems due to the limitation of data volume and the increase of training parameters. That is, the magnitude of the training set does not match the complexity of the model, and the weight learning iterations are overtraining, fitting the noise in the training data and the non-representative features in the training examples. Recently, a Deep Forest (DF) model called gcForest is proposed by Zhou and Feng, which is an ensemble of ensembles decision tree method and performs excellently in many experiments [30, 31]. The interpretable tree structure can solve the problem of non-differentiable. Additionally, compared to the time-consuming parameter adjustment, gcForest is far more efficient with fewer hyper-parameters.

In the gcForest model, it will go through multi-grained scanning first to get its corresponding transformed feature representation. Sliding windows are used to scan the low dimension features, and differently grained feature vectors will be generated by using multiple sizes of sliding windows. The instances extracted from the same size of windows will be used to train a completely-random tree forest and random forest. And then the class vectors are generated and concatenated as transformed features. The model also uses two different types of forests to represent the data. Random forest is an integrated model of random trees, introducing randomness to encourage diversity. While for completely random trees, they select and assign features completely randomly. The transformed training set will then be used to train the first grade of a cascade forest, which is the ensemble of ensembles method.

However, the scanning model in gcForest can only consider the original sequence, which will lead to features disturbing for the unknown relationship between two adjacent features. The microbiota datasets are too sparse and contain lots of 0 value in many flora features. When the training sets are put into a multi-grained scanning package, due to the not yet clear complicated relationships between each microbiota, it can extract representative new features sometimes but others not. Thus, the standard DF model still faces overfitting and ensemble diversity challenges when dealing with such “large p, small n” biology data. Many researchers have been exploring how to improve the DF algorithm of identification for special field [32, 33, 27]. Features are the key to determining similarity measure and classification prediction. To highlight some useful information and suppress the useless, it is necessary to reduce the input features. The original datasets can be transformed at the beginning of the algorithm to adapt to subsequent depth learning [34]. The affinity network model was put forward to learn from a limited number of training examples and generalizes well [35]. The kernel-based model can also offset the hyperplane by modifying the kernel function caused by the unbalanced data. [36] applied the kernel method to feature extraction and proposed kernel principal components analysis (kPCA) method. The experimental results show that kPCA can not only extract nonlinear features but also have better recognition results. To find nonlinear high-order correlations between the data and remove this correlation, it first upgrades the data and then reduces the dimension. KPCA is widely used in various fields such as industrial nonlinear process monitoring [37, 38] and image classification [39]. We systematically explored the disease identification by utilizing the kernel principal components analysis considering limited and unbalanced samples and a large number of features. To further improve the meta-genomic classification accuracy, we use the mixed data fused with associated metadata, such as gender, age, and other basic information as the diagnosis basis and fed them to the proposed model.

## Methods

The disease identification can be treated as a multi-class classification problem, and all the datasets we use here contain three categories. This section presents the datasets information and detailed procedures of the KPCCF method for disease identification. The four microbiota datasets used in our paper are introduced first. In the following subsection, the kernel principal components analysis method is applied to reduce the original dimension of the microbiota datasets. Secondly, we use cascade forests to preliminarily discriminate against the disease state of the sample with the reduced human gut microbiota. Finally, the overall procedure of KPCCF are detailed present.

### Microbiota datasets

In this work, four datasets derived from the standardized database of human intestinal microbiome study [40] are used to verify the effects of the gut microbiome on the occurrence of different diseases in humans. The datasets are related to four popular diseases, Clostridium Difficile Infection (CDI), Colorectal Cancer (CRC), Inflammatory Bowel Diseases (IBD), and Obesity (OB). CDI is the main cause of antibiotic-associated diarrhea. With the increase of its incidence rate, CDI has already become one of the important public health problems that threaten human-beings' health. CRC, the world's second-largest cancer, is malignant cancer caused by the accumulation of genetic mutations, which causes a massive proliferation and spread of more than 50%. IBD is caused by abnormal responses of the immune system of the genetically susceptible host to environmental factors, including Crohn's disease (CD) and ulcerative colitis (UC). Different disease states occur under the combined action of environmental factors and intestinal microbes. OB measures are the incidence of overweight/obesity (OW/OB). Table 1 shows the detailed divisions of the used datasets. Specifically, in the cdi_schubert dataset [41], the samples consist of 93 $CDI$, 89 $nonCDI$ and 154 $H$ samples, in which nonCDI represents patients with diarrhea who tested negative for CDI, CDI represents patients that suffer from CDI, and H represents the healthy samples. The crc_baxter dataset [15] consists of 120 $CRC$, 198 *adenoma* and 172 controls, in which CRC represents tumor disease infection, adenoma signifies adenoma infection, and H denotes the healthy samples. In ibd_papa[42], there are 24 $nonIBD$, 43 $UC$, and 23 $CD$, in which non-IBD controls are patients with gastrointestinal symptoms but no intestinal inflammation. While ob_goodrich dataset [43] possess 185 $OB$ (obesity), 336 $OW$ (overweight) and 428 controls.

The datasets all come from real-world cases. Each dataset contains a metadata table, an OTU (Operational Taxonomic Units) table, and other related information. The metadata table involves various physical characteristics such as gender, age, and disease state of the patient. OTU is an operation classification unit that artificially groups sequences according to a certain degree of similarity where different classification units are formed by clustering operations. Since the microbiota community has no explicit relationship so far, there are many types of research carried out using RNA sequencing [33], DNA sequencing [34], and clinical images [39]. Some experiments choose OTU as an additional supplement nowadays, however, only a few related types of research using microbiota OTU data for research, and the results obtained were not ideal. Our paper only used the OTU table for prediction to mine the relationship between the patients and their microbiota. Thus, our results only generated by OUT are more competitive.

To mine the microbiome data, the datasets need to be processed and converted into a sample-feature matrix first. The procedure of data processing is shown in Figure 1. *Step One*, we split the first column of the original OTU table by a semicolon, and connect the split series expanding the columns of the original OTU table. *Step Two*, according to the columns of the genus and the family level, the flora is hierarchically clustered, and the number of communities of different numbered samples is accumulated together. *Step Three*, the table obtained in the previous step is transposed to a sample-microbiota features table, and in *Step Four*, placing the disease state in the metadata set as the final column. The processed sample dataset is represented as a sample/feature dimension, and the last column is the annotation of the disease-state.

### Kernel principal components based feature reduction

The number of training samples needs to grow exponentially with the feature dimension [44]. That is, if $N$ training samples are enough to cover the one-dimensional feature space, then $N^2$ samples are needed to cover the two-dimensional feature space of the same density, $N^3$ samples are needed to cover the three-dimensional feature space, and so on. From the very beginning, as the feature dimension increases, the performance of the classifier will gradually increase. However, after the number of features reaching a certain point, the prediction accuracy gradually decreases. Both redundant features (which can be derived from other features), and irrelevant features (which do not affect model training) are catastrophic for machine learning algorithms. Dimensional disaster always leads to weak generalization, so it is necessary to first reduce the dimension to avoid overfitting. Removing unrelated features can not only reduce the difficulty and speed of learning tasks but also enhance the understanding between features and eigenvalues.

Since it is unclear about the biological mechanism of action and the relationship between every microbiota population, directly eliminating the ”useless” features may result in information omission. Therefore, we use the feature transformation method to reduce the dimension of data. During the features mapping from one-dimensional space to another, only the eigenvalues will change accordingly. Kernel Principal Components Analysis (kPCA) is a nonlinear extension of the Principal Components Analysis (PCA) algorithm. KPCA is considered being used to reduce the intestinal microbiota characteristics to slightly improve the data dimension. The process of kPCA is to raise the original dimension data to new $k$-dimensional, and the final goal is to make the data linear separable in the target dimension, which is the maximum separability of PCA. The kernel-based model can also offset the hyperplane by modifying the kernel function caused by the unbalanced data. By replacing the original data with a kernel function, it is possible to mine the nonlinear information contained in the datasets. It describes the correlation between multiple features and captures important information to achieve better results. What’s more, dimension reduction can also remove some noise and unnecessary details, and effectively speed up the training process.

We choose kPCA depending on the following considerations: (i) the calculation of the kernel function is independent of the feature dimension. The introduction of kernel function avoids the direct operation of high-dimensional feature space after transformation, greatly reducing the calculation amount and avoiding the ”dimensionality disaster”. Some kernel functions, such as RBF kernel, make the dimension of feature space infinite to improve the pattern classification or regression ability and (ii) there is no need to know the form and parameters of the nonlinear transformation function. The calculation of kernel function in the original input space essentially implicitly corresponds to high-dimensional nonlinear transformation function. The transformation overcomes the limitation of the nonlinear feature space dimension.

There are no obvious performance metrics to help choose the best kernel method and hyper-parameter values for kPCA, which is an unsupervised learning algorithm. We use the grid search method to select the kernel function and gamma values that will allow the task to perform optimally and get the best classification accuracy. There are many kinds of kernel functions, such as linear kernel functions, polynomial kernel functions, sigmoid kernel function, and Gaussian kernel functions, etc. Gaussian kernel functions, also called Radial Basis Function (RBF), are the most commonly used. By grid search method, we choose RBF as the kernel function of our model and set kPCA gamma 0.05, which obtained the highest predict accuracy. The RBF kernel is presented as:

$$K(x, x') = exp(-\frac{\|x - x'\|_2^2}{2\sigma^2}) \tag{1}$$

where $\|x - x'\|_2^2$ is the squared Euclidean distance between two feature vectors, $\sigma$ is a free parameter. It can map the input data into infinite dimensions. An equivalent but the simpler definition is to set a new parameter $\gamma = \frac{1}{2\alpha^2}$, then the expression can be expressed as:

$$K(x, x') = exp(-\gamma\|x - x'\|_2^2) \tag{2}$$

The value of the RBF kernel ranges from 0 to 1, which is a similar metric representation and decreases as the distance increases. The feature space of the kernel has an infinite number of dimensions. For $\sigma = 1$, its expansion is:

$$exp(-\frac{1}{2}\|x - x'\|_2^2) = \sum_{j=0}^{\infty} \frac{(x^T x')^j}{j!} exp(-\frac{1}{2}\|x\|_2^2) exp(-\frac{1}{2}\|x'\|_2^2) \tag{3}$$

### Ensemble classification model of cascade forest

The main goal of the paper is to explore the relationship between microbes and disease occurrence based on community and quantity of intestinal microbiota. However, the abundance matrix data of the microbiome is too sparse with the small sample size even after appropriate dimensionality reduction. That is, most microbes are limited to a relatively small number of samples. A deep model is needed to represent and mine the data. The integrated cascade forest model is the ensemble of both breadth and depth of the traditional forest model.

Cascade forest is an ensemble of ensembles method, which is composed of random forests and completely random forests in its structure. Completely random forest randomly selects a feature when splitting. Each random forest will output features with an important factor, then we rank the features after the average important factor for all forests and combine features of all levels according to each forest feature's importance. In each level, the entire model is validated on the training set. Compared to most deep neural networks with fixed model complexity, the cascade forest adaptively determines its model complexity by terminating training when it is sufficient. This makes it suitable for training data at different scales. Finally, averaging across all trees in the same forest, and the class distribution for each forest is generated.

### The overall procedure of Kernel Principal Components based Cascade Forest (KPCCF)

KPCCF model is composed of two modules: firstly, using kPCA to reduce the high dimension of the input "large p, small n" data; secondly, using the cascade forest depth model to improve the model's classification ability.

The overall procedure of the KPCCF algorithm is shown in Figure 2.

*Step One* We apply feature reduction to adjust the datasets better suitable for microbiota datasets. By RBF kernel function, the unknown correlation high-dimensional data will be transformed into approximately linearly separable data. Though not all categories are distinguished, it will still catch some similarity factors. The transformed feature vectors will then be used to train the following cascade forests respectively.

*Step Two* The features extracted from the previous stage by kPCA method are fed to the cascade forest. Each layer of cascade forest is composed of multiple forests and will produce a class vector as its output. The class vector will connect with the output produced by one branch of the former stage to be the inputs of the next layer. Then the next layer will produce another class vector, which will further connect with an output produced by another branch of kPCA. This process continues until reaching the termination condition, such as achieving the expected accuracy or reaching the maximum number of layers. After getting the final class vector, we will calculate the average value for all kinds of possibilities and select the class with the maximum aggregated to be the final classification result.

KPCCF is a novel decision tree aggregation method, and its prediction accuracy is highly competitive with deep neural networks in a wide range of tasks. Besides, the deep forest is easier to train because it has fewer hyper-parameters. Another advantage is that the model complexity of the deep forest can be automatically determined for different training datasets, making the deep forest work well even on small datasets. Moreover, the advanced feature reduction makes the cascade forest algorithm much more suited for disease prediction.

## Results

In order to verify the proposed method, in this section, we tested the performances of various classifiers derived respectively from KPCCF and other state-of-the-art methods, including Decision Tree (DT) [45], standard ensemble method RF [21], the normal deep learning algorithm CNN [28], and the original deep forest model gcForest(DF) [33, 27, 34] on the four datasets shown in Table 1, and evaluated the results through classification accuracy.

### Experiment design and parameter settings

As the downloaded four microbiota datasets are composed of the most basic hierarchical species microbiome, we preprocess the datasets according to procedures in Figure 1 and form the family level microbiota datasets. The newly-built ones are composed of a set of input features (the number of microbiota in the unit sample) and disease tags.

Cross-validation (CV) is a common statistical analysis method used to verify the performance of classifiers. In this work, we apply 4-fold CV to carry out the experiment, and for each dataset, the CV process has been conducted 20 times, and the average performance is evaluated as the final result. The samples in each dataset are randomly divided into 4 parts evenly. Each part of the samples is respectively used as a testing dataset and the remaining parts of samples make up the training dataset. During each fold, the training dataset is fed into different classifiers to train the model, then the testing dataset is used to test the trained classifiers. The results are evaluated through the prediction error and their square sum.

DT performs as the tree structure. It starts from the root node, then tests the corresponding feature attributes in each item to be classified, and selects the output branch according to its value until the leaf node is reached. The category stored by the leaf node is used as its result. The decision process is shown in Figure 3. During the process of CDI disease prediction, every feature of different microbiota leads to a specific decision.

One of the improved bagging DT algorithms, RF, is a classifier using multiple decision trees to train and predict samples. Select the category with the most votes in the classifier's voting results as the final classification result. For the random forest, the number of trees, max-depth is tested with the grid searching method. We set them 100 and 2 separately. All other parameters are left as default, such as max_features (default is auto), min_samples_split (default is 2) and min_samples_leaf (default is 1). We use the value of Gini impure to calculate properties and select the most appropriate node.

In the model of CNN, the original input 2-dimension sample-feature data vector needs to be expanded to the 3-dimension, that is, turned from 2*2 to 1*2*2 to make the dimension conform to the model's input. We decide how many hidden layers are best in the disease classification based on experimental tests, errors, and accuracy. 6 hidden layers are used in this study (including 3 convolution layers, 2 pooling layers, and 1 fully-connected layer). The multi-class classification of this experiment uses categorical cross-entropy as the loss function. By using the method of Stochastic Gradient Descent (SGD), recursively approximating the minimum deviation model, and using the chain derivation rule to deduct the nodes of the hidden layer, the ultimate goal is to make the loss of all training data as small as possible. The disease classification result is obtained at the output layer after the transformation of the hidden layer. The loss and the accuracy of CNN in different epoch are shown in Figure 4. We can see that the loss function is decreasing as epochs grow, and the accuracy outperforms consistently as dimension increasing. The loss value on the test set starts to rise again after 200 epochs. Because the data is too small, the accuracy changes slowly in the early stages. Through the curve of accuracy, it can be found that the fit has also begun to appear after 200 epochs.

For KPCCF training, suppose that the original 1-D microbiome input is of 100 raw features. In the feature reduction module, taking the cdi_schubert dataset at the family level as an example, each sample has 90 features. According to the number of the dataset features, we have varied the parameter number of components from 5 to 90 with the step size of 5, the accuracy is shown in Figure 5. We can abstract 30 principal components in the family level as it reaches its peak. However, in the process of dimension reduction, the features number in our four datasets varies from 49 to 93 in the family level and from 142 to 255 in the genus level. We cannot find the number of features that can optimize the final accuracy and the computing efficiency at the same time. As a result, we set the hyper-parameter "*n_components*" as mle at last, which means the number of features will be automatically selected to meet the required percentage of variance. That is, the model will select a certain number of principal components features to reduce dimensionality according to the variance distribution of the feature, which we find can balance the final accuracy and the compute efficiency.

After feature reduction by kPCA, the transformed training set will then be used to train the 1st-grade of a cascade forest. These data will be used to train two random forests and two completely-random tree forests. Each forest contains 30 trees generated by randomly selecting a feature for a split at each node of the tree and growing tree until each leaf node contains only the same class of instances. If there are three classes to be predicted, then each of the four forests will produce a three-dimensional class vector; thus, the next level of the cascade will receive augmented features. Compared to most deep neural networks with fixed model complexity, the cascade forest adaptively determines its model complexity by terminating training when it is sufficient. As a result, the KPCCF model has a few parameters to adjust.

### The performance comparison of various classifiers

In this paper, every dataset has been tested 20 times in all 6 methods, DT, RF, CNN, CF, DF, and KPCCF, with the data being divided differently. And we take the average as their final results. Taking the cdi_schubert dataset as an example, the confusion matrix of one experiment by six algorithms is shown in Figure 6. As we can see, various algorithms identify diseases with different sensitivity. DT and RF identify samples with CDI disease well, while CNN, DF, CF, and KPCCF algorithms can identify healthy samples well. Above all, KPCCF has the best results for its diagonal color is the lightest. In specific, KPCCF classifies 22 samples as nonCDI, while 8 of these are supposed to be CDI in reality. It predicts 17 to be CDI with 11 to be true. 45 of the training samples are diagnosed as healthy, and only one of them are wrong.

When the dataset is unbalanced, using accuracy measures to evaluate the classification performance is not enough, some other metrics, like "*precision*" and "*recall*", or a combination of the two. In the multi-category problem, the F1 score is divided into two types, which are $Macra\ F1\ score$ and $Micro\ F1\ score$ respectively. The n-class classification problem is divided into n two-category evaluations, and the F1 score of each two classifications is calculated. The average of n $F1\ scores$ is $Macra\ F1\ score$. $Macra\ F1\ score$ is heavily influenced by the small number of samples. Divided the n classification into n two-category evaluations, and the TP, FP, RN of the n two classification are added together. Then the evaluation accuracy and the recall rate are calculated. The final calculated $F1\ score$ is $Micro\ F1\ score$. In the case of uneven data samples, the use of $Micro\ F1\ score$ is more reasonable.

We use accuracy ($Acc$), variance ($Var$) and Micro F1 score as model evaluations. As can be seen in Table 2, the $Acc$ and the $Micro\ F1\ score$ of the KPCCF algorithm is generally better than the other five existing algorithms. In all datasets, CNN always got the lowest accuracy except in the ob_goodrich dataset. It's probably because the ob_goodrich dataset has a relatively larger dataset, while CNN is easier to over-fit, especially when the datasets are extremely small like what used in this article. When the sample number of the dataset is extremely small, such as the ibd_papa dataset, KPCCF showed an overwhelming advantage whose accuracy reached up to 0.57. It's a 3-class classification problem with only less than 100 samples. And it has much more hyper-parameters to adjust. The predictions of CF and DF models are not very stable, while kPCA can improve the situation. It is also noticeable that DF model performs well in the ibd_papa and ob_goodrich datasets but poorly in other two datasets. This is because their features are relatively smaller compared to their samples' size. Thus, by use of feature reduction method reasonably, cascade forest, which is a deep forest model, may produce sensible results on the datasets.

To more intuitively display the results in the table, we visualize some of the results. In the multi-class classification problem, the Micro F1 score is more accurate to measure the algorithms. The Micro F1 score of 4 prediction results in family level respectively is shown in Figure 7. In the thermal map, the darker the color, the larger the value. It can be easily found that CDI disease is the most adaptive to classification methods for its color always much darker than other datasets. While CRC disease gets prediction far from satisfied. That's may because the disease cannot be easily classified by microbiota.

The biggest advantage of KPCCF is that: 1) it has excellent performance even with a small amount of data as it's the ensemble of RF, 2) it has fewer hyper-parameters compared to DNN, and 3) compared to the multi-scanning stage of DF, it discovers nonlinear high-order correlations between data and remove this correlation without knowing the relationships between microbiota community in advance.

## Discussion

### Extended study based on metadata

To explore the relationship between metadata and disease, here, we use four datasets by fusing their microbiota data and the metadata as the mixed datasets covering CDI, CRC, IBD and OB diseases, and train the KPCCF diagnosis model again. Since the information in each metadata is different, we add age and gender feature in the cdi_schubert and ibd_papa dataset, add BMI and age and gender in the crc_baxter dataset, add age in ob_goodrich dataset. Based on the mixed datasets, the various models are tested via similar settings as before. The performance of microbiota data only and mixed data fused with metadata is shown in Table 3. Results show that the concatenation improves the accuracy score in CRC, IBD, and OB. This means these three diseases may have a great relationship with samples' gender, age, and other characteristics and simply concatenating them brings better result. The prediction accuracy of CRC and OB increased by 0.05 reaching 0.48 and 0.52 respectively, which is great progress. Because the results improve a lot after adding gender and other information. In specific, we find that older people are more likely to get sick in these three datasets. While there is no obvious relationship between CDI and age and gender for the predicted accuracy even decrease. However, the accuracy of CDI prediction still ranks highest.

### Comparative study between the genus level and the family level microbiota

Using the family level will be more beneficial to the application as we analyzed before. To verify this, we compare the prediction accuracy of the genus level and the family level in all datasets. Similarly, the KPCCF model has been tested 20 times via the genus level dataset and family level dataset.

According to the comparison of KPCCF prediction results in the genus level and the family level respectively shown in Figure 8, it is found that in all of the four datasets, the family level performs more stable than genus level, although the prediction of genus level is sometimes a little more accurate in classification. What's more, the family level has fewer features, thus time-saving and further avoiding overfitting.

### Comparative study among various dimension reduction methods

To validate the usefulness of our used kPCA method, by substituting it with Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Linear Discriminant Analysis (LDA), Least Absolute Shrinkage and Selection Operator (LASSO) dimension reduction methods, respectively, we conduct CV experiments with cascade forest on 4 datasets for comparison. The dimension reduction process is shown in figure 9.

PCA dimension reduction requires the largest $d$ eigenvectors of the sample covariance matrix $X^TX$ and then using the matrix of the largest $d$ eigenvectors to make low dimensional projection dimensionality reduction. SVD can also obtain the matrix of the largest $d$ eigenvectors of the covariance matrix $X^TX$, but SVD has another advantage. SVD is especially effective when the sample size is large. In fact, PCA only uses the right singular matrix of SVD, but the left singular matrix can also be used for row number compression. In contrast, the right singular matrix can be used for the compression of the number of columns, that is, the feature dimension.

The principles of LDA and PCA are different. PCA is an unsupervised algorithm projected to the direction by the sort of data variance. The assumption is that the larger the variance, the more information there is. While for LDA, it is projected after the selection of the smallest intra-class variance and the largest variance between classes. Considering specific purposes and scenarios, in classification problems, the feature reduction criteria for LDA are more reasonable.

LASSO raised for the problem that the ridge regression cannot be parameterized, and it can select parameters by parameter reduction to achieve dimension reduction. The penalty term is a norm, and some parameters can be forced to 0 to achieve the purpose of parameter selection.

We use the $Micro\ F1\ score$ to evaluate different feature reduction models. The $Micro\ F1\ score$ of 5 algorithms prediction results in each dataset is shown in Figure 10. As we can see, the kPCA model in red color has the most prominent performance among these 5 methods with $Micro\ F1\ score$ 0.71, 0.48, 0.57 and 0.48 respectively. LASSO's performance ranks in second place with $Micro\ F1\ score$ 0.7, 0.41, 0.56 and 0.46 respectively, which are very close to the best performance on most datasets. While the $Micro\ F1\ score$ in PCA, SVD, and LDA fluctuating in different data sets.

## Conclusion

Considering the genus level vast microbiota species and the difficulty of sequencing, it is more advantageous to make a predictive analysis at the family level. In this work, we propose a KPCCF model to solve the problem of disease identification based on the family-level microbiome. To prove the superiority of the proposed model, we conduct the multi-class classification experiment on four different real microbiota datasets and compare its performance with other state-of-the-art algorithms, including DT, RF, CNN, CF, and DF algorithms. The results confirm that our improved cascade forest model KPCCF performs comparatively better, while cascade forest can adapt larger datasets and get better results. Furthermore, we carry out the extended study by combining the microbiota data with the corresponding metadata and find the insertion of the metadata can effectively improve the accuracy of disease identification. In the end, we explore different mainstream feature reduction algorithm and find kPCA is the best selection for our microbiota datasets.

The contributions of our work are summarized as: (1) we introduce the kPCA method into the cascade forest algorithm, which can both effectively reduce the feature dimension and improve the classification accuracy; (2) instead of traditional two-class disease diagnosis problem, we explore a multi-class classification model to solve the disease identification problem with more than three disease states; and (3) in practical application, we only utilize numbers of microbiota in the family level for supervised learning and find ways to improve disease identification accuracy, which is a great challenge. However, due to the difference between individuals, when there are small number of samples, the trained model may lack generalization ability. In our future works, we will focus on improving the generalization ability of our KPCCF model. One feasible way is using transfer learning to construct more samples from the samples with different diseases or health states.

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Availability of data and materials**
Not applicable. The datasets used in this paper are from the standardized database of the human intestinal microbiome study: MicrobiomeHD, https://zenodo.org/record/1146764#.XDv1O_xS9sN.

### Competing interests

The authors declare that they have no competing interests.

### Funding

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences [XDB13040400], the China 863 program [SS2015 AA020109-4], Shenzhen Science and Technology Grants [JCYJ2013040111411 1457 and JCYJ20130401170306884], Shenzhen Peacock Plan [KQCX201306281 12914301 and KQCX20130628112914291], and Key Laboratory of HumanMachine-Intelligence Synergic Systems, Chinese Academy of Sciences. This was also partly supported by the MOE Humanities Social Sciences Fund [13YJC790 105 to Y.L.] and Doctoral Research Fund of HBUT [BSQD13050 to Y.L.].

### Author's contributions

JJ, JZ, XW and YY conceived the project, conducted the major analysis part and drafted the manuscript. JZ wrote the experimental code and conducted the experiments. YY was involved in the model optimization. All authors read and approved the final manuscript.

### Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 71671186 and 71571185. Many thanks to the reviewers for proposing sound advice that is really helpful in improving our paper.

### Author details

[1]College of Systems Engineering, National University of Defense Technology, Changsha, China. [2]Channing Division of Network Medicine, Harvard Medical School, Boston, American. [3]Center for Polymer Studies, Boston University, Boston, American.

### References

1. Dan, K., Costello, E.K., Rob, K.: Supervised classification of human microbiota. Fems Microbiology Reviews **35**(2), 343–359 (2011)
2. Qin J, R.J.e.a. Li R: A human gut microbial gene catalogue established by metagenomic sequencing. Nature **464**(7285), 59–65 (2010)
3. Ilseung, C., Blaser, M.J.: The human microbiome: at the interface of health and disease. Nature Reviews Genetics **13**(4), 260–70 (2012)
4. Koh, H., Blaser, M.J., Li, H.: A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. Microbiome **5**(1), 45 (2017)
5. Morgan, X.C., Tickle, T.L., Sokol, H., Gevers, D., Devaney, K.L., Ward, D.V., Reyes, J.A., Shah, S.A., Leleiko, N., Snapper, S.B.: Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biology **13**(9), 79 (2012)
6. Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Vincent, M., Mardis, E.R., Gordon, J.I.: An obesity-associated gut microbiome with increased capacity for energy harvest. Nature **444**(7122), 1027–1031 (2006)
7. Walters, W.A., Xu, Z., Knight, R.: Meta-analyses of human gut microbes associated with obesity and ibd. Febs Letters **588**(22), 4223–4233 (2014)
8. Sze, M.A., Schloss, P.D.: Looking for a signal in the noise: Revisiting obesity and the microbiome. Mbio **7**(4), 01018–16 (2016)
9. Finucane, M.M., Sharpton, T.J., Laurent, T.J., Pollard, K.S.: A taxonomic signature of obesity in the microbiome? getting to the guts of the matter. Plos One **9**(1), 84689 (2014)
10. Dae-Wook K, E.I.Z.e.a. Gyoon P J: Reduced incidence of prevotella and other fermenters in intestinal microflora of autistic children. Plos One **8**(7), 68322 (2013)
11. Son, J.S., Ling, Z. J, Rowehl, L.M., Xinyu, T., Yuanhao, Z., Wei, Z., Leighann, L.K., Gadow, K.D., Grace, G., Robertson, C.E.: Comparison of fecal microbiota in children with autism spectrum disorders and neurotypical siblings in the simons simplex collection. Plos One **10**(10), 0137725 (2015)
12. Hooper, L.V., Dan, L. R, Macpherson, A.J.: Interactions between the microbiota and the immune system. Science **336**(6086), 1268–73 (2012)
13. Hsiao, E.Y., Mcbride, S.W., Sophia, H., Gil, S., Hyde, E.R., Tyler, M.C., Codelli, J.A., Janet, C., Reisman, S.E., Petrosino, J.F.: Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell **155**(7), 1451–1463 (2013)
14. Wang T, Q.Y.e.a. Cai G: Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. ISME JOURNAL **6**(2), 320–329 (2011)
15. Baxter, N.T., Ruffin, M.T., Rogers, M.A.M., Schloss, P.D.: Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Medicine **8**(1), 37 (2016)
16. Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N.: Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular Systems Biology **10**(11), 766 (2015)
17. Sharpton, T.J.: An introduction to the analysis of shotgun metagenomic data. Frontiers in Plant Science **5**(209), 209 (2014)
18. Meta HIT. http://www.metahit.eu/
19. HMP. http://www.hmpdacc.org/
20. Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A., Alm, E.J.: Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nature Communications **8**(1), 1784 (2017)
21. Pasolli, E., Truong, D.T., Malik, F., Waldron, L., Segata, N.: Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. Plos Computational Biology **12**(7), 1004977 (2016)
22. Hinton GE, T.Y.W. Osindero S: A fast learning algorithm for deep belief nets. Neural Computation **18**(7), 1527–1554 (2006)
23. Lecun Y, H.G. Bengio Y: Deep learning. Nature **521**(7553), 436–444 (2015)

24. Ditzler, G., Polikar, R., Rosen, G.: Multi-layer and recursive neural networks for metagenomic classification. IEEE Transactions on Nanobioscience **14**(6), 608 (2015)
25. Angermueller, C., Pärnamaa, T., Parts, L., Stegle, O.: Deep learning for computational biology. Molecular Systems Biology **12**(7), 878 (2016)
26. Ditzler, G., Polikar, R., Rosen, G.L.: Multi-layer and recursive neural networks for metagenomic classification. IEEE Transactions on NanoBioscience **14**, 608–616 (2015)
27. Q. Zhu, M.P.e.a. Q. Zhu: The phylogenetic tree based deep forest for metagenomic data classification. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 279–282 (2018)
28. Fioravanti, D., Giarratano, Y., Maggio, V., Agostinelli, C., Chierici, M., Jurman, G., Furlanello, C.: Phylogenetic convolutional neural networks in metagenomics. Bmc Bioinformatics **19**(2), 49 (2018)
29. Rhee, S., Seo, S., Kim, S.: Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification, 3527–3534 (2018). doi:10.24963/ijcai.2018/490
30. Zhou, Z.-H., Feng, J.: Deep forest: Towards an alternative to deep neural networks, 3553–3559 (2017). doi:10.24963/ijcai.2017/497
31. Zhou, Z.H., Feng, J.: Deep forest (2017)
32. Han, L., Haihong, Z., Erxin, Y., Yuming, B., Huiying, L.: A clothes classification method based on the gcforest, 429–432 (2018). doi:10.1109/ICIVC.2018.8492801
33. Guo, Y., Liu, S., Li, Z., Shang, X.: Bcdforest: a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data. Bmc Bioinformatics **19**(Suppl 5), 118 (2018)
34. Zhu, Q., Pan, M., Liu, L., Li, B., He, T., Jiang, X., Hu, X.: An ensemble feature selection method based on deep forest for microbiome-wide association studies. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 248–253 (2018)
35. Ma, T., Zhang, A.: Affinitynet: Semi-supervised few-shot learning for disease type prediction. In: AAAI (2018)
36. Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Smola, A.J., Müller, K.R.: Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. IEEE Transactions on Pattern Analysis  Machine Intelligence **25**(5), 623–633 (2003)
37. Lee, J.M., Yoo, C.K., Choi, S.W., Vanrolleghem, P.A., Lee, I.B.: Nonlinear process monitoring using kernel principal component analysis. Chemical Engineering Science **59**(1), 223–234 (2004)
38. Deng, X., Tian, X.: Nonlinear process fault pattern recognition using statistics kernel pca similarity factor. Neurocomputing **121**(18), 298–308 (2013)
39. Romero, A., Gatta, C., Camps-Valls, G.: Unsupervised deep feature extraction for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing **54**(3), 1349–1362 (2016). doi:10.1109/TGRS.2015.2478379
40. MicrobiomeHD. https://zenodo.org/record/1146764#.XDv1O_xS9sN
41. Schubert, A.M., Rogers, M.A.M., Cathrin, R., Jill, M., Petrosino, J.P., Young, V.B., Aronoff, D.M., Schloss, P.D.: Microbiome data distinguish patients with clostridium difficile infection and non-c. difficile-associated diarrhea from healthy controls. Mbio **5**(3), 01021 (2014)
42. Papa, E., Docktor, M., Smillie, C., Weber, S., Preheim, S.P., Gevers, D., Giannoukos, G., Ciulla, D., Tabbaa, D., Ingram, J.: Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. Plos One **7**(6), 39242 (2012)
43. Goodrich, J.K., Waters, J.L., Poole, A.C., Sutter, J.L., Omry, K., Ran, B., Michelle, B., William, V.T., Rob, K., Bell, J.T.: Human genetics shape the gut microbiome. Cell **159**(4), 789–799 (2014)
44. Dixon, B., Candade, N.: Multispectral landuse classification using neural networks and support vector machines: one or the other, or both? International Journal of Remote Sensing **29**(4), 1185–1206
45. Wu, M., Hughes, M.C., Parbhoo, S., Zazzi, M., Roth, V., Doshi-Velez, F.: Beyond sparsity: Tree regularization of deep models for interpretability. In Proceedings of the ThirtySecond AAAI Conference on Artificial Intelligence (2018)

### Figures

**Figure 1 Introduction to the datasets.** Firstly, split the first column of the original OTU table by a semicolon, and connect the split series expanding the columns of the original OTU table. Secondly, hierarchically cluster the microbiota, and accumulate different numbered samples. Thirdly, transpose the table obtained in the previous step. Fourthly, placing the disease state in the metadata set as the final column.

**Figure 2 The overall procedure of KPCCF.** Firstly, apply the RBF kernel function to adjust the datasets better suitable for microbiota datasets. Secondly, input the extracted outputs features of the previous stage into cascade forest.

### Tables

**Figure 3 Decision process of the cdi_schubert dataset.** The model tests the corresponding feature attributes in each item to be classified, and selects the output branch according to its value until the leaf node is reached. During the process of CDI disease prediction, every feature of different microbiota leads to a specific decision.

**Figure 4 The loss and the accuracy of CNN in the different epoch of the ibd_papa dataset.** The loss function is decreasing as epochs grow, and the accuracy outperforms consistently as dimension increasing.

**Figure 5 Accuracy of different principal components.** We have varied the parameter number of components from 5 to 90 with the step size of 5. We can abstract 30 principal components in the family level as it reaches its peak.

**Figure 6 The confusion matrix of 3 classification.** Above all, KPCCF has the best results for its diagonal color is the lightest.

**Figure 7 The $Micro\ F1\ score$ of 4 prediction results in family level respectively.** CDI disease is the most adaptive to classification methods.

**Figure 8 The accuracy of KPCCF prediction results in genus level and family level respectively.** The family level performs more stable than genus level.

**Figure 9 Process of comparative study among various feature reduction methods.** We make comparition between Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Linear Discriminant Analysis (LDA), Least Absolute Shrinkage, Selection Operator (LASSO) dimension reduction methods and our used kPCA method.

**Figure 10 The $Micro\ F1\ score$ of 5 feature reduction models prediction results in each dataset.** The kPCA model in red color has the most prominent performance among these 5 methods with $Micro\ F1\ score$.

**Table 1** Number of datasets samples and features

| ID | Data Sources | Disease label and sample size | f-level Features | g-level Features |
|---|---|---|---|---|
| 1 | cdi_schubert [41] | CDI(93), nonCDI(89), H(154) | 80 | 198 |
| 2 | crc_baxter [15] | CRC(120), H(172), adenoma(198) | 93 | 255 |
| 3 | ibd_papa [42] | nonIBD(24), UC(43), CD(23) | 49 | 142 |
| 4 | ob_goodrich [43] | OB(185), OW(336), H(428) | 79 | 199 |

**Table 2** The performance comparison of different models in disease identification

| Disease | cdi_schubert | | | crc_baxter | | | ibd_papa | | | ob_goodrich | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Var | F1 | Acc | Var | F1 | Acc | Var | F1 | Acc | Var | F1 |
| DT | 0.66 | 0.043 | 0.68 | 0.40 | 0.035 | 0.41 | 0.48 | 0.091 | 0.48 | 0.39 | 0.021 | 0.39 |
| RF | 0.63 | 0.038 | 0.65 | 0.41 | 0.033 | 0.41 | 0.52 | 0.089 | 0.52 | 0.47 | 0.029 | **0.48** |
| CNN | 0.56 | 0.068 | 0.54 | 0.38 | 0.048 | 0.37 | 0.47 | 0.090 | 0.43 | 0.43 | 0.045 | 0.41 |
| CF | 0.67 | 0.053 | 0.69 | 0.40 | 0.042 | 0.4 | 0.53 | 0.082 | 0.54 | 0.46 | 0.026 | 0.44 |
| DF | 0.61 | 0.037 | 0.64 | 0.39 | 0.042 | 0.37 | 0.53 | 0.074 | **0.57** | 0.46 | 0.022 | 0.46 |
| KPCCF | **0.69** | 0.057 | **0.71** | **0.43** | 0.040 | **0.48** | **0.57** | 0.072 | **0.57** | **0.47** | 0.012 | **0.48** |

**Table 3** The prediction accuracy of microbiota data only and mixed data fused with metadata

| Disease | microbiota data | mixed data |
|---|---|---|
| cdi_schubert | 0.69 | 0.68 |
| crc_baxter | 0.43 | 0.48 |
| ibd_papa | 0.57 | 0.59 |
| ob_goodrich | 0.47 | 0.52 |