# Identifying Microbe-Disease Association Based on a Novel Back-Propagation Neural Network Model

Hao Li, Yuqi Wang, Zhen Zhang, Yihong Tan, Zhiping Chen, Xiangyi Wang, Tingrui Pei, and Lei Wang

**Abstract**—Over the years, numerous evidences have demonstrated that microbes living in the human body are closely related to human life activities and human diseases. However, traditional biological experiments are time-consuming and expensive, so it has become a research topic in bioinformatics to predict potential microbe-disease associations by adopting computational methods. In this study, a novel calculative method called BPNNHMDA is proposed to identify potential microbe-disease associations. In BPNNHMDA, a novel neural network model is first designed to infer potential microbe-disease associations, its input signal is a matrix of known microbe-disease associations, and its output signal is matrix of potential microbe-disease associations probabilities. And moreover, in the novel neural network model, a new activation function is designed to activate the hidden layer and the output layer based on the hyperbolic tangent function, and its initial connection weights are optimized by adopting Gaussian Interaction Profile kernel (GIP) similarity for microbes, which can improve the training speed of BPNNHMDA efficiently. Finally, in order to verify the performance of our prediction model, different frameworks such as the Leave-One-Out Cross Validation (LOOCV) and $k$-Fold Cross Validation ($k$-Fold CV) are implemented on BPNNHMDA respectively. Simulation results illustrate that BPNNHMDA can achieve reliable AUCs of 0.9242, $0.9127 \pm 0.0009$ and $0.8955 \pm 0.0018$ in LOOCV, 5-Fold CV and 2-Fold CV separately, which are superior to previous state-of-the-art methods. Furthermore, case studies of inflammatory bowel disease (IBD), asthma and obesity demonstrate that BPNNHMDA has excellent prediction ability in practical applications as well.

**Index Terms**—Microbe, disease, prediction, back-propagation

## 1 Introduction

Microorganisms have been widely found in the oceans, soils, human bodies and other places, and their existences have profound impacts on human life [1]. With the rapid development of high-through sequencing technologies and modern bioinformatics, researches on microbiology have attracted increasing attention from the scientific and medical communities [2]. Recent studies have indicated that there are trillions of microbes in the human body, which substantially outnumber the number of human cells [3]. Meanwhile, microbes participate in different levels of metabolic activities in the human body and are interdependent with the host. For example, microbiota plays important roles in supporting normal digestion and host metabolism [4] and is essential for the development of gut associated lymphoid tissue [5]. Additionally, the gut microbiota shapes the host immune system while the immune system reciprocally shapes and modifies the gut microbiota [6]. Therefore, the ”health” of human microbiome in human body is an important factor for human health.

On the surface, human life activities depend on microorganisms, but in fact, the hosts and their living environment always affect the survival of microorganisms. For, instance, the use of antibiotics and western-style high-fat diets may alter microbial composition [7]. In addition, smoking, stress, radiation, etc. are also the causes of microbial changes [8]. Dysbiosis, which means disruption of the normal microbiome, is associated with immune, metabolic and even neurological diseases [9], [10], [11]. With the rapid advancement of multiomics technologies, many studies have demonstrated that microbial communities are associated with complex diseases. It has been reported that changes in the composition of the gut microbiota may be associated with the pathogenesis of various neurological disorders, including stress, autism, depression, Parkinson’s disease and Alzheimer’s disease [12]. In addition, inflammatory bowel disease (IBD) [13], irritable bowel syndrome (IBS) [14], obesity [15], etc. are considered as potential microbial-based diseases as well. Furthermore, evidences have shown that microbiota can regulate diseases [16]. For example, studies

- *H. Li and L. Wang are with the College of Computer Engineering and Applied Mathematics, Changsha University, Changsha 410022, China, and also with the Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan 411105, China. E-mail: 412198735@qq.com, wanglei@xtu.edu.cn.*
- *Y. Wang and T. Pei are with the Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan 411105, China. E-mail: wangyuqi@smail.xtu.edu.cn, peitingrui@xtu.edu.cn.*
- *Z. Zhang is with the College of Electronic Information and Electrical Engineering, Changsha University, Changsha 410022, China. E-mail: tanglaoya456@126.com.*
- *Y. Tan, Z. Chen, and X. Wang are with the College of Computer Engineering and Applied Mathematics, Changsha University, Changsha 410022, China. E-mail: {yhtan, zpchen, xywang}@ccsu.edu.cn.*

*Manuscript received 23 Oct. 2019; revised 27 Mar. 2020; accepted 3 Apr. 2020. Date of publication 13 Apr. 2020; date of current version 8 Dec. 2021.*
*(Corresponding authors: Lei Wang.)*
*Digital Object Identifier no. 10.1109/TCBB.2020.2986459*

suggest a role for Lactobacillus and Bifidobacterium species in the regulation of anxiety, mood, cognition, pain, and depressive-like symptoms [17], [18]. Asmaa a. Althani *et al.* believed that microbiome may play a significant role in inducing various disease states. Therefore, a deeper understanding of microbe related pathological relationships may provide new ideas for the study of new treatment and prevention strategies for diseases, as well as promote global human health [19].

Considering the inseparable relationships between microbes and human health, many databases and projects of microbes and diseases, including the Human Microbiome Project (HMP) [20] and the Earth Microbial Community Project (EMP) [21], have been developed successively in recent years. Here, the HMP is an interdisciplinary project initiated by the United States, Europe and Asia to provide deeper understandings of microbial composition and its significant role in human disease. EMP was launched in August 2010 with the goal of building a global catalogue of the microbial diversity of the earth [22]. Besides, Ma *et al.* established a human microbe-disease association database (HMDAD) [23], in which, a total of 483 microbe-disease associations between 39 diseases and 292 microbes were covered. Certainly, these projects and databases facilitate the study of complex relationships between microbes and human diseases greatly.

However, traditional biomedical validation experiments are very expensive and time-consuming. Previously, correlation prediction research in various fields of computational biology has achieved satisfactory success, such as lncRNA-disease association prediction [24], [25], [26], [27], [28], miRNA-disease association prediction [29], [30], [31], [32], [33], drug-target interaction prediction [34] and synergistic drug combination identification [35]. These methods provide inspiration for research in the field of microbe-disease association prediction. Hence, for the past few years, accumulating computational methods have been proposed to discover valuable microbe-disease correlation information in advance. Among them, KATZHMDA was the first model proposed for the prediction of potential microbe-disease associations, in which, the topological information of the known microbe-disease association network was adopted to infer potential relationships between microbes and diseases by using the relationship prediction method for social networks [36]. Huang and You *et al.* developed a microbe-disease prediction model by combining two single recommendation models based on neighbor information and graph topology (called NGRHMDA)[37]. Huang *et al.* put forward a computational model called PBHMDA to infer unknown microbe-disease associations by traversing all possible paths between microbes and diseases in a heterogeneous network [38]. Li and Wang *et al.* designed a model named BWNMHMDA to discover underlying microbe-disease associations through calculating the two-way weighted path scores between microbes and diseases in a two-way weighted network as corresponding microbe-disease potential similarities [39]. In addition, Shen *et al.* adopted the restart random walk algorithm in the Spearman correlation-based microbe network and symptom similarity-based disease network to score each candidate microbe-disease pair [40]. Wang *et al.* proposed a prediction model called NBLPIHMDA to identify possible associations between microbes and diseases by implementing a bidirectional label propagation scheme on a heterogeneous network to obtain a disease-microbe correlation score matrix [41]. All these above state-of-the-art methods have demonstrated that computational methods can achieve satisfactory performances in identifying potential microbe-disease associations, which inspires researchers to further explore more novel and effective computational methods for microbe-disease association prediction.

In recent years, neural networks have attracted great interests and have been successfully applied to different fields such as finance, medicine, engineering, geology and physics [42]. Among them, the Back Propagation Neural Network (BPNN) is one of the most popular models and is widely applied to prediction and classification problems [43]. In this manuscript, we aimed to design a novel BPNN model for microbe-disease association prediction. First, we constructed a 3-layer BPNN with 292 nodes per layer, in which, the information of known associations between each microbe and all diseases would be used as the input signals of corresponding nodes in the input layer. Next, we designed a new activation function to activate the hidden layer and the output layer based on the hyperbolic tangent function. Thereafter, the weights and node biases of the BPNN would be continuously updated during the backpropagation process until the whole network reached convergent state. And then, based on the outputs of the convergent network, a microbe-disease correlation score matrix could be finally obtained. Moreover, we adopted the Gaussian Interaction Profile kernel (GIP) similarity to optimize the initial weights of BPNN. And as a result, the training speed of BPNNHMDA could be improved by more than 10 percent. Finally, for evaluating the performance of our prediction model, different frameworks such as the Leave-One-Out Cross Validation (LOOCV) and $k$-Fold Cross Validation ($k$-Fold CV) would be implemented for BPNNHMDA based on the dataset downloaded from the HMDAD database. Simulation results illustrated that BPNNHMDA could achieve reliable AUCs of 0.9242, 0.9127 ± 0.0009 and 0.8955 ± 0.0018 in LOOCV, 5-Fold CV and 2-Fold CV separately, which outperformed previous state-of-the-art methods significantly. Additionally, we further conducted case studies of inflammatory bowel disease, asthma and obesity on BPNNHMDA, and simulation results demonstrated that BPNNHMDA could effectively identify potential microbe-disease associations as well.

## 2 Material

In BPNNHMDA, since known microbes-disease associations would be taken as the valid input data, then, we first downloaded known microbe-disease associations from the HMDAD database in this section. After removing duplicated associations, we finally obtained 450 known microbe-disease associations including 39 human diseases and 292 microbes from 61 publications. For convenience, we denoted the dataset of 450 known microbe-disease associations as $S_{MD}$, the dataset of 39 human diseases as $S_D$, and the dataset of 292 microbes as $S_M$ respectively. Thus, we could construct a 292 × 39 dimensional adjacency matrix $A$ as well. In the adjacency matrix $A$, there is $A(i,j)$=1, if and only if there is a known association in $S_{MD}$ between the $i$th microbe in $S_M$

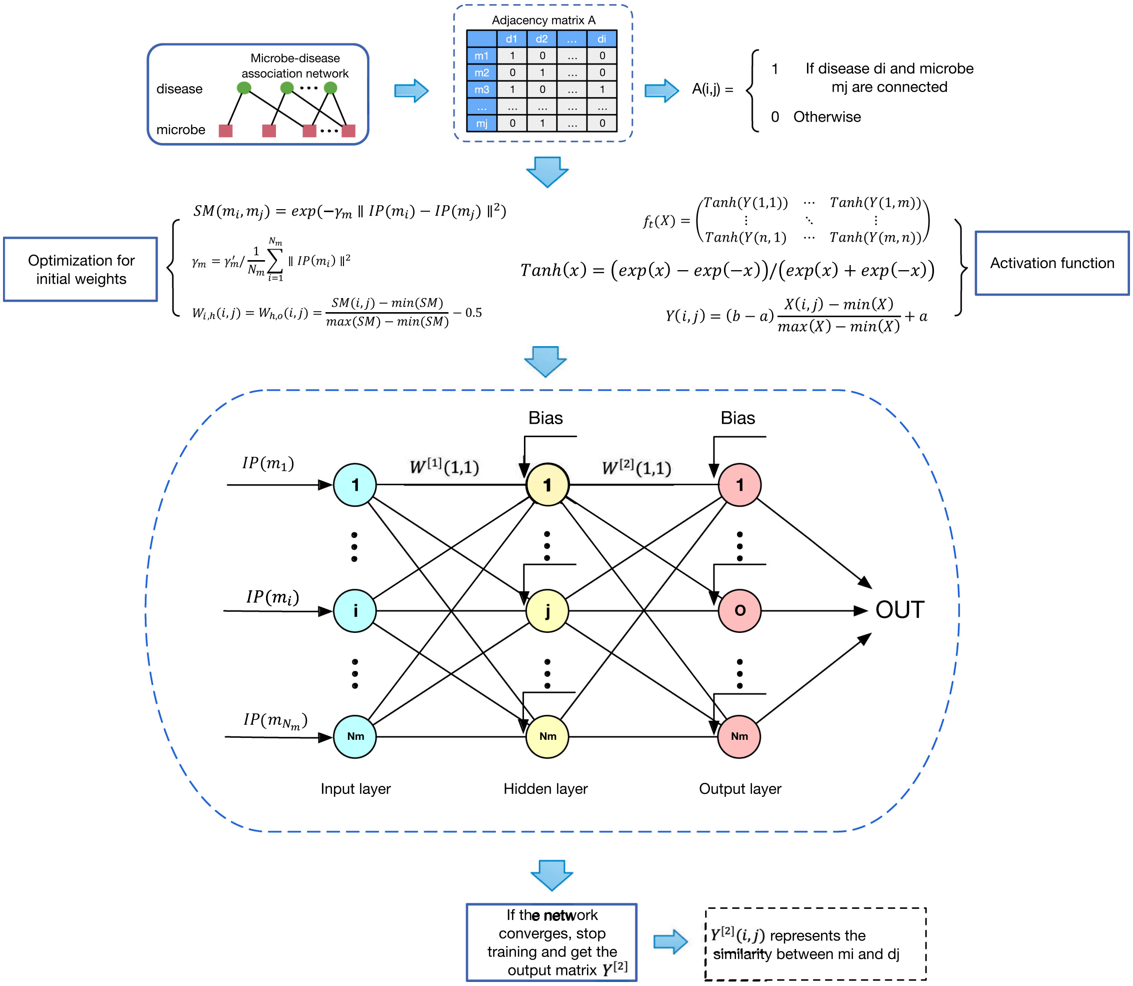

Microbe-disease association network
disease
microbe
Adjacency matrix A
d1 d2 … di
m1 1 0 … 0
m2 0 1 … 0
m3 1 0 … 1
… … … … …
mj 0 1 … 0
A(i,j) =
1 If disease di and microbe mj are connected
0 Otherwise
Optimization for initial weights
$SM(m_i,m_j) = exp(-\gamma_m \parallel IP(m_i) - IP(m_j) \parallel^2)$
$\gamma_m = \gamma_m' / \frac{1}{N_m}\sum_{i=1}^{N_m} \parallel IP(m_i) \parallel^2$
$W_{i,h}(i,j) = W_{h,o}(i,j) = \frac{SM(i,j) - min(SM)}{max(SM) - min(SM)} - 0.5$
$f_t(X) = \begin{pmatrix} Tanh(Y(1,1)) & \cdots & Tanh(Y(1,m)) \\ \vdots & \ddots & \vdots \\ Tanh(Y(n,1) & \cdots & Tanh(Y(m,n)) \end{pmatrix}$
$Tanh(x) = (exp(x) - exp(-x))/(exp(x) + exp(-x))$
$Y(i,j) = (b-a)\frac{X(i,j) - min(X)}{max(X) - min(X)} + a$
Activation function
Bias
Bias
$IP(m_1)$
$IP(m_i)$
$IP(m_{N_m})$
$W^{[1]}(1,1)$
$W^{[2]}(1,1)$
1
i
Nm
1
j
Nm
1
O
Nm
OUT
Input layer
Hidden layer
Output layer
If the network converges, stop training and get the output matrix $Y^{[2]}$
$Y^{[2]}(i,j)$ represents the similarity between mi and dj

Fig. 1. The flowchat of BPNNHMDA.

and the $j$th disease in $S_D$, otherwise, there is $A(i,j) = 0$. Thereafter, the adjacency matrix $A$ would be utilized as the valid input data of BPNNHMDA ultimately.

## 3 Methods

Back Propagation Neural Network [44] is one of the most well-known supervised learning neural networks and is characterized by its back propagation learning algorithm [45]. Up to now, BPNN has been widely used in various prediction model designs, such as stock market prediction [46], time series forecasting [47], cancer prediction [48], etc. Inspired by the excellent prediction performance of BPNN, in this manuscript, we designed a prediction model BPNNHMDA applicable to human microbe-disease association prediction based on an improved BPNN. The key ideas of BPNNHMDA were to use known microbe-disease association pairs as effective input data of the BPNN, and then, the whole network would be updated iteratively by adopting the back propagation algorithm to predict potential association scores for each pair of microbe and disease.

### 3.1 Model Design of BPNNHMDA

#### 3.1.1 Architecture of BPNNHMDA

As illustrated in the following Fig. 1, BPNNHMDA consists of three layers including the input layer, the output layer and the hidden layer. Considering that neural networks with the number of hidden nodes equaling to the number of input nodes will have better predictive results [45], and in addition, we have obtained 292 microbes and 39 diseases from the HMDAD database, therefore, we finally designed 292 nodes for each layer in BPNNHMDA. Here, the $i$th node in the input layer represents the $i$th microbe in $S_M$, and the input data of the $i$th node in the input layer consists of these known associations between the $i$th microbe and all diseases, namely the $i$th row of the adjacency matrix $A$, which is a 39 dimensional vector.

Comparing with traditional BPNN, the feature of BPNNHMDA lies in that the input data of the whole neural network is a 292 × 39 dimensional matrix. The reason is that all nodes in each pair of neighboring layers are interconnected. That is to say, not only each node in the input layer is connected to all nodes in the hidden layer, but also each node in the hidden layer is connected to all nodes in the output layer. Thus, every node in the hidden layer will receive a unique 39 dimensional output from each of these 292 nodes in the input layer respectively, and then, after processing, it will further forward the 39 dimensional output of itself to all 292 nodes in the output layer. Therefore, the input data of the whole neural network in BPNNHMDA will be a 292 × 39 dimensional matrix, which will be propagated through the whole neural network in this kind of matrix form as well.

#### 3.1.2 The Setting of Initial Weights and Biases of BPNN

The method for controlling a neural network is to set and adjust its connection weights and node biases. In traditional BPNN, the initial connection weights are usually set to random values. However, since BPNN uses gradient descent method to adjust the network error and weights, then there are local minimum points of error. After the initial connection weights have been set randomly for BPNN, the errors will reach the minimum points along the slopes of the error functions. That is to say, the initial values of connection weights will directly affect the training efficiency and convergence speed of BPNN. And moreover, if the initial weights have been selected improperly, the convergence direction of BPNN will be divergent, which may cause network oscillation [49]. Therefore, it is necessary to optimize the initial connection weights of BPNN obviously.

In our method, the optimization of the initial connection weights consists of two major steps. First, based on the assumption that two microbes will be more similar to each other, if there are more common human diseases having been proven to be associated with both of them, we will adopt the Gaussian Interaction Profile kernel similarity [36] to obtain a similarity matrix $S_M$ for microbes as follows:

$$SM(i,j) = exp(-\gamma_m \| IP(m_i) - IP(m_j) \|^2). \tag{1}$$

Here, $m_i$ and $m_j$ denote the $i$th and $j$th microbes respectively in $SM$. $IP(m_i)$ and $IP(m_j)$ denote the $i$th and $j$th row in the adjacency matrix $A$, respectively. $||X||$ indicates the norm of the vector $X$. Moreover, the parameter $\gamma_m$ can be further calculated according to the following formula (2).

$$\gamma_m = \gamma'_m / \left( \frac{1}{N_m} \sum_{i=1}^{N_m} \| IP(m_i) \|^2 \right). \tag{2}$$

Where $\gamma'_m$ is a parameter for controlling the Gaussian kernel bandwidth, and in this manuscript, we will set $\gamma'_m$ to 1 according to relevant studies [50].

Thus, based on above newly obtained similarity matrix $SM$, we can construct two identical initial weight matrices $W^{[1]}$ and $W^{[2]}$ as follows:

$$W^{[1]}(i,j) = W^{[2]}(i,j) = \frac{SM(i,j) - min(SM)}{max(SM) - min(SM)} - 0.5. \tag{3}$$

Here, $min(SM)$ and $max(SM)$ represent the minimum and maximum values of the elements in the matrix $SM$ respectively.

Thereafter, based on these two 292 × 292 dimensional matrices $W^{[1]}$ and $W^{[2]}$, we will adopt $W^{[1]}(i,j)$ as the weight between the $i$th node of the input layer and the $j$th node of the hidden layer in BPNNHMDA, and $W^{[2]}(i,j)$ as the weight between the $i$th node of the hidden layer and the $j$th node of the output layer in BPNNHMDA, respectively. Additionally, in order to improve the activity of nodes, we will set an initial bias with random value between $[-\rho, \rho]$ for each node in the hidden layer and output layer in BPNNHMDA as well. Here, the parameter $\rho$ is considered as a factor that may affect the prediction performance and stability of BPNNHMDA. For convenience, we will represent the biases of hidden layer nodes and output layer nodes as two 292 dimensional vectors $\theta^{[1]}$ and $\theta^{[2]}$ separately. Here, $\theta_i^{[1]}$ and $\theta_j^{[2]}$ represent the bias value of the $i$th node of the hidden layer and the $j$th node of the output layer in BPNNHMDA respectively.

#### 3.1.3 Activation Function

In the neural networks, the activation function is introduced to increase the nonlinearity of the network, and the activation functions commonly used in previous studies include sigmoid (Logistic) function, hyperbolic tangent (Tanh) function, sine or cosine function, etc. There are some heuristic rules for selecting activation functions according to previous researches. For example, Klimasauskas [51] recommended Logistic activation functions for classification problems involving learning average behavior, and Tanh functions for problems involving learning deviations from the mean (e.g., prediction problems). However, using traditional Tanh function as activation function also has some disadvantages. For example, if the input value of Tanh function is too large or too small, the gradient will disappear, which will lead to the loss of feedforward layer feature information. Therefore, the input value of Tanh function shall be normalized to a smaller symmetric interval. In addition, traditional Tanh function is only suitable for single-valued activation, while BPNNHMDA propagates information in the form of matrix between neighboring layers. Hence, based on the traditional Tanh function, we designed an improved activation function $f_t$ for both hidden layer and output layer in BPNNHMDA as follows:

$$f_t(X) = \begin{pmatrix} Tanh(Y(1,1)) & \cdots & Tanh(Y(1,n)) \\ \vdots & \ddots & \vdots \\ Tanh(Y(m,1) & \cdots & Tanh(Y(m,n)) \end{pmatrix} \tag{4}$$

$$Tanh(x) = (exp(x) - exp(-x))/(exp(x) + exp(-x)) \tag{5}$$

$$Y(i,j) = (b-a)\frac{X(i,j) - min(X)}{max(X) - min(X)} + a. \tag{6}$$

In above Equation (4), $X$ represents an $m \times n$ dimensional matrix that needs to be activated. Additionally, Equation (6) adopts the batch normalization to linearly convert value ranges of all elements in $X$ to the interval [$a$, $b$]. When the activation function $f_t$ is used to activate the hidden layer and output layer in BPNNHMDA, the parameters $a$ and $b$ will be set to $-\alpha$ and $\alpha$ respectively. Obviously, the parameter $\alpha$ controls the range of input values of the Tanh function given in above Equation (5), which affects the sensitivity of the activation function $f_t$. Therefore, the parameter $\alpha$ is a factor that may affect the prediction performance of BPNNHMDA as well.

### 3.2 Training Algorithm of BPNNHMDA

Algorithm 1 describes the BPNNHMDA method to predict potential microbe-disease associations. In BPNNHMDA, the input value of each node is a 39 dimensional vector, and network signals will be propagated in the form of 292 × 39 dimensional matrixes between neighboring layers. For convenience of expression, we present some nomenclatures that will be used in the following sections first:

$I^{[1]}$: The input matrix of the hidden layer in BPNNHMDA.
$Y^{[1]}$: The output matrix of the hidden layer in BPNNHMDA.
$I^{[2]}$: The input matrix of the output layer in BPNNHMDA.
$Y^{[2]}$: The output matrix of the output layer in BPNNHMDA.

In addition, the errors of nodes in the hidden layer and output layer are represented by 39 dimensional vectors as follows:

$e^{[1]}$: The error of the hidden layer in BPNNHMDA.
$e^{[2]}$: The error of the output layer in BPNNHMDA.

### *3.2.1 Feed-Forward Computation*

As described above, the adjacency matrix $A$ will be adopted as the input signals of the input layer in BPNNHMDA. However, considering that in the HMDAD database, there are not only some microbes related with a lot of diseases, but also some microbes related with few diseases. For example, Bacteroidetes is related with 10 diseases, while Clostridiales is related with 1 disease only. Therefore, in order to keep the same amount of useful information for each node in the input layer of BPNNHMDA, we will first adopt the across channel normalization scheme [52] to normalize the adjacency matrix $A$ as follows: Let $IP(m_i)$ denote the $i$th row in the adjacency matrix $A$, and $NZ(X)$ represent the number of non-zero elements in the vector $X$, then based on the across channel normalization scheme, for the $i$th node in the input layer of BPNNHMDA, it can obtain a normalized vector $NI_i$ =( $NI_{i1}$, $NI_{i2}$,, $NI_{iN_D}$) as its output according to the following Equation (7):

$$NI_{ij} = IP_j(m_i)/NZ(IP(m_i)) \quad (i = 1, 2, \dots, N_M). \tag{7}$$

Here, $N_D$ and $N_M$ represent the total numbers of diseases in $S_D$ and microbes in $S_M$ respectively. Apparently, there are $N_D$ = 39 and $N_M$ = 292 in this manuscript.

Next, for hidden layer of BPNNHMDA, it can obtain a normalized matrix $Y^{[1]}$ as its output according to the following Equations:

$$Y^{[1]}(i,j) = \frac{Y^{[1]}(i,j) - min(Y^{[1]})}{max(Y^{[1]}) - min(Y^{[1]})} \tag{8}$$

$$Y^{[1]} = f_t\left(I^{[1]}\right) \tag{9}$$

$$I_j^{[1]} = \sum_{i=1}^{N_M} W^{[1]}(i,j) \times NI_i + \theta_j^{[1]} \quad (j = 1, 2, \dots, N_M). \tag{10}$$

Here, let $X$ be a matrix, then $X_j$ denotes the $j$th row in $X$, and $X(i,j)$ represents the element locating at the $i$th row and the $j$th column in $X$. Additionally, $\theta_j^{[1]}$ denotes the $j$th element in the vector $\theta^{[1]}$, i.e., the bias of the $j$th node in the hidden layer of BPNNHMDA.

Similarly, for output layer of BPNNHMDA, it can obtain a normalized matrix $Y^{[2]}$ as its output according to the following Equations:

$$Y^{[2]}(i,j) = \frac{Y^{[2]}(i,j) - min\left(Y^{[2]}\right)}{max(Y^{[2]}) - min(Y^{[2]})} \tag{11}$$

$$Y^{[2]} = f_t\left(I^{[2]}\right) \tag{12}$$

$$I_j^{[2]} = \sum_{i=1}^{N_M} W^{[2]}(i,j) \times Y_i^{[1]} + \theta_j^{[2]} (j = 1, 2, \dots, N_M). \tag{13}$$

Where $\theta_j^{[2]}$ denotes the $j$th element in the vector $\theta^{[2]}$, i.e., the bias of the $j$th node in the output layer of BPNNHMDA.

### *3.2.2 Back Propagation to the Output Layer*

For each node in the output layer of BPNNHMDA, its error can be calculated based on its target output and actual output. For instance, let $m_i$ be the $i$th node in the output layer of BPNNHMDA. Let $IP(m_i)$ = $\{P_1, P_2,, P_{N_D}\}$ and $T_i$ = $\{T_{i1}$, $T_{i2},, T_{iK}\}$, where there is $P_{T_{ij}}$ = 1, and $N_{T_i}$ denote the number of elements in $T_i$, then the error of the $i$th node $m_i$ in the output layer of BPNNHMDA can be calculated as follows:

$$e_i^{[2]} = \frac{\sum_{t \in T_i}\left(1 - \left(Y^{[2]}(i,t)\right)^2\right) * \left(1 - Y^{[2]}(i,t)\right)}{N_{T_i}} \quad (i = 1, 2, \dots, N_M). \tag{14}$$

Based on above Equation (14), we can obtain the mean absolute deviation (*MAD*) of BPNNHMDA as follows:

$$MAD = \frac{\sum_{i=1}^{Nm}\left|e_i^{[2]}\right|}{N_M}. \tag{15}$$

In this manuscript, we will adopt $MAD$ as the measure to evaluate the final network error of BPNNHMDA.

### *3.2.3 Back Propagation to the Hidden Layer*

Similar to above Section 3.2.2, for each node in the hidden layer of BPNNHMDA, its error can be calculated based on its target output and actual output as well. Therefore, through considering both the errors of nodes in the output layer and the connection weights between nodes in the hidden layer and nodes in the output layer, the error of the $i$th node in the hidden layer of BPNNHMDA can be calculated as follows:

$$e_i^{[1]} = \frac{\sum_{t \in T_i}\left(1 - \left(Y^{[1]}(i,t)\right)^2\right)}{N_{T_i}} * \sum_{k=1}^{N_M}\left(e_k^{[2]} * W^{[2]}(i,k)\right) \tag{16}$$
$$(i = 1, 2, \dots, N_M).$$

### *3.2.4 Updating of the Connection Weights and Biases*

In BPNNHMDA, the connection weights and biases including $W^{[1]}$, $W^{[2]}$, $\theta^{[1]}$ and $\theta^{[2]}$ can be updated iteratively according to the following formulas:

$$W^{[1]}(i,j) = W^{[1]}(i,j) + l * e_j^{[1]} * \sum_{t \in T_i} NI(i,t) \tag{17}$$

$$W^{[2]}(i,j) = W^{[2]}(i,j) + l * e_j^{[2]} * \sum_{t \in T_i} Y^{[1]}(i,t) \tag{18}$$

$$\theta^{[1]} = \theta^{[1]} + l * e^{[1]} \tag{19}$$

$$\theta^{[2]} = \theta^{[2]} + l * e^{[2]}. \tag{20}$$

Where the parameter $l \in$ [0,1] is the learning rate, according to latter simulation results, its value will be set to 0.1 in this manuscript.

#### *3.2.5 Determining Convergence Conditions for the Network Error of BPNNHMDA*

During the training process, in the ideal case, the network error of BPNNHMDA will be gradually reduced by the adjustment of the back propagation algorithm until an optimal solution is reached and stabilized. However, it is not that the network error is as small as possible, since a too small network error may cause over-fitting, which may weaken the prediction performance of BPNNHMDA on the contrary. Hence, in order to avoid that the network error of BPNNHMDA may begin to increase after the optimal solution has been reached, we set the following convergence rules, i.e., the training process will stop when one of the following conditions is met:

1) The mean absolute deviation ($MAD$) of BPNNHMDA is lower than $10^{-3}$.
2) The times of iterations of the connection weights is more than 100, and at the same time, the value of $MAD$ will no longer continue decreasing.
3) The times of iterations of the connection weights has reached 1,000.

Obviously, if one of the above three conditions is met, then the values of elements in both $W^{[1]}$ and $W^{[2]}$ will be normalized to the range of [-0.5,0.5], while the values of elements in both $\theta^{[1]}$ and $\theta^{[2]}$ will be normalized to the range of $[-\rho,\ \rho]$ simultaneously. Moreover, while the training process has reached one of these termination conditions, it is obvious that the output matrix of BPNNHMDA can be used as the final microbe-disease correlation score matrix. Especially, for any given microbe $m_i$ and disease $d_j$, the potential similarity score between them can be calculated as follows:

$$sim\left(m_i, d_j\right) = Y^{[2]}(i,j). \tag{21}$$

**Algorithm 1.** BPNNHMDA

```
Input: microbe-disease adjacency matrix $A$, parameters $\alpha$ and $\rho$.
Output: output matrix of output layer $Y^{[2]}$
Step1: Calculate microbe GIP similarity matrix $SM$ by Equations (1) and (2);
Step2: Calculate normalized initial weight matrix $W^{[1]}$ and $W^{[2]}$ by Equation (3);
Step3: Generate random initial bias vector $\theta^{[1]}$ and $\theta^{[2]}$;
Step4: Batch normalize $W^{[1]}$ and $W^{[2]}$ to [-0.5, 0.5] by Equation (3);
Step5: Batch normalize $\theta^{[1]}$ and $\theta^{[2]}$ to $[-\rho,\ \rho]$ by Equation (3);
Step6: Calculate across channel normalization matrix $NI$ by Equation (7);
Step7: Calculate the input and output matrix of hidden layer by Equations (8), (9), and (10);
Step8: Calculate the input and output matrix of output layer by Equations (11), (12), and (13);
Step9: Calculate the error of each output layer node by Equation (14);
Step10: Calculate network error $MAD$ by Equation (15);
Step11: Calculate the error of each hidden layer node by Equation (16);
Step12: Update the connection weights and biases by Equations (17), (18), (19), and (20);
Step13: Repeat Step4-12 until the convergence condition is met;
Step14: Obtain the predicted association matrix $Y^{[2]}$;
```

# 4 RESULTS

## 4.1 Performance Evaluation

In this section, in order to estimate the performance of BPNNHMDA, we implemented different frameworks such as the LOOCV and $k$-Fold CV on dataset downloaded from the HMDAD database. In LOOCV, each known microbe-disease association will be taken as a test sample in turn, while the remaining 449 known microbe-disease associations are used as training samples, and besides, all unconfirmed microbe-disease pairs will constitute candidate samples. Finally, after performing the BPNNHMDA, the similarity scores of the test sample will be ranked with candidate samples. Similarly, in $k$-Fold CV, we will divide all known microbe-disease association pairs into $k$ equal parts, and then, each part will be taken as the test samples in turn, while the remaining $k$-1 parts are regarded as the training samples, and besides, all unconfirmed microbe-disease pairs will be taken as the candidate samples. Finally, after performing the BPNNHMDA, the similarity scores of the test samples will be ranked with candidate samples as well. Moreover, in both LOOCV and $k$-Fold CV, true positive rate (TPR, sensitivity) and false positive rate (FPR, 1-specificity) will be calculated to plot the receiver operating characteristics (ROC) curve by setting various thresholds. Here, sensitivity denotes the percentage of the test samples ranked above the given thresholds, while specificity denotes the percentage of candidate samples with ranks below the given thresholds. Thereafter, the area under the ROC curve (AUC) will be used to evaluate the performance of our prediction model. Obviously, the closer the AUC value is to 1, the better the prediction performance of BPNNHMDA will be, and the AUC value of 0.5 indicates pure random prediction performance. Moreover, during simulation, we will implement the LOOCV, 5-Fold CV and 2-Fold CV for 50 times, 100 times and 100 times respectively, and take the average value of these AUCs as our final value of AUC. And as a result, BPNNHMDA can achieve reliable AUCs of 0.9242, 0.9127 ± 0.0009 and 0.8955 ± 0.0018 under the frameworks of LOOCV, 5-Fold CV and 2-Fold CV respectively, which demonstrated that BPNNHMDA has reliable and stable prediction performance. We further performed both BPNNHMDA and BPNNHMDA without initial weights optimization for 100 times respectively. As a result, we found that the time required for training of BPNNHMDA increased by at least 10 percent after removing the initial weights optimization. Furthermore, we also adopted the microbe cosine similarity [53] for initial weights optimization for comparison. The simulation results showed that the training time of BPNNHMDA based on Gaussian interaction profile kernel similarity is nearly 24 percent less than that based on cosine similarity.

## 4.2 Analysis of the Parameters $\alpha$ and $\rho$

In above Sections 3.1.2 and 3.1.3, we defined two parameters $\alpha$ and $\rho$, which may affect the prediction performance of BPNNHMDA. Among them, the parameter $\alpha$ is utilized to control the range of input values of the Tanh function, which will affect the sensitivity of the activation function. Obviously, too large or too small activation sensitivity is detrimental to the activation process, which will weaken the prediction performance of BPNNHMDA on the contrary.

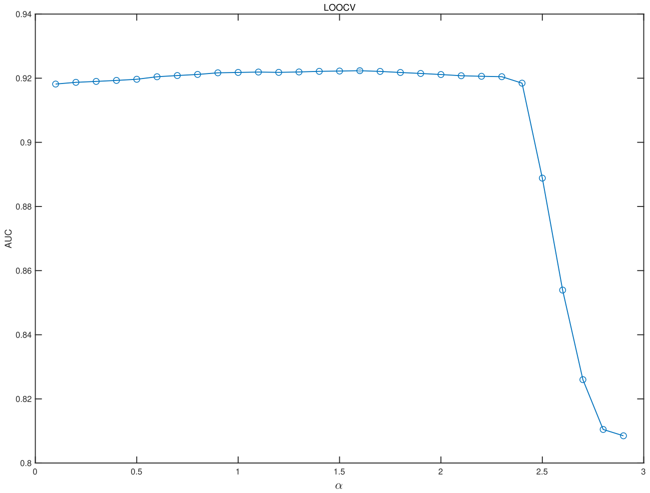


Fig. 2. The mean AUCs in LOOCV while $\alpha$ varying from 0.1 to 2.9 with increment of 0.1 and $\rho$ = 4. In addition, the solid point represents the largest AUC of 0.9223 achieved by BPNNHMDA while $\alpha$ = 1.6.

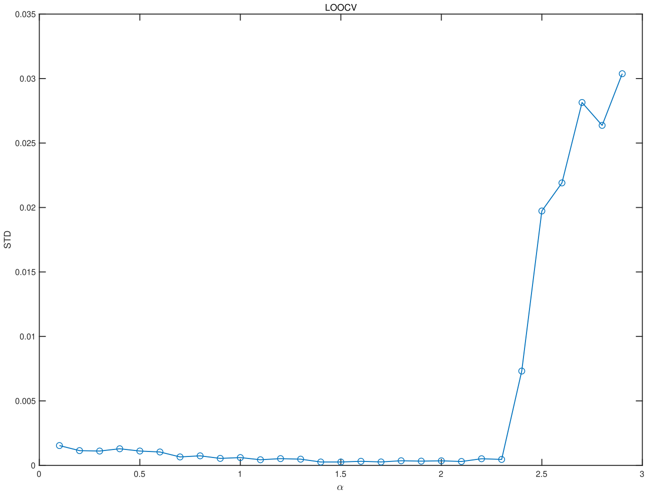


Fig. 3. The average standard deviation of AUCs in LOOCV while $\alpha$ varying from 0.1 to 2.9 with increment of 0.1 and $\rho$ = 4.

Additionally, the parameter $\rho$ is used to control the range of initial values of biases. Considering that the function of biases is to improve the node activity, it is obvious that if the value of $\rho$ is too large, then the predictive stability of BPNNHMDA may be reduced, but if the value of $\rho$ is too small, then the node activity in BPNNHMDA will not be improved effectively. In order to estimate the effects of these two parameters, during simulation, we will keep $\rho$ constant while analysing $\alpha$, and keep $\alpha$ unchanged while analysing $\rho$. And as a result, the following Fig. 2 shows the average AUCs in LOOCV while $\rho$ = 4 and the parameter $\alpha$ varies from 0.1 to 2.9. And the following Fig. 3 illustrates the average standard deviation (STD) of AUCs in LOOCV while $\rho$ = 4 and the parameter $\alpha$ varies from 0.1 to 2.9. From observing the Fig. 2, it is easy to see that the AUCs increase slightly when $\alpha$ varies from 0.1 to 1.6, and then the AUCs will decrease significantly when $\alpha$ varies from 2.4 to 2.9. Additionally, from observing the Fig. 3, it is easy to see that the average standard deviation of AUCs increases significantly when $\alpha$ $>2.3$, which means that the prediction performance of BPNNHMDA becomes unstable when $\alpha$ is bigger than 2.3. Moreover, it is obvious that when $\alpha$ is set to 1.6, BPNNHMDA can achieve the largest AUC of 0.9223 and maintain stable performance.

Similarly, the following Table 1 shows the average AUCs and STDs achieved by BPNNHMDA in LOOCV while $\alpha$ = 1.6 and the parameter $\rho$ varies from 1 to 9. Obviously, from observing the Table 1, we can see that the best AUC of 0.9258 is achieved by BPNNHMDA when $\rho$ is set to 8. In addition, with the increasing of the value of $\rho$, the average standard deviation will increase synchronously, which demonstrates that larger random interval of bias will make the stability of BPNNHMDA weaker. In this manuscript, we consider the average standard deviation of AUCs less than 0.001 as a reliable result. Therefore, we will set $\rho$ to 6 during latter simulations.

Next, we further calculate the mean AUCs achieved by BPNNHMDA in LOOCV while both $\alpha$ and $\rho$ are set to different values, and the simulation results are shown in the following Fig. 4. In Fig. 4, the parameter $\alpha$ varies from 0.1 to 3.1 while the parameter $\rho$ varies from 1 to 9 simultaneously. After excluding parameter combinations with standard deviation greater than 0.001, we find that BPNNHMDA can achieve the largest AUC of 0.9242 when $\alpha$ = 1.6 and $\rho$ = 6.

### 4.3 Comparison With State-of-the-Art Methods

In this section, we will further evaluate the performance of BPNNHMDA by comparing it with state-of-the-art competing methods including BWNMHMDA [39], NBLPIHMDA [41], LRLSHMDA [54], and KATZHMDA [36] based on the dataset downloaded from the HMDAD database. And as a result, the following Fig. 5 and Table 2 illustrate the comparison results between them under the framework of LOOCV and $k$-Fold CV respectively. From observing the Fig. 5, it is easy to see that BPNNHMDA can achieve reliable AUC of 0.9242 in LOOCV, which is superior to the AUCs achieved by BWNMHMDA (with AUC of 0.9127), NBLPIHMDA (with AUC of 0.9041), LRLSHMDA (with AUC of 0.8909) and KATZHMDA (with AUC of 0.8382) respectively. And in addition, in 5-Fold CV and 2-Fold CV, BPNNHMDA can achieve reliable AUC of 0.9127 ± 0.0009 and 0.8955 ± 0.0018 respectively, which outperforms the AUCs achieved by BWNMHMDA (with AUCs of 0.8967 ± 0.0027 and 0.8668 ± 0.0043), NBLPIHMDA (with AUCs of 0.8958 ± 0.0027 and 0.8799 ± 0.0062), LRLSHMDA (with AUCs of 0.8794 ± 0.0029 and 0.8595 ± 0.0056) and KATZHMDA (with AUCs of 0.8301 ± 0.0033 and 0.8171 ± 0.0051) as well. Furthermore,

TABLE 1
The Mean AUCs in LOOCV While $\alpha$ = 1.6 and the Parameter $\rho$ Varies From 1 to 9 With Increment of 1

| $\rho$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| AUC | 0.9184 | 0.9195 | 0.9209 | 0.9224 | 0.9231 | 0.9242 | 0.9248 | 0.9258 | 0.9239 |
| STD | 0.00002 | 0.00008 | 0.0001 | 0.0002 | 0.0003 | 0.0007 | 0.0013 | 0.001 | 0.0082 |

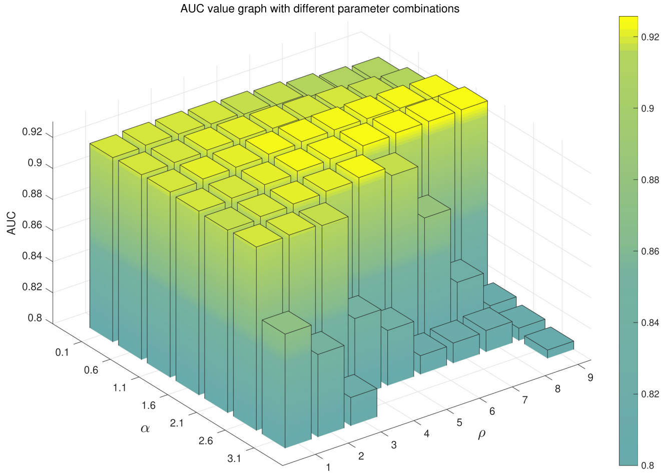


Fig. 4. The mean AUCs achieved by BPNNHMDA in LOOCV while $\alpha$ varies from 0.1 to 3.1 with increment of 0.5 and $\rho$ varies from 1 to 9 with increment of 1.

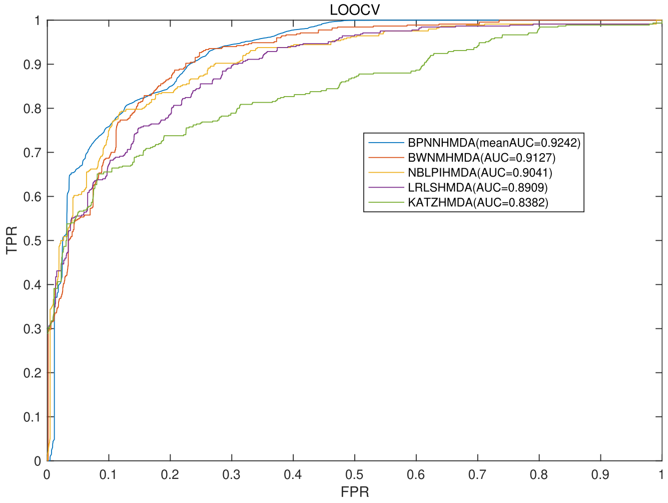


Fig. 5. Comparison of AUCs achieved by BPNNHMDA, BWNMHDA, NBLPIHMDA, LRLSHMDA, KATZHMDA in LOOCV.

TABLE 2
Comparison of AUCs Achieved by BPNNHMDA, BWNMHMDA, NBLPIHMDA, LRLSHMDA, KATZHMDA in 5-Fold CV and 2-Fold CV

| Method | AUC of 5-Fold CV | AUC of 2-Fold CV |
|---|---|---|
| BPNNHMDA | 0.9127 ± 0.0009 | 0.8955 ± 0.0018 |
| BWNMHMDA | 0.8967 ± 0.0027 | 0.8668 ± 0.0043 |
| NBLPIHMDA | 0.8958 ± 0.0027 | 0.8799 ± 0.0062 |
| LRLSHMDA | 0.8794 ± 0.0029 | 0.8595 ± 0.0056 |
| KATZHMDA | 0.8301 ± 0.0033 | 0.8171 ± 0.0051 |

TABLE 3
P-Values Achieved by Paired t-Test the AUCs of 19 Diseases

| Method | BWNMHMDA | NBLPIHMDA | LRLSHMDA | KATZHMDA |
|---|---|---|---|---|
| P-value | 4.89E-03 | 1.33E-02 | 6.44E-03 | 1.42E-05 |

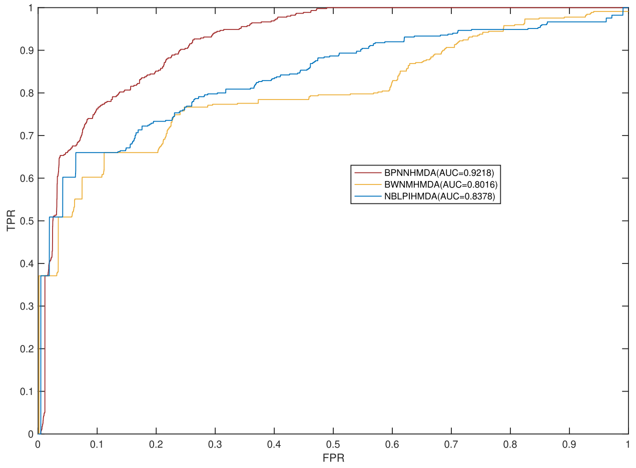


Fig. 6. The performance of BPNNHMDA, BWNMHMDA and NBLPIHMDA on prediction of new microbe-associated diseases.

in order to test the prediction performance of BPNNHMDA on the real dataset, we calculated AUC values for various diseases in BPNNHMDA and other prediction methods based on the LOOCV framework. Diseases associated with too few microbes are not enough to evaluate the performance of the prediction methods, therefore the diseases related with less than four known microbes would be excluded from the experiment. We finally collected prediction results from 19 different diseases, and these results were published in the Supplementary Table S1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2020.2986459. The average AUC values of BPNNHMDA, BWNMHMDA, NBLPIHMDA, LRLSHMDA, and KATZHMDA are 81.35, 65.37, 73.11, 71.71, and 56.91 percent, respectively. We then performed the paired t-test in terms of the AUCs of 19 different diseases. The significance test results indicated that the performance of BPNNHMDA is significantly better than other competing methods (p-value $< 0.05$, as illustrated in the Table 3). In the LOOCV framework, in order to evaluate the performance of BPNNHMDA for predicting the associations between new microbe and related diseases, for any given microbe $m_i$, we will exclude all known associations of microbe $m_i$ when calculating the score of association with microbe $m_i$ and all diseases, and only relying on the remaining associations. As illustrated in the Fig. 6, BPNNHMDA can reach a promising AUC of 0.9218 for new microbe-associated diseases prediction, while the AUCs of BWNMHMDA and NBLPIHMDA are 0.8016 and 0.8378, respectively. That is to say, BPNNHMDA has better prediction performance than all these state-of-the-art competing methods.

## 5 Case Study

In this section, we selected three kinds of common human diseases including inflammatory bowel disease, asthma and obesity to further verify the prediction performance of BPNNHMDA. During simulation, we excluded 450 known associations from the prediction results of BPNNHMDA, and selected the top 15 microbes associated with these three kinds of diseases as our case studies. Among these three kinds of diseases, first, inflammatory bowel disease is a chronically devoid of gastrointestinal diseases that has caused huge costs to the healthcare system and society [55], and accumulating evidences show that the incidence of IBD is increasing in recent years [56].

TABLE 4
Top 15 Potential IBD-Related Microbes Predicted by BPNNHMDA and all of These Microbes Have Been Supported by Literature Evidences

| Rank | Microbe | Evidence |
|---|---|---|
| 1 | Bacteroidetes | PMID: 25307765, 28842640 |
| 2 | Prevotella | PMID: 25307765, 24013298 |
| 3 | Firmicutes | PMID: 25307765, 28842640 |
| 4 | Helicobacter pylori | PMID: 22221289 |
| 5 | Haemophilus | PMID: 24013298 |
| 6 | Clostridium coccoides | PMID: 19235886 |
| 7 | Lactobacillus | PMID: 26340825, 17897884 |
| 8 | Staphylococcus aureus | PMID: 23885156 |
| 9 | Clostridium difficile | PMID: 28785153 |
| 10 | Enterobacteriaceae | PMID: 23013615 |
| 11 | Veillonella | PMID: 28842640, 24013298 |
| 12 | Clostridia | PMID: 25307765 |
| 13 | Bacteroides | PMID: 27999802 |
| 14 | Faecalibacterium prausnitzii | PMID: 28683448 |
| 15 | Staphylococcus | PMID: 28174737 |

TABLE 6
Top 15 Potential Obesity-Related Microbes Predicted by BPNNHMDA and 13 Out of These 15 Microbes Have Been Supported by Literature Evidences

| Rank | Microbe | Evidence |
|---|---|---|
| 1 | Proteobacteria | PMID: 25407880 |
| 2 | Prevotella | PMID: 28521862 |
| 3 | Helicobacter pylori | PMID: 26409735 |
| 4 | Actinobacteria | Unconfirmed |
| 5 | Haemophilus | PMID: 2788987 |
| 6 | Clostridium coccoides | PMID: 23147032 |
| 7 | Lactobacillus | PMID: 28555008 |
| 8 | Lachnospiraceae | PMID: 28286339 |
| 9 | Clostridium difficile | Unconfirmed |
| 10 | Enterobacteriaceae | PMID: 28916564 |
| 11 | Veillonella | PMID: 30765893 |
| 12 | Bacteroides | PMID: 28628112 |
| 13 | Clostridia | PMID: 25271283 |
| 14 | Staphylococcus | PMID: 29576948 |
| 15 | Faecalibacterium prausnitzii | PMID: 19849869 |

We verified the top 15 candidate IBD related microbes predicted by BPNNHMDA, and as a result, all of them have been proved to have different degrees of association with IBD, and the detailed evidences are illustrated in the following Table 4. From observing the Table 4, it is easy to see that comparing with the microbiome of healthy people, the Firmicutes (Ranking third in the prediction list) in the microbiome of IBD patients would increase significantly, whereas Bacteroidetes (Ranking first in the prediction list) would decrease on the contrary [57]. In another experiment, Prevotella (Ranking second in the prediction list), Haemophilus (Ranking fifth in the prediction list), and Veillonella (Ranking eleventh in the prediction list) were found to be largely responsible for dysbacteriosis of the salivary microbiome in IBD patients [58]. Additionally, Sonnenberg A *et al.* confirmed that there exists negative correlation between Helicobacter pylori (Ranking fourth in the prediction list) and IBD [59].

Second, asthma is defined as a real public health problem affecting the world and population [60]. Now there are more than 300 million people worldwide suffering from asthma, with approximately 180,000 deaths per year, and it is conservatively estimated that there will be more 100 million asthma patients by 2025 [60], [61]. Huang and Boushey [62] believed that the pathogenesis of asthma is closely related to the microbiota, so we chose asthma as a case study and listed the top 15 potential asthma related microbes predicted by BPNNHMDA in the following Table 5. From observing the Table 5, it is easy to see that there are 14 out of these 15 microbes having been documented in association with the onset of asthma. For instance, both Firmicutes (Ranking first in the prediction list) and Actinobacteria (Ranking second in the prediction list) were found more frequently in sputum samples from non-asthmatic patients than from asthmatic patients [63]. Meanwhile, Lactobacillus (Ranking fourth in the prediction list) was found to be highly abundant in asthma patients [64]. By contrast, Streptococcus (Ranking 13th in the prediction list) and Veillonella (Ranking 9th in the prediction list) were found to be dominant in the oropharynx of healthy people [65]. In the experiment of Wang *et al.*, they found Faecalibacterium prausnitzii (Ranking 12th in the prediction list) was enriched in the gut of healthy people, but that was depleted in asthma cases [66].

TABLE 5
Top 15 Potential Asthma-Related Microbes Predicted by BPNNHMDA and 14 Out of These 15 Microbes Have Been Supported by Literature Evidences

| Rank | Microbe | Evidence |
|---|---|---|
| 1 | Firmicutes | PMID:23265859 |
| 2 | Actinobacteria | PMID:23265859 |
| 3 | Clostridium coccoides | PMID: 21477358 |
| 4 | Lactobacillus | PMID:20592920 |
| 5 | Lachnospiraceae | PMID:26512904 |
| 6 | Staphylococcus aureus | PMID:17950502 |
| 7 | Clostridium difficile | PMID:21872915 |
| 8 | Enterobacteriaceae | PMID:28947029 |
| 9 | Veillonella | PMID:25329665 |
| 10 | Clostridia | PMID:22047069 |
| 11 | Bacteroides | PMID: 18822123 |
| 12 | Faecalibacterium prausnitzii | PMID: 30208875 |
| 13 | Streptococcus | PMID: 25329665,17950502 |
| 14 | Fusobacterium | [63] |
| 15 | Bacteroides vulgatus | Unconfirmed |

Finally, according to statistics, there are currently more than 1.9 billion people obese or overweight in the world. The total prevalence of childhood obesity is 5.0 percent, and the adult prevalence rate is as high as 12.0 percent. [67], [68]. Obesity is more likely to cause health complications such as insulin resistance, type 2 diabetes, cardiovascular disease, liver disease, cancer, and neurodegeneration [68]. As illustrated in the following Table 6, out of the top 15 potential obesity-related microbes predicted by BPNNHMDA, there are 13 microbes having been proved to be associated with obesity. For instance, Park *et al.* [69] analyzed the gut microflora in beagle dogs, and Proteobacteria (Ranking first in the prediction list) was found to dominate the gut microbiota of dogs in the obese group. And additionally, Helicobacter pylori (Ranking third in the prediction list), which infects the human stomach causing H. pylori infection, was found to be significantly and positively associated with obesity in China as well [70].

Moreover, case studies of these three kinds of diseases are also implemented on those competing methods listed in

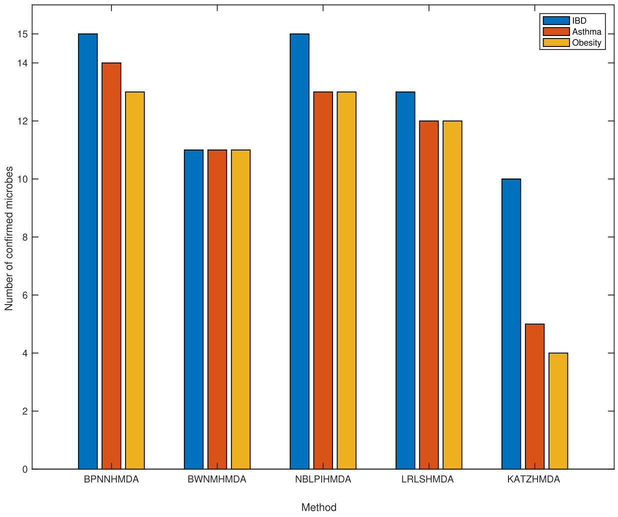


Fig. 7. Comparison of case studies between BPNNHMDA and competing methods including BWNMHMDA, NBLPIHMDA, LRLSHMDA and KATZHMDA.

above Section 4.3. And as shown in the following Fig. 7, it is obvious that among the top 15 potential IBD-related, asthma-related and obesity-related microbes predicted by BPNNHMDA, there are 15, 14 and 13 microbes having been confirmed to be associated with IBD, asthma, and obesity respectively, which are superior to BWNMHMDA (11, 11 and 11), NBLPIHMDA (15, 13 and 13), LRLSHMDA (13, 12 and 12) and KATZHMDA (10, 5 and 4). Apparently, comparative results of case studies reconfirmed the excellent prediction performance of BPNNHMDA. In addition, for more details about the top 15 potential microbes predicted by above state-of-the-art competing methods and the literature evidences corresponding to them, please see Supplementary Table S2, available online. And furthermore, we published the rankings of all potential microbes predicted by BPNNHMDA in the Supplementary Table S3, available online, as well, and hoped that these associations might provide some help for the future researches of relevant researchers.

## 6 Conclusion

Microorganisms are abundant in the human body and form an interdependent relationship with the host. Numerous studies have indicated that microbial activity is involved in human life, while the dysbiosis of human microbiome will lead to human diseases. Hence, studying the complex relationships between microbes and human diseases can better understand the pathology of the disease as well as the prevention and treatment of human diseases. Therefore, microbe-disease researches have become a hot topic in bioinformatics, microbiology and medicine in recent years. However, traditional experimental methods are time-consuming, expensive and blind. Therefore, exploring potential microbe-disease associations by computation methods is becoming more and more necessary and important.

In this article, a novel approach called BPNNHMDA was proposed to identify potential microbe-disease associations based on an improved BPNN. In BPNNHMDA, we first designed a unique three-layer neural network structure, in which, known microbe-disease associations were taken as the input signals. Next, we designed a new activation function to activate the hidden layer and the output layer in the improved BPNN based on the hyperbolic tangent function. Besides, we optimized initial connection weights of the improved BPNN by adopting GIP similarity for microbes. After that, the connection weights and biases in the improved BPNN would be updated iteratively according to the back propagation learning algorithm until the whole network had reached convergent state. Finally, in order to verify the predictive performance of BPNNHMDA, we implemented LOOCV, 5-Fold CV and 2-Fold CV on the dataset downloaded from the HMDAD database. Simulation results illustrated that BPNNHMDA could achieve better performance than state-of-the-art methods. Moreover, results of the case study demonstrated the excellent prediction performance of BPNNHMDA in practical applications as well. Through analysis, the major reasons that BPNNHMDA could achieve reliable prediction performance might be due to its following characteristics: (1) GIP similarity was used to obtain the optimal initial connection weights of BPNN, which improved the efficiency of training process. (2) The input signals of BPNNHMDA were normalized across channels, and in addition, input data, connection weights and biases were all strictly normalized during the training process, which guaranteed the prediction performance of BPNNHMDA apparently. (3) The improved activation function based on Tanh function was adopted to our BPNN model, which made up for the deficiency of traditional activation function and ensured the validity of data activation in BPNNHMDA.

As far as we know, in this manuscript, BPNN is adopted for the first time to infer potential microbe-disease associations. Thus, there are certainly some shortcomings in BPNNHMDA. For example, the GIP similarity was used to optimize initial connection weights of BPNNHMDA, although it could improve the training process, but it fixed the convergent direction of the neural network as well, which mean that some poorly optimized weights may not be conducive to effective prediction of some potential microbe-disease associations. In addition, the learning rate of BPNN was fixed in BPNNHMDA, which was not conducive to the training process either. Conservative and small learning rate would reduce the training speed, while large learning rate may make the neural network miss the optimal error. Therefore, it is necessary to improve the adaptive change ability of learning rate to reduce the training time and improve the prediction performance of BPNNHMDA in future works.

## Acknowledgments

The authors thank the anonymous referees for suggestions that helped improve the paper substantially. The project is partly sponsored by the National Natural Science Foundation of China (No.61873221, No.61672447), the Natural Science Foundation of Hunan Province (No.2018JJ4058, No.2019JJ70010), and the Hunan province science and technology project Fund (2018TP1036).

## References

[1] S. R. Gill *et al.*, “Metagenomic analysis of the human distal gut microbiome,” *Science*, vol. 312, no. 5778, pp. 1355–1359, 2006.

[2] J. A. Gilbert and C. L. Dupont, “Microbial metagenomics: Beyond the genome,” *Annu. Rev. Marine Sci.*, vol. 3, no. 1, pp. 347–371, 2011.

[3] R. Sender, S. Fuchs, and R. Milo, "Revised estimates for the number of human and bacteria cells in the body," *PLOS Biol.*, vol. 14, no. 8, pp. 1–14, 2016.

[4] I. Sekirov, S. L. Russell, L. C. M. Antunes, and B. B. Finlay, "Gut microbiota in health and disease," *Physiol. Rev.*, vol. 90, no. 3, pp. 859–904, 2010.

[5] P. Barko, M. McMichael, K. Swanson, and D. Williams, "The gastrointestinal microbiome: A review," *J. Veterinary Intern. Med.*, vol. 32, no. 1, pp. 9–25, 2017.

[6] K. M. Maslowski, "Metabolism at the centre of the host–microbe relationship," *Clin. Exp. Immunol.*, vol. 197, no. 2, pp. 193–204, 2019.

[7] K. M. Maslowski and C. R. Mackay, "Diet, gut microbiota and immune responses," *Nat. Immunol.*, vol. 12, no. 1, pp. 5–9, 2010.

[8] M. R. Mason, P. M. Preshaw, H. N. Nagaraja, S. M. Dabdoub, A. Rahman, and P. S. Kumar, "The subgingival microbiome of clinically healthy current and never smokers," *The ISME J.*, vol. 9, no. 1, pp. 268–272, 2014.

[9] X. Li, K. Watanabe, and I. Kimura, "Gut microbiota dysbiosis drives and implies novel therapeutic strategies for diabetes mellitus and related metabolic diseases," *Front. Immunol.*, vol. 8, 2017, Art. no. 1882.

[10] M. Scriven, T. Dinan, J. Cryan, and M. Wall, "Neuropsychiatric disorders: Influence of gut microbe to brain signalling," *Diseases*, vol. 6, no. 3, 2018, Art. no. 78.

[11] M. Valles-Colomer *et al.*, "The neuroactive potential of the human gut microbiota in quality of life and depression," *Nat. Microbiol.*, vol. 4, no. 4, pp. 623–632, 2019.

[12] N. Kim, M. Yun, Y. J. Oh, and H.-J. Choi, "Mind-altering with the gut: Modulation of the gut-brain axis with probiotics," *J. Microbiol.*, vol. 56, no. 3, pp. 172–182, 2018.

[13] J. Kelsen and G. D. Wu, "The gut microbiota and IBD," in *Pediatric Inflammatory Bowel Disease*. New York, NY, USA: Springer, 2012, pp. 35–42.

[14] E. M. Quigley, "Review: Do patients with functional gastrointestinal disorders have an altered gut flora?" *Ther. Advances Gastroenterol.*, vol. 2, no. 4_suppl, pp. S23–S30, 2009.

[15] H. Zhang *et al.*, "Human gut microbiota in obesity and after gastric bypass," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 7, pp. 2365–2370, 2009.

[16] J. F. Cryan and T. G. Dinan, "Mind-altering microorganisms: The impact of the gut microbiota on brain and behaviour," *Nat. Rev. Neurosci.*, vol. 13, no. 10, pp. 701–712, 2012.

[17] L. Desbonnet, L. Garrett, G. Clarke, B. Kiely, J. Cryan, and T. Dinan, "Effects of the probiotic bifidobacterium infantis in the maternal separation model of depression," *Neuroscience*, vol. 170, no. 4, pp. 1179–1188, 2010.

[18] D. J. Davis *et al.*, "Lactobacillus plantarum attenuates anxiety-related behavior and protects against stress-induced dysbiosis in adult zebrafish," *Sci. Rep.*, vol. 6, no. 1, 2016, Art. no. 33726.

[19] A. A. Althani *et al.*, "Human microbiome and its association with health and diseases," *J. Cellular Physiol.*, vol. 231, no. 8, pp. 1688–1694, 2016.

[20] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "The human microbiome project," *Nature*, vol. 449, no. 7164, pp. 804–810, 2007.

[21] J. A. Gilbert *et al.*, "Meeting report: The terabase metagenomics workshop and the vision of an earth microbiome project," *Standards Genomic Sci.*, vol. 3, no. 3, pp. 243–248, 2010.

[22] J. A. Gilbert, J. K. Jansson, and R. Knight, "The earth microbiome project: Successes and aspirations," *BMC Biol.*, vol. 12, no. 1, 2014, Art. no. 69.

[23] W. Ma *et al.*, "An analysis of human microbe–disease associations," *Brief. Bioinformatics*, vol. 18, no. 1, pp. 85–97, 2016.

[24] P. Ping, L. Wang, L. Kuang, S. Ye, M. F. B. Iqbal, and T. Pei, "A novel method for LncRNA-disease association prediction based on an lncRNA-disease association network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 2, pp. 688–693, Mar./Apr. 2019.

[25] J. Li *et al.*, "A novel approach for potential human LncRNA-disease association prediction based on local random walk," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2019.2934958.

[26] X. Chen, Y.-Z. Sun, H. Liu, L. Zhang, J.-Q. Li, and J. Meng, "RNA methylation and diseases: Experimental results, databases, web servers and computational models," *Brief. Bioinformatics*, vol. 20, no. 3, pp. 896–917, 2019.

[27] X. Chen and G.-Y. Yan, "Novel human lncRNA-disease association inference based on lncRNA expression profiles," *Bioinformatics*, vol. 29, no. 20, pp. 2617–2624, Oct. 2013.

[28] X. Chen, C. C. Yan, X. Zhang, and Z.-H. You, "Long non-coding RNAs and complex diseases: From experimental results to computational models," *Brief. Bioinformatics*, vol. 18, no. 4, pp. 558–576, 2017.

[29] H. Zhao *et al.*, "Prediction of microRNA-disease associations based on distance correlation set," *Bioinformatics*, vol. 19, no. 1, 2018, Art. no. 141.

[30] X. Chen, C.-C. Zhu, and J. Yin, "Ensemble of decision tree reveals potential miRNA-disease associations," *PLoS Comput. Biol.*, vol. 15, no. 7, 2019, Art. no. e1007209.

[31] X. Chen, L. Wang, J. Qu, N.-N. Guan, and J.-Q. Li, "Predicting miRNA–disease association based on inductive matrix completion," *Bioinformatics*, vol. 34, no. 24, pp. 4256–4265, 2018.

[32] X. Chen, D. Xie, Q. Zhao, and Z.-H. You, "MicroRNAs and complex diseases: From experimental results to computational models," *Brief. Bioinformatics*, vol. 20, no. 2, 2019, Art. no. e1006418.

[33] X. Chen, J. Yin, J. Qu, and L. Huang, "MDHGI: Matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction," *PLoS Comput. Biol.*, vol. 14, no. 8, pp. 1–24, 2018.

[34] X. Chen *et al.*, "Drug–target interaction prediction: Databases, web servers and computational models," *Brief. Bioinformatics*, vol. 17, no. 4, pp. 696–712, 2016.

[35] X. Chen, B. Ren, M. Chen, Q. Wang, L. Zhang, and G. Yan, "NLLSS: Predicting synergistic drug combinations based on semi-supervised learning," *PLoS Comput. Biol.*, vol. 12, no. 7, 2016, Art. no. e1004975.

[36] X. Chen, Y.-A. Huang, Z.-H. You, G.-Y. Yan, and X.-S. Wang, "A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases," *Bioinformatics*, vol. 33, no. 5, pp. 733–739, 2017.

[37] Y.-A. Huang, Z.-H. You, X. Chen, Z.-A. Huang, S. Zhang, and G.-Y. Yan, "Prediction of microbe–disease association from the integration of neighbor and graph with collaborative recommendation model," *J. Translational Med.*, vol. 15, no. 1, 2017, Art. no. 209.

[38] Z.-A. Huang *et al.*, "PBHMDA: Path-based human microbe-disease association prediction," *Front. Microbiology*, vol. 8, 2017, Art. no. 233.

[39] H. Li *et al.*, "A novel human microbe-disease association prediction method based on the bidirectional weighted network," *Front. Microbiol.*, vol. 10, 2019, Art. no. 676.

[40] X. Shen, Y. Chen, X. Jiang, X. Hu, T. He, and J. Yang, "Predicting disease-microbe association by random walking on the heterogeneous network," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2016, pp. 771–774.

[41] L. Wang, Y. Wang, H. Li, X. Feng, D. Yuan, and J. Yang, "A bidirectional label propagation based computational model for potential microbe-disease association prediction," *Front. Microbiol.*, vol. 10, 2019, Art. no. 684.

[42] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Boca Raton, FL, USA: CRC Press, 2018.

[43] M. Cilimkovic, "Neural networks and back propagation algorithm," *Institute of Technology Blanchardstown, Blanchardstown Road North Dublin*, vol. 15, 2015.

[44] D. Rumelhart, G. Hinton, and R. Williams, "Learning international representations by error propagation, parallel distributed processing: Explorations in the microstructures of cognition," 1986.

[45] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks:," *Int. J. Forecast.*, vol. 14, no. 1, pp. 35–62, 1998.

[46] A. H. Moghaddam, M. H. Moghaddam, and M. Esfandyari, "Stock market index prediction using artificial neural network," *J. Econ. Finance Administ. Sci.*, vol. 21, no. 41, pp. 89–93, 2016.

[47] L. Wang, Y. Zeng, and T. Chen, "Back propagation neural network with adaptive differential evolution algorithm for time series forecasting," *Expert Syst. Appl.*, vol. 42, no. 2, pp. 855–863, 2015.

[48] C.-H. Hsu, G. Manogaran, P. Panchatcharam, and S. Vivekanandan, "A new approach for prediction of lung carcinoma using back propagation neural network with decision tree classifiers," in *Proc. IEEE 8th Int. Symp. Cloud Service Comput.*, 2018, pp. 111–115.

[49] J. Li, J. H. Cheng, J. Y. Shi, and F. Huang, "Brief introduction of back propagation (BP) neural network algorithm and its improvement," in *Proc. Int. Conf. Intell. Soft Comput.*, 2012, pp. 553–558.

[50] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[51] C. C. Klimasauskas, "Applying neural networks. Part 3: Training a neural network," in *Proc. Artif. Intell.*, 1991, pp. 20–24.

[52] N. Meade, "Neural network time series forecasting of financial markets," *Int. J. Forecast.*, vol. 11, no. 4, pp. 601–602, 1995.

[53] C. Wu, R. Gao, D. Zhang, S. Han, and Y. Zhang, "PRWHMDA: Human microbe-disease association prediction by random walk on the heterogeneous network with PSO," *Int. J. Biol. Sci.*, vol. 14, no. 8, pp. 849–857, 2018.

[54] F. Wang *et al.*, "LRLSHMDA: Laplacian regularized least squares for human microbe–disease association prediction," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 7061.

[55] J. Burisch, T. Jess, M. Martinato, and P. L. Lakatos, "The burden of inflammatory bowel disease in europe," *J. Crohn's Colitis*, vol. 7, no. 4, pp. 322–337, 2013.

[56] Y.-Z. Zhang, "Inflammatory bowel disease: Pathogenesis," *World J. Gastroenterol.*, vol. 20, no. 1, 2014, Art. no. 91.

[57] M. L. Santoru *et al.*, "Cross sectional evaluation of the gut-microbiome metabolome axis in an italian cohort of IBD patients," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 9523.

[58] H. S. Said *et al.*, "Dysbiosis of salivary microbiota in inflammatory bowel disease and its association with oral immunological biomarkers," *DNA Res.*, vol. 21, no. 1, pp. 15–25, 2013.

[59] A. Sonnenberg and R. M. Genta, "Low prevalence of helicobacter pylori infection among patients with inflammatory bowel disease," *Alimentary Pharmacol. Ther.*, vol. 35, no. 4, pp. 469–476, 2012.

[60] C. Nunes, A. M. Pereira, and M. Morais-Almeida , "Asthma costs and social impact," *Asthma Res. Pract.*, vol. 3, 2017, Art. no. 1.

[61] M. Masoli, D. Fabian, S. Holt, and R. B. and, "The global burden of asthma: Executive summary of the GINA dissemination committee report," *Allergy*, vol. 59, no. 5, pp. 469–478, 2004.

[62] Y. J. Huang and H. A. Boushey, "The microbiome in asthma," *J. Allergy Clin. Immunol.*, vol. 135, no. 1, pp. 25–30, 2015.

[63] H. T. Dang, S. A. Kim, H. K. Park, J. W. Shin, S.-G. Park, and W. Kim, "Analysis of oropharyngeal microbiota between the patients with bronchial asthma and the non-asthmatic persons," *J. Bacteriol. Virol.*, vol. 43, no. 4, 2013, Art. no. 270.

[64] J. Yu *et al.*, "The effects of lactobacillus rhamnosus on the prevention of asthma in a murine model," *Allergy Asthma Immunol. Res.*, vol. 2, no. 3, pp. 199–205, 2010.

[65] H. Park, J. W. Shin, S.-G. Park, and W. Kim, "Microbial communities in the upper respiratory tract of patients with asthma and chronic obstructive pulmonary disease," *PLoS One*, vol. 9, no. 10, 2014, Art. no. e109710.

[66] Q. Wang *et al.*, "A metagenome-wide association study of gut microbiota in asthma in UK adults," *BMC Microbiol.*, vol. 18, no. 1, 2018, Art. no. 114.

[67] G. O. Collaborators, "Health effects of overweight and obesity in 195 countries over 25 years," *N. Engl. J. Med.*, vol. 377, no. 1, pp. 13–27, 2017.

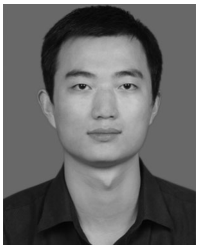
[68] A. R. Saltiel and J. M. Olefsky, "Inflammatory mechanisms linking obesity and metabolic disease," *J. Clin. Invest.*, vol. 127, no. 1, pp. 1–4, 2017.

[69] H.-J. Park, S.-E. Lee, H.-B. Kim, R. Isaacson, K.-W. Seo, and K.-H. Song, "Association of obesity with serum leptin, adiponectin, and serotonin and gut microflora in beagle dogs," *J. Vet. Intern. Med.*, vol. 29, no. 1, pp. 43–50, 2014.

[70] Y. Zhang, T. Du, X. Chen, X. Yu, L. Tu, and C. Zhang, "Association between helicobacter pylori infection and overweight or obesity in a chinese population," *The J. Infection Developing Countries*, vol. 9, no. 09, pp. 945–953, 2015.

**Hao Li** is currently working toward the postgraduate degree in the College of Information Engineering, Xiangtan University, China. His current research interest is bioinformatics.

**Yuqi Wang** is currently working toward the postgraduate degree in the College of Information Engineering, Xiangtan University, China. Her current research interest is bioinformatics.



**Zhen Zhang** is currently a associate professor of College of Electronic Information and Electrical Engineering in Changsha University, China. His current research interest include bioinformatics.

**Yihong Tan** is a professor of College of Computer Engineering and Applied Mathematics of Changsha University. His research focuses on machine learning.

**Zhiping Chen** is a professor of College of Computer Engineering and Applied Mathematics of Changsha University. His research focuses on machine learning.

**Xiangyi Wang** is currently working toward the undergraduate degree in the College of Computer Engineering and Applied Mathematics, Changsha University, China. Her current research interest is bioinformatics.

**Tingrui Pei** received the PhD degree in signal and information processing from the Beijing University of Posts and Telecommunications, P.R. China, in 2004. From 2006 to 2007, he moved to Japan as a visiting scholar in Waseda University, Tokyo, Japan. Currently, he is a full professor in Xiangtan University, P.R.China. His main research areas include bioinformatics and Internet of Things.

**Lei Wang** received the PhD degree in computer science from Hunan University, P.R.China, in 2005. From 2005 to 2007, he was a postdoctoral fellow in Tsinghua University, P.R.China. After that, he moved to USA and Canada as a visiting scholar in Duke University, Durham, North Carolina and Lakehead University, Thunder Bay, Canada successively. From 2005 to 2011, he had been an associate professor with the College of Software in Hunan University, Changsha, China. From 2011 to 2018, he had been a full professor with the College of Information Engineering, Xiangtan University, Xiangtan, China. Currently, he is a full professor and an academic leader of Computer Engineering in Changsha University, P.R.China. He has published more than 100 peer-reviewed articles. His main research areas include bioinformatics and Internet of Things.