OXFORD

Briefings in Bioinformatics, 2023, **24(4)**, 1–14

https://doi.org/10.1093/bib/bbad255

Advance access publication date 5 July 2023

Problem Solving Protocol

Predicting potential microbe–disease associations based on multi-source features and deep learning

Liugen Wang, Yan Wang, Chenxu Xuan, Bai Zhang, Hanwen Wu and Jie Gao

Corresponding author. Jie Gao, School of Science, Jiangnan University, Wuxi, Jiangsu 214122, China. Tel.: +86-510-85910532; E-mail: gaojie@jiangnan.edu.cn

Abstract

Studies have confirmed that the occurrence of many complex diseases in the human body is closely related to the microbial community, and microbes can affect tumorigenesis and metastasis by regulating the tumor microenvironment. However, there are still large gaps in the clinical observation of the microbiota in disease. Although biological experiments are accurate in identifying disease-associated microbes, they are also time-consuming and expensive. The computational models for effective identification of diseases related microbes can shorten this process, and reduce capital and time costs. Based on this, in the paper, a model named DSAE_RF is presented to predict latent microbe–disease associations by combining multi-source features and deep learning. DSAE_RF calculates four similarities between microbes and diseases, which are then used as feature vectors for the disease-microbe pairs. Later, reliable negative samples are screened by *k*-means clustering, and a deep sparse autoencoder neural network is further used to extract effective features of the disease-microbe pairs. In this foundation, a random forest classifier is presented to predict the associations between microbes and diseases. To assess the performance of the model in this paper, 10-fold cross-validation is implemented on the same dataset. As a result, the AUC and AUPR of the model are 0.9448 and 0.9431, respectively. Furthermore, we also conduct a variety of experiments, including comparison of negative sample selection methods, comparison with different models and classifiers, Kolmogorov–Smirnov test and *t*-test, ablation experiments, robustness analysis, and case studies on Covid-19 and colorectal cancer. The results fully demonstrate the reliability and availability of our model.

Keywords: microbe–disease associations, *k*-means clustering, deep sparse autoencoder neural network, random forest

INTRODUCTION

Microbial communities consisting of bacteria, archaea, fungi, viruses and protozoa commonly colonize the human body [1]. Accumulated research has shown that the microbial communities in the human body are strongly associated with the occurrence of diseases. And the imbalance or dysbiosis of microbial ecosystems will disturb the transcription, translation and DNA repair mechanisms of the host [2]. For instance, the gut microbiota may be one of the main factors in multiple sclerosis, which is an autoimmune disease of the central nervous system [3]. The gut microbiome can also be used to enhance the effectiveness of glioma treatments [4]. These studies suggest that the gut microbiome represents a new and relatively unexplored field as a potential predictive biomarker for some diseases.

In addition, oral microbes are also one of the main communities in human microbial communities, and many oral diseases are based on microbes [5]. Experiments have found that oral microbial transplantation might be a novel option for the treatment of periodontitis [6]. *Porphyromonas gingivalis*, an anaerobic bacterium in the oral microbiota, may be involved in three distinct stages of oral squamous cell carcinoma [7].

As described above, the occurrence of many complex diseases in the human body is closely related to the microbial community, and microbes can also affect tumorigenesis and metastasis by regulating the tumor microenvironment (TME). If disease-associated microbes can be identified, precision medicine and cancer treatment may be possible. In summary, the identification of latent microbe–disease associations will have a longer-term rationale and practical implications, not only to better develop an understanding of the mechanisms of disease formulation and progression, but also to provide new medical solutions to diseases. However, the reported microbe–disease associations do not allow us to fully understand the mechanisms of disease pathogenesis. Therefore, some computational approaches to predict potential microbe–disease correlations have been widely adopted. At present, these forecasting models are mainly divided into the following categories: random walk-based methods, path-based methods, matrix factorization-based methods and artificial intelligence-based methods, including machine learning and deep learning.

Random walk-based methods have been widely used in bioinformatics, such as lncRNA (miRNA)-disease prediction. Currently, Zou *et al.* proposed a new method to predict microbe–disease relationships through a bi-random walk on the heterogeneous network [8], the model combined the known microbe–disease associations and Gaussian interaction profile kernel (GIP) similarity to establish biological networks, and then obtained prediction scores for microbe and disease networks by random walk,

Liugen Wang is with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, Jiangsu 214122, China
Yan Wang, Chenxu Xuan, Bai Zhang, Hanwen Wu, and Jie Gao are with the School of Science, Jiangnan University, Wuxi, Jiangsu 214122, China

Received: April 3, 2023. **Revised:** May 30, 2023. **Accepted:** June 20, 2023

ultimately integrating these two scores as microbe–disease association probabilities. Peng *et al*. developed a framework to predict potential microbe–disease correlations based on multiple features and bi-random walks [9], the model combined non-negative matrix decomposition, neural network and random walk to obtain the characteristic representation of microbes and diseases, then input the characteristics into the logistic regression (LR) model to obtain the microbe–disease association score. Wang *et al*. adopted an embedding algorithm to predict latent microbe–disease correlations, which obtained association scores by random walk and LINE methods after computing multiple similarities [10].

Path-based methods typically take the weights of potential paths as scores of unknown associations by considering indirect paths across a graph or network. Chen *et al*. predicted the latent correlations between microbes and diseases through the calculation of the KATZ measure after calculating the GIP similarity between microbes and diseases [11]. Since the existing microbe–disease correlation matrix is severely sparse, Li *et al*. used the bipartite network recommendation to reconstruct the correlation matrix, and then predicted the diseases-associated microbes through the KATZ measure [12]. Similarly, considering the sparsity of the network, Yin *et al*. calculated the functional similarity through the 16S rRNA sequence information of microbes and combined it with the GIP similarity of microbes. For disease similarity, they integrated semantic and GIP similarity. Yin *et al*. calculated the functional similarity through the 16S rRNA sequence information of microbes and combined it with the GIP similarity of microbes. For disease similarity, they integrated semantic and GIP similarity. The projection scores of microbe and disease spaces were calculated based on the network consistency projection, and, finally, the latent microbes of diseases were identified by the label propagation algorithm [13].

Matrix factorization methods also play a key role in predicting underlying microbe–disease relationships. Yang *et al*. designed a novel approach to predict latent microbe–disease correlations using bilinear matrix factorization after computing microbial GIP and cosine similarity, disease GIP, cosine and phenotypic similarity [14]. Similarly, Xu *et al*. combined the functional similarity of microbes with GIP similarity, and the phenotype similarity of diseases with GIP similarity. They presented a model of cooperative weighted nonnegative matrix factorization with graph Laplacian normalization to predict latent microbe–disease correlations by introducing a Tikhonov (L_2) regularization term [15].

Recently, machine learning and deep learning have also been applied to related microbial work for disease prediction. Based on the known network of microbe–disease correlations, an LRLSHMDA model [16] and an Adaptive Boosting model named ABHMDA [17] were employed to predict the underlying microbe–disease correlations. Among them, the LRLSHMDA model constructed classifiers in the microbe and disease space, respectively, and then integrated the scores of the two classifiers as the final microbe–disease prediction probability. Different from LRLSHMDA, the ABHMDA model selected reliable negative samples through *k*-means clustering, and then integrated multiple weak classifiers to predict the association probability of strong classifiers. In addition, BPNNHMDA, a calculative model based on an improved back-propagation neural network, was proposed to recognize latent microbe–disease associations, for which they designed a new activation function and optimized the initial connection weights by the GIP similarity of microbes, and set the same number of neurons in the input layer, hidden layer and output layer of the neural network to obtain better prediction results [18].

These methods have been used with some success in identifying microbe–disease associations. However, as mentioned above, they have various disadvantages. Some models only use the similarity of individual microbes or diseases, which cannot well characterize the features of microbes or diseases. When selecting training samples, some models randomly select a certain number of negative samples, which is also inaccurate. At the same time, substantial literature gaps remain regarding the clinical observation of the microbiota in disease. Based on this background, we propose the DSAE RF model, hoping to quickly identify potential microbe–disease associations, and effectively shorten the time of biological experiments. At the same time, it also contributes to the diagnosis, treatment, and prognosis of diseases, and provides new insights for precision medicine and cancer treatment.

DSAE_RF model combines deep sparse autoencoder neural network (DSAE) and random forest (RF). We calculate four similarities between microbes and diseases, which are then used as features of disease-microbe pairs. In this study, we regard the known microbe–disease correlations as positive samples and otherwise as negative samples. Then, after selecting negative samples with an equal quantity of positive samples by *k*-means clustering, later, the positive and negative samples are input into the DSAE as training samples to extract features. In addition, we introduce RF as the final classifier model because it can work effectively on large datasets and have high training speed and accuracy.

As a result, the AUC and AUPR of DSAE_RF are 0.9448 and 0.9431 in 10-fold cross-validation, respectively, which indicates that the prediction performance of the model can be effective. Simultaneously, ablation studies, comparison of negative sample selection methods, comparisons with recently published models, robustness analysis, statistical tests, comparisons with other classifiers, and case studies on Covid-19 and colorectal cancer are listed to further validate the predictive ability and reliability of DSAE_RF. In addition, the results will contribute to the insight into the development of complex diseases and will promote the development of relevant drugs for disease prevention, diagnosis and treatment.

Overall, our main contributions are concluded as follows:

(i) We have constructed and integrated four similarity networks of microbes and diseases in order to better predict the relevant microbes of diseases. Experiments have confirmed the effectiveness of integrating different networks.

(ii) We select reliable negative samples by *k*-means clustering, which can eliminate the problem of sample imbalance to a certain extent. Meanwhile, experimental results have confirmed that reliable negative sample selection can improve model performance.

(iii) We extract effective features through neural networks, which can not only improve the performance of the model, but also effectively shorten the running time of the model.

(iv) We use RF to predict potential microbe–disease associations, the reliability of the model has been confirmed through experiments, indicating that our model can provide new insights into the diagnosis, treatment and prognosis of diseases, and can formulate new solutions for precision medicine.

RESULTS

Performance evaluation

In the work, we choose cross-validation to evaluate the performance of DSAE_RF. In order to increase the utilization of data,

Table 1. AUC, AUPR, Recall, Pre, Acc, F1-score of 10 experiments and their means

Test set	AUC (%)	AUPR (%)	Recall (%)	Pre (%)	ACC (%)	F1-score (%)
1	95.71	96.15	88.75	88.85	88.79	88.78
2	94.92	94.42	86.42	86.40	86.46	86.41
3	93.78	92.56	87.12	87.14	87.13	87.12
4	92.98	92.71	84.14	84.08	84.13	84.10
5	94.84	95.18	86.44	86.42	86.46	86.43
6	94.46	94.80	87.13	87.14	87.13	87.12
7	93.80	93.19	85.82	85.76	85.79	85.78
8	93.88	93.79	86.79	86.75	86.78	86.77
9	94.74	94.65	85.83	85.91	85.89	85.86
10	95.68	95.68	88.31	88.39	88.33	88.32
Average	94.48 ± 0.83	94.31 ± 1.16	86.67 ± 1.24	86.69 ± 1.28	86.69 ± 1.25	86.67 ± 1.26

at the same time, considering the computational cost of leave-one-out cross-validation (LOOCV) and repeated k-fold cross-validation, we finally adopt k-fold cross-validation. Furthermore, k-fold cross-validation can effectively reduce the contingency of the results and thus improve the accuracy of the model. For k-fold cross-validation, the selection range of k is usually from 2 to 10, among which the larger k is the more accurate estimation of generalization performance, so we finally choose $k = 10$. Then, we estimate the performance of our model using different categories of criteria, namely AUC, AUPR, Recall, precision (Pre), accuracy (ACC) and F1-score [19]. Among them, AUC and AUPR are the areas under the receiver operating characteristic curve (ROC) and precision-recall curve (PR). They can be computed as follows:

$$TPR = Recall = \frac{TP}{TP + FN}, \tag{1}$$

$$FPR = \frac{FP}{FP + TN}, \tag{2}$$

$$Pre = \frac{TP}{TP + FP}, \tag{3}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

$$F1 - score = \frac{2}{{}^{1}/_{Pre} + {}^{1}/_{Recall}} = \frac{2TP}{2TP+FP+FN}. \tag{5}$$

where TP, FP, TN, and FN stand for the amount of positive samples correctly predicted, the amount of positive samples falsely predicted, the amount of negative samples correctly predicted and the amount of negative samples false predicted in the model, respectively.

Based on 10-fold cross-validation, the AUC, AUPR, Recall, Pre, ACC, F1-score of the 10 experiments and their respective means are shown in Table 1. In addition, the ROC and PR curves of 10 experiments are also plotted in Figure. 1. The average results are as follows: AUC is 94.48%, AUPR is 94.31%, Recall is 86.67%, Pre is 86.69%, ACC is 86.69% and F1-score is 86.67%. Their standard deviations are 0.83, 1.16, 1.24, 1.28, 1.25 and 1.26%, respectively. The capability of DSAE_RF with high accuracy depends on the feature selection approach and classifier selection. Combined with the features extracted by the neural network, the RF classifier has better classification ability.

Comparison of negative sample selection methods

To confirm the reliability of our extracted negative samples, we compare the results with randomly selected negative samples. Furthermore, to ensure the accuracy of the experiment, 10 experiments are conducted with random sampling and the average of the 10 experiments is compared with our model, and the results are shown in Figure 2. From the results, it can be observed that the performance of the negative samples selected by k-means clustering is significantly improved, where the AUC and AUPR are improved by about 2%.

Ablation experiments

In this paper, a set of ablation experiments are implemented to validate the contribution of different species similarities. The similarities are divided into four groups, namely Group A ($DS + F\_MS$), Group B ($GIP\_MS + GIP\_DS$), Group C ($COS\_MS + COS\_DS$) and Group D ($SIG\_MS + SIG\_DS$). The training strategy is the same as our final model. The experiment results are shown in Table 2.

It can be observed that the properties of the model decrease by about 6% if only Group A similarities, i.e., diseases semantic similarity and microbes' functional similarity. The results suggest that combining different kinds of topological features of microbes and diseases will help to increase the predictive properties of the model. In that case, the topological features of microbes and diseases can better identify potentially relevant microbes for disease prevention, treatment and prognosis.

Furthermore, compared with our model, the AUC and AUPR decrease by 0.01 and 0.46%, respectively, in the absence of Group A, decrease by 1.4 and 1.32%, respectively, in the absence of Group B, decrease by 1.07 and 1.75%, respectively, in the absence of Group C, and decrease by 1.28 and 1.75%, respectively, in the absence of Group D. Although the performance of evaluation metrics (e.g., ACC) of our model is slightly lower than that of this experiment without considering Group A, we still believe that the absence of Group A will have an influence on the properties of our model, considering the more robust performance of AUC and AUPR. It is concluded that the ablation studies demonstrate the fundamental and significant contributions of the four groups of similarity to our model.

Comparison among different classifiers

To confirm the effectiveness of the features extracted by DSAE, we compare the features processed by DSAE with those not processed by DSAE. In this paper, commonly used basic classifiers such as the support vector machine (SVM), LR, Naive Bayes (NB), Decision Tree (DT), Gradient Boosting Decision Tree (GBDT) and K Nearest Neighbors (KNN) are compared. The results are shown in Figure 3.

From the results in Figure 3, it can be observed that the prediction accuracy is remarkably improved after DSAE extracts features with the same parameters and classifiers, which is

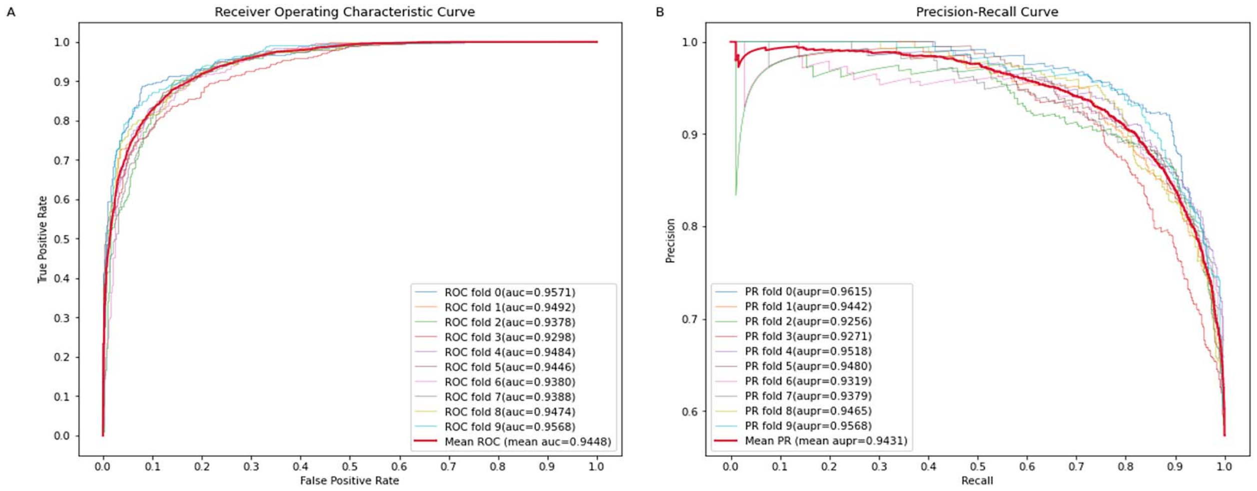

Figure 1. (A) ROC curves and its average curve of 10 experiments. (B) PR curves and its average curve of 10 experiments.

Table 2. Results of ablation experiments on our model

Group				AUC (%)	AUPR (%)	Recall (%)	Pre (%)	ACC (%)	F1-score (%)
A	B	C	D						
√				89.15	88.39	79.80	79.84	79.78	79.76
√	√	√	×	93.20	92.56	85.17	85.16	85.17	85.15
√	√	×	√	93.41	92.56	86.04	86.07	86.03	86.01
√	×	√	√	93.08	92.99	84.19	84.18	84.19	84.17
×	√	√	√	94.47	93.85	87.24	87.29	87.24	87.22
√	√	√	√	94.48	94.31	86.67	86.69	86.69	86.67

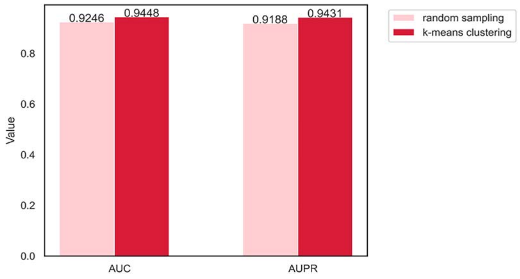

Figure 2. Comparison of different negative sample sampling methods.

most obvious in the NB classifier, where the AUC and AUPR are improved by about 3% after DSAE extracts features. The results show that DSAE can effectively mine hidden information and extract more effective features. Furthermore, to better illustrate the effectiveness of DSAE in learning features, we generate a feature space visualization based on DSAE and the original feature representation in 2D space using the T-SNE algorithm (see Figure 4). The main purpose of the T-SNE algorithm is to visualize and explore high-dimensional data, which has been successfully used in bioinformatics [20, 21]. As shown in Figure 4, the sample definition in the DSAE representation is much clearer, while the sample definition without DSAE learning is quite confusing. This means that our model will be able to identify classes individually by using these representations at a relatively small cost, which means that the features learned by DSAE are salient and effective.

In addition, to estimate the performance of the DSAE_RF, we further implement comparative experiments of different classifiers, such as the SVM, LR, NB, DT, GBDT, KNN and the latest machine learning model proposed by Bukhari SNH *et al.* (see Table 3); for convenience, we abbreviate this model as BDT (the model is an ensemble machine learning model based on DT) [22]. We conduct the experiment on the same dataset using the identical feature extraction method. Apart from this, the 10 experimental results of all the comparison classifiers are shown in Figure 5. We can see that our model has higher performance than the other seven classifiers for all evaluation metrics. The mentioned results show that RF is more reliable as the final model classifier.

Comparison with other methods

To further estimate the capability of the DSAE_RF model as well as to take into account the computational resource issue, we compare nine state-of-the-art methods for microbe–disease associations prediction based on 10-fold cross-validation, including the following models: BiRWRHMDA [8], KATZHMDA [11], KATZBNRA [12], LRLSHMDA [16], ABHMDA [17], BPNNHMDA [18], NCPHMD [23], NTSHMDA [24] and NBLPIHMDA [25]. It is worth noting that all models use the datasets in the paper.

The DSAE_RF model and all the comparison methods are trained and tested using the same dataset in the cross-validation. Simultaneously, the best parameters for each method are used in the implementation. As shown in Figure 6, our model attains the highest average AUC of 0.9448 and AUPR of 0.9431, higher than all comparison methods. In addition, the AUC and AUPR values of all methods compared in the 10 experiments are plotted in Figures 7

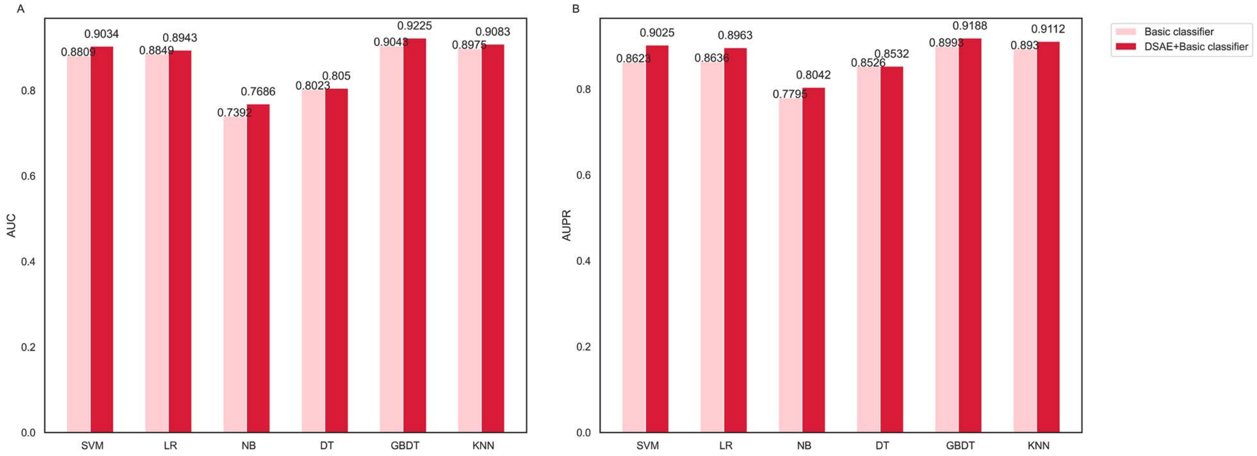

Figure 3. Comparison of results of base classifiers.

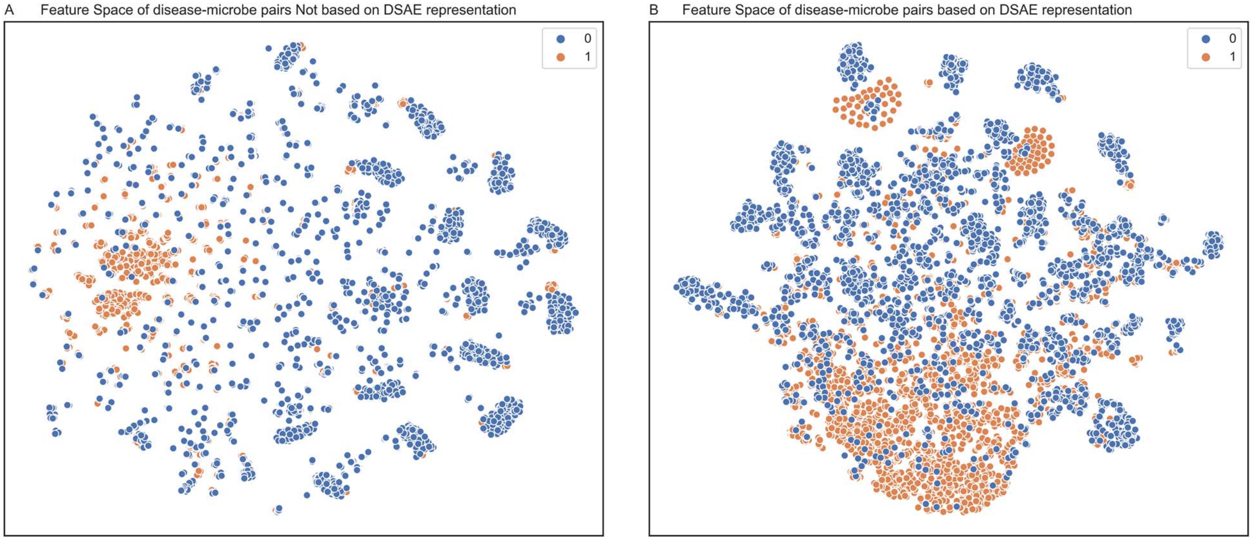

Figure 4. Feature space visualization. (A) Feature space visualization not based on DSAE representation. (B) Feature space visualization based on DSAE representation.

Table 3. Average results of different classifiers based on 10-fold cross-validation

classifiers	AUC (%)	AUPR (%)	Recall (%)	Pre (%)	ACC (%)	F1-score (%)
DSAE+SVM	90.34	90.25	81.42	81.66	81.41	81.36
DSAE+LR	89.43	89.63	80.72	80.76	80.73	80.70
DSAE+NB	76.86	80.42	72.53	72.58	72.55	72.51
DSAE+DT	80.50	85.32	80.50	80.51	80.53	80.50
DSAE+GBDT	92.25	91.88	83.62	83.68	83.62	83.60
DSAE+KNN	90.83	91.12	83.55	83.56	83.56	83.54
DSAE+BDT	81.21	85.84	81.21	81.23	81.24	81.21
DSAE_RF	94.48	94.31	86.67	86.69	86.69	86.67

and 8. These results suggest that our model outshines the other nine state-of-the-art comparative approaches, indicating DSAE_RF is a valid and robust computational method in predicting microbe–disease correlations.

Robustness analysis

In addition, to measure the generalization ability of our model within other datasets, the newly published MDADP database [26] is used to test our model, and the consequences are presented in Table 4. Furthermore, we also plot the average AUC and AUPR of different approaches in the MDADP database.

As illustrated in Figure 9, although our model performs slightly poorer than ABHMDA in terms of AUC, it is higher than that of the other eight models, exceeding LRLSHMDA 13.62%, NCPHMDA 14.56%, NTSHMDA 17.87%, BiRWRHMDA 23.09%, NBLPIHMDA 33.29%, KATZHMDA 12.19%, KATZBNRA 25.8% and BPNNHMDA 23.42%, respectively. For AUPR, our model outperforms all models, with AUPR 0.44% better than that of the second best ABHMDA and 83.16% higher than that of the worst performed KATZBNRA. These

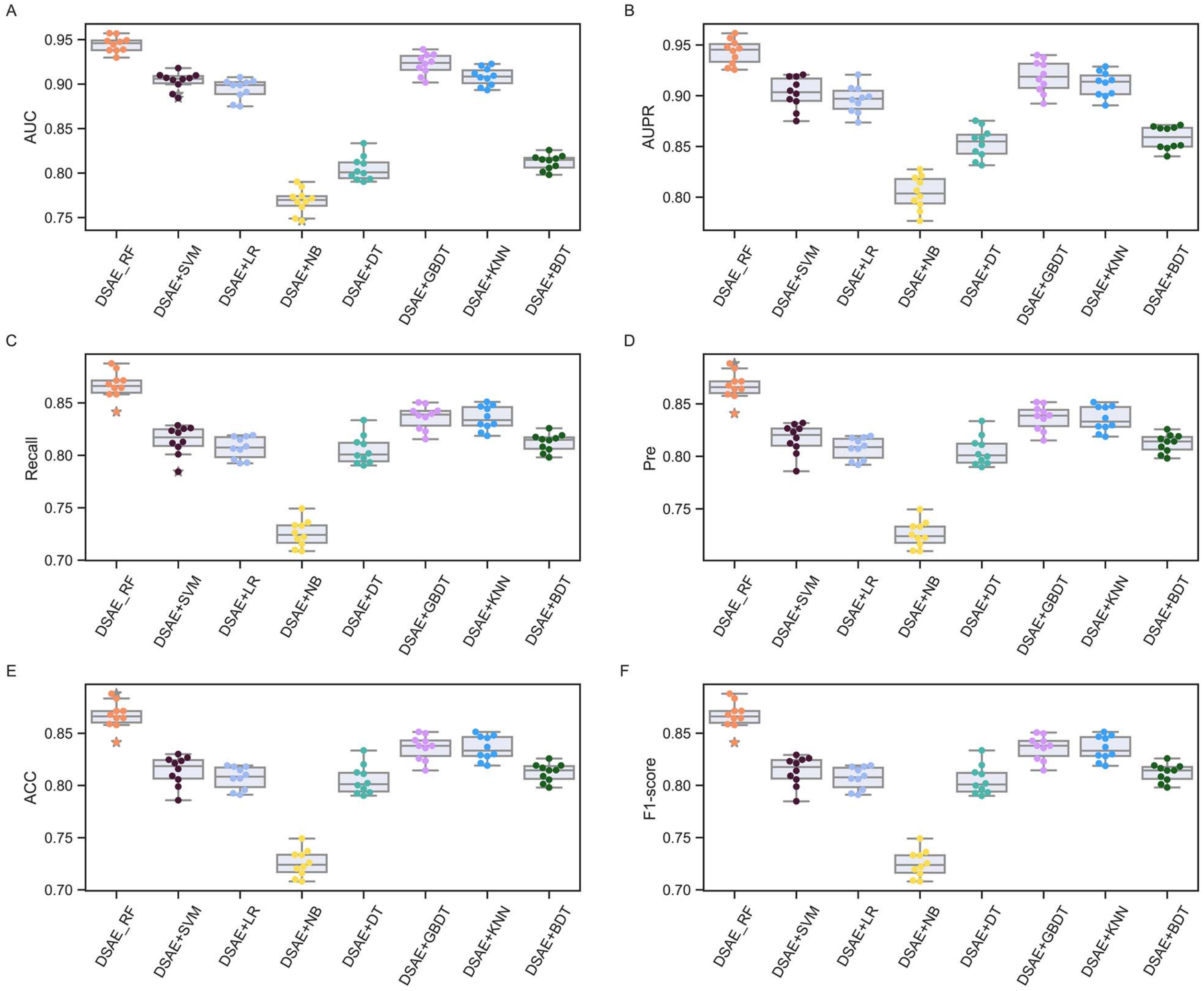

Figure 5. Comparison results between DSAE_RF and the other seven classifiers. (A) AUC, (B) AUPR, (C) ACC, (D) Pre, (E) ACC, (F) F1-score.

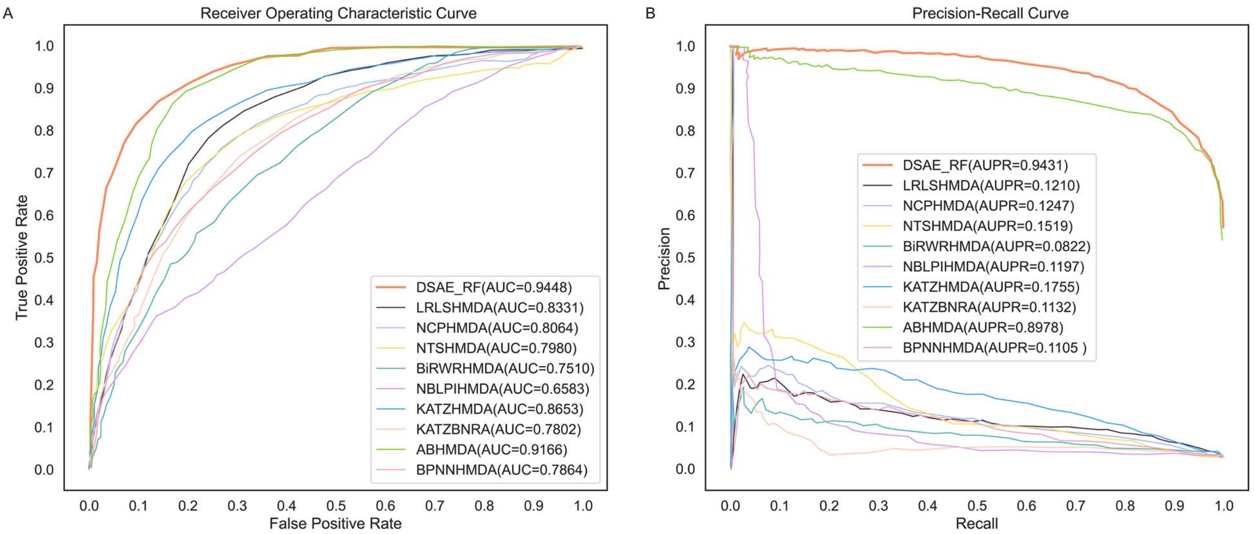

Figure 6. (A) The ROC curves of all the methods. (B) The PR curves of all the methods.

experimental results again confirm the validity and robustness of our model.

Kolmogorov–Smirnov test and *t*-test

In the study, although the AUC and AUPR of our model outperformed other advanced models, in order to avoid chance, we use statistical tests, i.e., Kolmogorov–Smirnov (KS) test and *t*-test to show that this is not coincident but is statistically significant. KS test is a non-parametric statistical test method, which is mainly used for the test of quantitative data. In addition, the KS test is also commonly used to apply whether the sample conforms to a known distribution. A *t*-test is usually used for the

Table 4. AUC, AUPR, Recall, Pre, ACC, F1-score by our model on MDADP database

Test set	AUC (%)	AUPR (%)	Recall (%)	Pre (%)	ACC (%)	F1-score (%)
1	93.70	93.77	83.89	83.91	83.90	83.90
2	95.41	94.63	86.99	86.94	87.01	86.96
3	95.08	94.42	86.14	86.13	86.16	86.13
4	95.03	94.59	87.70	87.55	87.57	87.56
5	92.11	92.07	82.25	82.21	82.20	82.20
6	92.14	91.38	84.19	84.18	84.18	84.18
7	92.75	92.52	83.15	83.15	83.05	83.05
8	93.85	93.91	84.60	84.55	84.46	84.46
9	92.01	91.66	83.72	84.06	83.85	83.78
10	93.37	94.32	83.46	83.52	83.57	83.49
Average	93.54 ± 1.23	93.33 ± 1.22	84.61 ± 1.68	84.62 ± 1.63	84.60 ± 1.66	84.57 ± 1.66

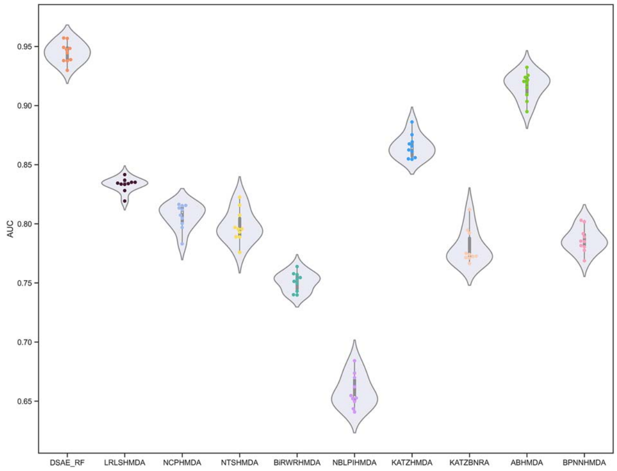

Figure 7. AUC distribution of 10 experiments of our model and nine models compared.

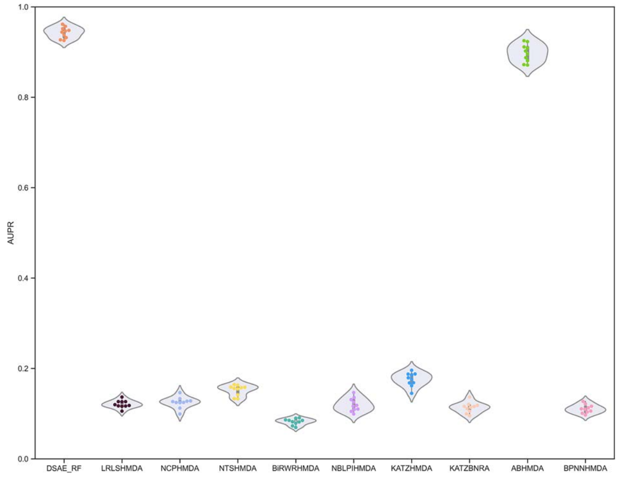

Figure 8. AUPR distribution of 10 experiments of our model and nine models compared.

detection of quantitative data and its main purpose is to compare whether there are significant differences between data samples. It is mainly tested by the difference of sample means, which is a comparison of the difference between two means. But the use of a *t*-test needs to satisfy the premise that the data obey normal distribution and the variance is unknown. Therefore, we need to use the KS test to judge whether the AUC and AUPR of all models conform to the normal distribution, that is, whether the distributions of AUC and AUPR satisfy the conditions of the *t*-test. If the *P*-value is greater than 0.05, it is considered to conform to the normal distribution, otherwise, it does not conform to the normal distribution.

The KS test results of the AUC and AUPR of all models are listed in Table 5. Meanwhile, the kernel density distributions of AUC and AUPR of all models are shown in Figures 10 and 11. The consequences indicate that the AUC and AUPR of all models are approximately normal distribution, which can be tested by the *t*-test. The results of the significance *t*-test indicate that DSAE_RF performs remarkably well than the other compared approaches (*P*-value <0.05, see Table 6).

Case studies

Case studies on predicted potential microbe–disease associations

As mentioned earlier, microbes play a key function in the progression of many diseases. To estimate the capability of DSAE_RF in identifying the underlying microbe–disease correlations, all known microbe–disease correlations are used to evaluate the latency of the correlations predicted by our model.

Covid-19 is a major respiratory disease, which has emerged for 3 years. Although Covid-19 can be preliminarily controlled, it still causes damage to human organs [27, 28]. At present, some researches have confirmed that microbes play an indispensable part in the treatment of Covid-19, which may be a new way to the treatment of Covid-19 [29]. For instance, the gut microbiota-derived symbiotic formula can accelerate the formation of antibodies against Covid-19, reduce the viral load of the nasopharynx, reduce inflammatory immune markers and restore intestinal ecological imbalance [30]. The respiratory microbiota can be utilized as a biomarker to predict the severity of Covid-19 [31].

Colorectal cancer is the second leading cause of cancer death in the United States, and rates are increasing year by year in young and middle-aged adults [32]. A growing number of evidence show that gut microbiota is strongly involved in the incidence and subsequent survival of colorectal cancer [33]. Some studies have determined *Enterotoxigenic Bacteroides fragilis* [34] and *Fusobacterium nucleatum* [35], which can be used as potential targets of colorectal cancer treatment.

Therefore, in this work, we implement the above-mentioned two case studies of Covid-19 and colorectal cancer to verify the reliability of our model. Among the top 10 predicted Covid-19-related microbes, seven microbes are confirmed by new publications (see Table 7). For example, Ahmadi *et al*. treated

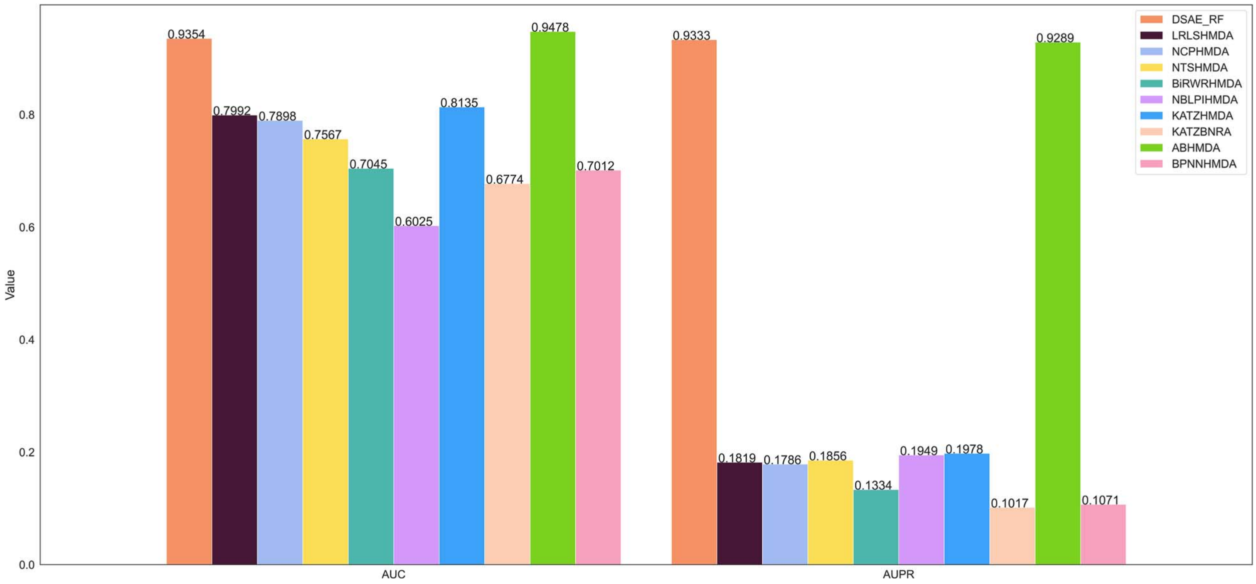

Figure 9. The average AUC and AUPR of different methods in the MDADP database.

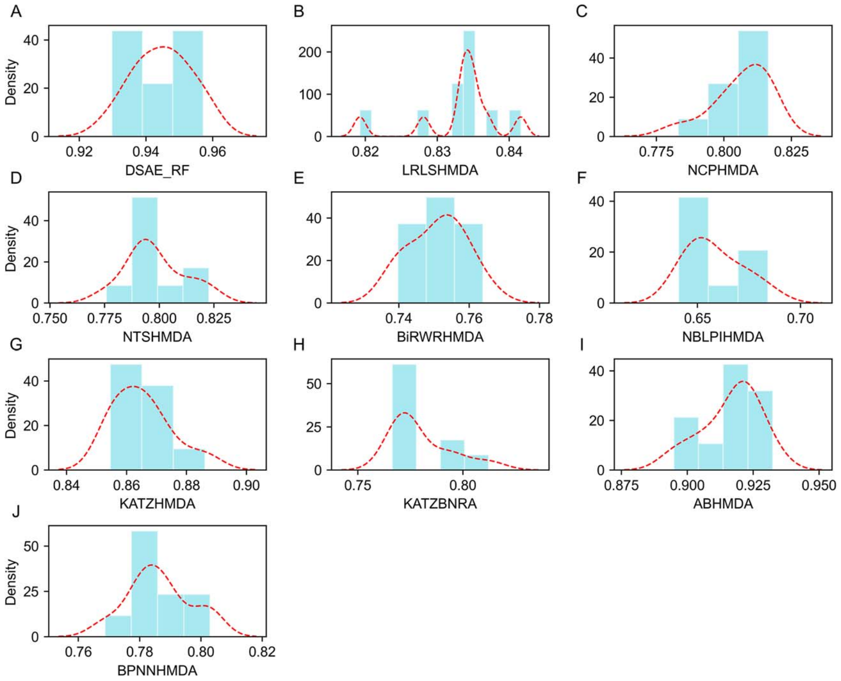

Figure 10. The kernel density distributions of AUC of all models.

Table 5. The KS test results of AUC and AUPR of all models

P-value	DSAE_RF	LRLSHMDA	NCPHMDA
AUC	0.9477	0.2249	0.7832
AUPR	0.9834	0.8150	0.3172
P-value	NTSHMDA	BiRWRHMDA	NBLPIHMDA
AUC	0.5732	0.9712	0.7287
AUPR	0.3039	0.8905	0.9739
P-value	KATZH-MDA	KATZBNRA	ABHMDA
AUC	0.9639	0.1613	0.8389
AUPR	0.9680	0.9118	0.9977
P-value	BPNNHMDA		
AUC	0.9759		
AUPR	0.9996		

Caco-2 cells overnight with a cell-free supernatant of a heat-inactivated form of *Bacteroides fragilis* and considered the genes (ACE, AGTR1, ACE2 and TMPRSS2) for enrichment analysis by GEO2R, DAVID. They find that *Bacteroides fragilis* down-regulated ACE, AGTR1 and ACE2 genes in vivo, heat-inactivated and cell-free supernatants. Whereas ACE, ATR1 and ACE 2 are fundamental genes driving the upregulation of bioprocesses in patients with Covid-19, this literature suggests that the expression of ACE, ATR1 and ACE2 may be reduced by *Bacteroides fragilis* and its subsequent biologics. In addition, this study shows that *Bacteroides fragilis* can be considered as a treatment strategy for Covid-19 [36]. However, Covid-19-*Bacteroides coprocola*, Covid-19-*Treponema amylovorum* and Covid-19-*Neisseria mucosa* correlations have not been experimentally verified.

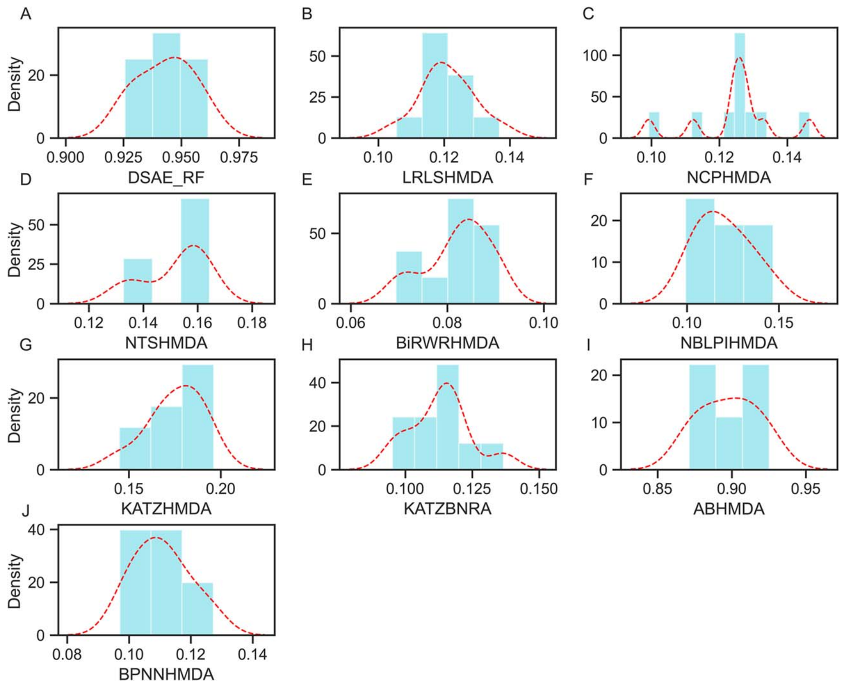

Figure 11. The kernel density distributions of AUPR of all models.

Table 6. Comparative results of *t*-tests

Methods	***P*-value**	
	AUC	**AUPR**
LRLSHMDA	9.6070E-10	8.2060E-17
NCPHMDA	6.2752E-10	1.3121E-15
NTSHMDA	2.1910E-09	9.0195E-16
BiRWRHMDA	1.1390E-12	5.5542E-17
NBLPIHMDA	3.3315E-12	2.3317E-15
KATZHMDA	8.0392E-08	8.3443E-15
KATZBNRA	7.0629E-10	1.0788E-16
ABHMDA	5.7549E-05	6.3061E-05
BPNNHMDA	2.5256E-11	3.0006E-16

In addition, it can also be observed that among the prediction results of microbes related to colorectal cancer, 8 of the top 10 microbes can be verified from the literature (see Table 7). For example, according to Xu *et al.* [37], in their findings, *Actinomyces* play an influential function in the promotion of colorectal cancer and thus have the possibility to be a target for antitumor therapy. They use 16S rRNA sequencing analysis to determine that *Actinomyces* are a critical microbiota in the colorectal cancer population. The results of the correlation analysis also show that *Actinomyces* co-exist with different preneoplastic microbial taxa. These results demonstrate the predictive performance of DSAE_RF and are thus very useful for screening candidate microbes for diseases.

Case studies on predicted new microbes-related diseases

For every disease, it is considered a new disease. All its associated microbes are eliminated for predicting the new microbes associated with that disease. The top 10 results of Covid-19 and colorectal cancer are listed in Table 8. Five of them have been confirmed in the literature, demonstrating their associations with Covid-19, which may be beneficial for Covid-19 treatment. Seven of the top candidate microbes are found to be related to colorectal cancer, which is supported by the literature. And they are involved in the pathogenesis of colorectal cancer. These results can be used to assist biologists in discovering true novel disease-related microbes, effectively reducing the wet-lab experimental time. In summary, DSAE_RF has obtained excellent performance in the prediction of neo-disease-related microbes and may offer innovative and reliable therapeutic strategies for cancers.

DISCUSSION AND CONCLUSION

The recognition of potential microbe–disease associations not only contributes to disease diagnosis, treatment and prognosis, but also facilitates microbe-oriented treatments in precision medicine. Although many efforts have been made in this field, there are still some limitations, including poor prediction accuracy, complicated data fusion for feature extraction and unreliable negative sample selection.

In the study, we present a new deep learning methodology named DSAE_RF, for predicting disease-associated microbes. Specifically, after completing the data collection and reorganization, we compute the four similarities between microbes and diseases separately, and fuse the four similarities into the sample features of disease-microbe pairs. Then, the reliable negative samples are selected by *k*-means clustering, which is conducive to improving the forecast accuracy of the model. Next, we learn the useful features of the samples through the DSAE neural network. Finally, the RF classifier is applied to predict underlying microbe–disease correlations. Ablation experiments, comparative experiments, robustness analysis, statistical tests and case studies indicate that the DSAE_RF model is credible and prospective in recognizing underlying target microbes of diseases.

Although our model shows good performance, there are still some limitations, and further improvements are needed in the future. On the one hand, we simply fuse the similarities of different species, which cannot well represent microbes and diseases. In the future, we will present a stronger fusion features method to

Table 7. The top 10 potential associated microbes for predicting Covid-19 and colorectal cancer

Covid-19			**Colorectal cancer**	
Rank	**Microbes**	**Evidence**	**Microbes**	**Evidence**
1	*Bacteroides fragilis*	PMID: 36174833	*Tm7*	PMID: 28373465
2	*Akkermansia*	PMID: 36011823	*Eggerthella*	PMID: 32126968
3	*Bifidobacterium longum*	PMID: 34258278	*Betaproteobacteria*	PMID: 27797671
4	*Bacteroides*	PMID: 36012406	*Dietzia maris*	Unconfirmed
5	*Tannerella*	PMID: 34835510	*Streptococcus constellatus*	PMID: 34434477
6	*Bacteroides coprocola*	Unconfirmed	*Porphyromo nadaceae*	PMID: 27880935
7	*Streptococcus*	PMID: 36183582	*Actinomyces*	PMID: 31171880
8	*Treponema amylovorum*	Unconfirmed	*Stenotrophomonas maltophilia*	PMID: 25447194
9	*Neisseria mucosa*	Unconfirmed	*Streptobacillus*	Unconfirmed
10	*Dialister*	PMID:36011823	*Cellulosilyticum*	PMID: 35008173

Table 8. The top 10 new associated microbes for predicting Covid-19 and colorectal cancer

Covid-19			**Colorectal cancer**	
Rank	**Microbes**	**Evidence**	**Microbes**	**Evidence**
1	*Lyrodus pedicellatus*	Unconfirmed	*Vivictivallis*	Unconfirmed
2	*Bifidobacterium longum*	PMID: 34643888	*Butiricimonas*	Unconfirmed
3	*Candida glabrata*	PMID: 36135649	*Micrococcus luteus*	PMID: 22117164
4	*Tm7*	PMID: 33672177	*Porphyromonadaceae*	PMID: 35664963
5	*Eubacterium saburreum*	Unconfirmed	*Gemella sanguinis*	PMID: 29214046
6	*Cardiobacterium valvarum*	Unconfirmed	*Delftia*	PMID: 34440574
7	*Leclercia*	Unconfirmed	*Candidatus brocadia*	Unconfirmed
8	*Bacteroides fragilis*	PMID: 35082504	*Escherichia coli*	PMID: 31554963
9	*Ebv*	PMID: 34088334	*Streptococcus sobrinus*	PMID: 24010070
10	*Mobiluncus curtisii*	Unconfirmed	*Prevotella*	PMID: 33488574

Table 9. The basic information about three databases

Database	**Microbes**	**Diseases**	**Associations**
HMDAD	292	39	450
Disbiome	1582	351	8645
Peryton	1396	43	4172

make up for this deficiency. On the other hand, we only consider the features of microbes and diseases, and ignore the effects of other biomolecules such as lncRNA on microbes and diseases. In the near future, we will consider introducing omics data to better identify latent correlations between microbes and diseases.

MATERIALS AND METHODS

Human microbe–disease associations

In the paper, we integrate multiple biological data into our prediction model of DSAE_RF. The human microbe–disease associations datasets are retrieved from HMDAD [38], Disbiome [39] and Peryton [40] databases, respectively. After deleting redundant data, the number of all microbes, diseases and their associations included in the three datasets is presented in Table 9.

Ultimately, after identifier standardization, de-duplication, simplification and removal of irrelevant items, some basic characteristics of the microbe–disease association dataset are summarized in Table 10. In addition, we use an adjacency matrix *MD* of size $m \times n$ to store all microbe–disease associations. If microbe m_i is associated with the disease d_j, $MD\ (i, j) = 1$, otherwise $MD\ (i, j) = 0$.

Table 10. The basic characteristics of microbe–disease associations

	Name	**Number**
Total	Microbes	1177
	Diseases	134
	Associations	4499
Average degree	Microbes	3.8
	Diseases	33.6
Max degree	Microbes	59
	Diseases	255

Similarity calculation

In the paper, we utilize four different methods to compute the similarity between microbes and diseases. Firstly, we calculate the disease semantic similarity. On this basis, we compute the microbes' functional similarity. Then, the cosine similarity, Gaussian interaction profile kernel similarity and sigmoid kernel function similarity between microbes and diseases are calculated according to the known microbe–disease correlations. Details of the similarity calculation can be found in the Supplementary Document.

Multi-source features fusion

The effective fusion of multi-source similarity is also an important task for us to apply deep learning methods. It has been estimated that feature fusion can produce more significant characteristics that comprehensively capture the characteristics of microbes and diseases. In this study, we fuse multiple microbe similarities

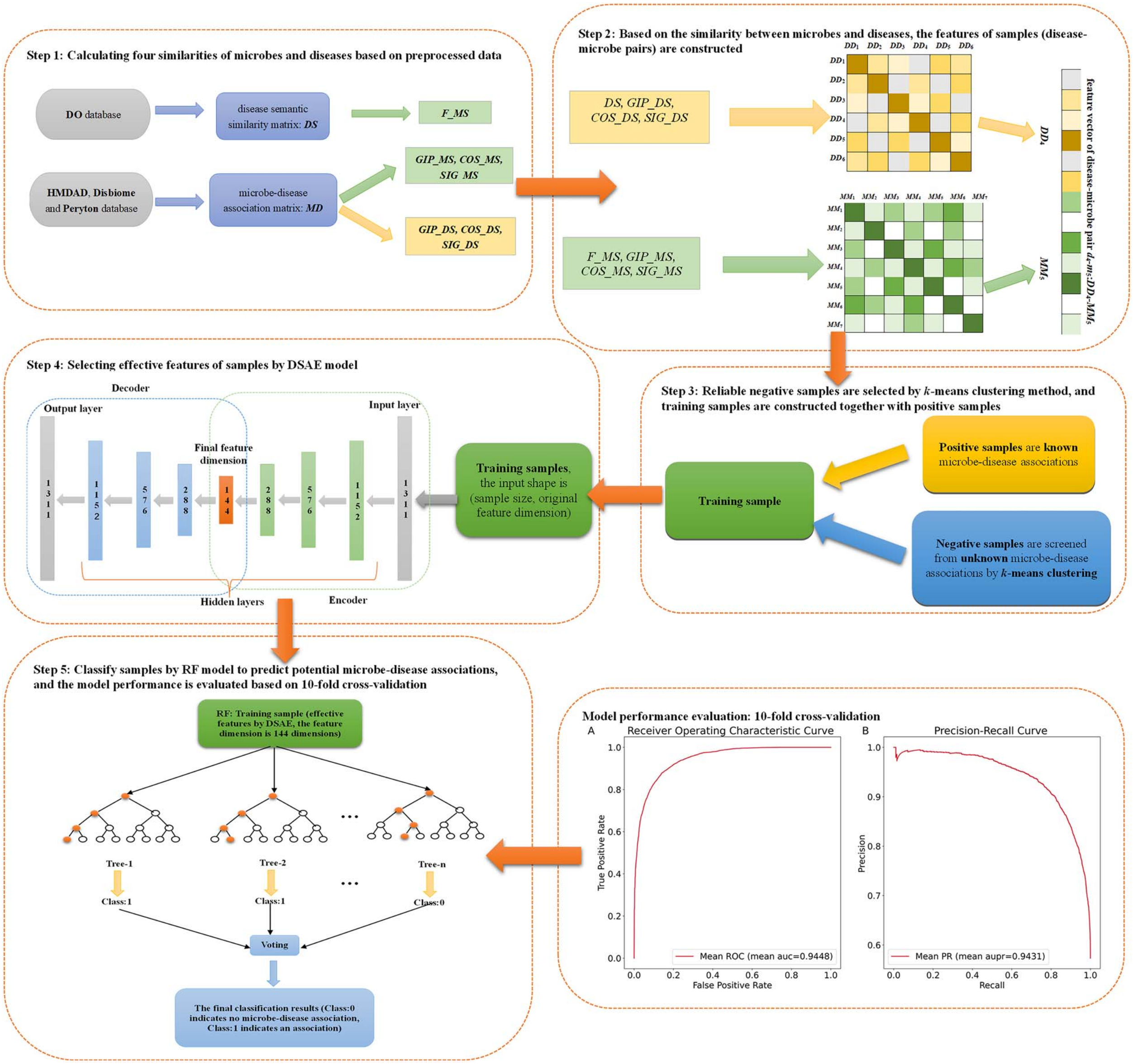
Step 1: Calculating four similarities of microbes and diseases based on preprocessed data
DO database
disease semantic similarity matrix: DS
F_MS
HMDAD, Disbiome and Peryton database
microbe-disease association matrix: MD
GIP_MS, COS_MS, SIG_MS
GIP_DS, COS_DS, SIG_DS
Step 2: Based on the similarity between microbes and diseases, the features of samples (disease-microbe pairs) are constructed
DS, GIP_DS, COS_DS, SIG_DS
DD1 DD2 DD3 DD4 DD5 DD6
DD4
F_MS, GIP_MS, COS_MS, SIG_MS
MM1 MM2 MM3 MM4 MM5 MM6 MM7
MM5
feature vector of disease-microbe pair d4-m5:DD4-MM5
Step 4: Selecting effective features of samples by DSAE model
Decoder
Output layer
Input layer
Final feature dimension
1311
1152
576
288
144
288
576
1152
1311
Hidden layers
Encoder
Training samples, the input shape is (sample size, original feature dimension)
Step 3: Reliable negative samples are selected by k-means clustering method, and training samples are constructed together with positive samples
Positive samples are known microbe-disease associations
Training sample
Negative samples are screened from unknown microbe-disease associations by k-means clustering
Step 5: Classify samples by RF model to predict potential microbe-disease associations, and the model performance is evaluated based on 10-fold cross-validation
RF: Training sample (effective features by DSAE, the feature dimension is 144 dimensions)
Tree-1
Tree-2
Tree-n
Class:1
Class:1
Class:0
Voting
The final classification results (Class:0 indicates no microbe-disease association, Class:1 indicates an association)
Model performance evaluation: 10-fold cross-validation
A
Receiver Operating Characteristic Curve
True Positive Rate
False Positive Rate
Mean ROC (mean auc=0.9448)
B
Precision-Recall Curve
Precision
Recall
Mean PR (mean aupr=0.9431)
0.0
0.2
0.4
0.6
0.8
1.0
0.6
0.7
0.8
0.9
1.0

Figure 12. Flowchart of DSAE_RF model to predict potential microbe–disease associations.

into microbe functional similarity, and so do disease semantic similarity.

For the similarities of microbes, we combine microbe functional similarity *F_MS*, microbe GIP similarity *GIP_MS*, microbe cosine similarity *COS_MS* and microbe sigmoid function kernel similarity *SIG_MS* to form microbe similarity *MM*. The microbe similarity of m_i and m_j can be calculated as follows:

$$MM\left(m_i, m_j\right) = \frac{F\_MS + GIP\_MS + COS\_MS + SIG\_MS}{4}. \quad (6)$$

Similarly, for the similarity of diseases *DD* (d_i, d_j), we calculate as follows:

$$DD\left(d_i, d_j\right) = \frac{DS + GIP\_DS + COS\_DS + SIG\_DS}{4}. \quad (7)$$

Reliable negative sample selection

In this paper, there are only 4499 positive samples, accounting for about 2.8% of the total sample. Therefore, selecting balanced and high-quality negative samples can improve the efficiency of the model in the prediction of microbe–disease correlations. We introduce the method from Peng *et al.* to select negative samples, namely the *k*-means clustering method [17, 41]. Specifically, we divide the negative samples into *k* fractions and randomly select some samples from every fraction as the negative samples, whereas the positive samples remain invariant. According to previous research, when *k* is 23, the effect of the model is the best [17]. Therefore, we let *k* = 23. To keep the balance of training samples, the quantity of negative samples should be basically the same as the quantity of positive samples. In the end, we choose 4508 negative samples and 4499 positive samples, that is, these 9007 samples as the training samples.

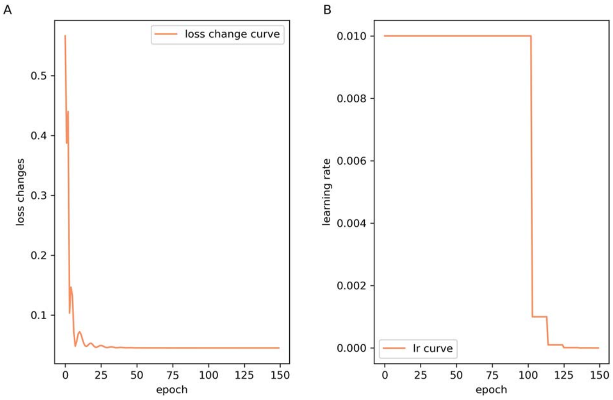

Figure 13. The changes of loss value and learning rate with epoch.

Feature extraction based on DSAE

Recently, deep learning has been extensively applied in bioinformatics, and feature extraction is the most basic and important research problem of deep learning. As an unsupervised model, a sparse autoencoder neural network can effectively learn the latent information of samples [42], which introduces a Kullback–Leibler (KL) divergence as a sparse penalty term on the basis of autoencoder neural network to learn relatively sparse features [43].

In this paper, the feature vectors of each disease-microbe pair are constructed for disease similarity *DD* and microbe similarity *MM*. Specifically, let $DD = DD_1, DD_2, \cdots DD_n$, $\mathrm{MM} = MM_1, MM_2, \cdots MM_m$, for any disease-microbe pair d_i-m_j, DD_i, MM_j is regarded as its feature vector whose dimension is 1311 (see Figure 12), where DD_i and MM_j represent the ith row of *DD* and the *j*th row of *MM*, respectively.

After constructing the feature vectors of samples, considering that their feature dimension is too large, we construct a DSAE neural network to obtain the most significant characteristics of disease-microbe pairs. Its basic structure is shown in Figure. 12. DSAE mainly includes an input layer, multiple hidden layers and an output layer. It is worth noting that the quantity of neurons in the input and output layers are the same, 1311. Through the experiments, we set up seven hidden layers. Finally, we reduce the feature vector dimension of each sample to 144. Details of the DSAE can be found in the Supplementary Document.

In addition, the epoch numbers are chosen based on the fact that the loss becomes small enough, and the learning rate changes. In Figure 13(A), with the increase of epochs, the loss of model drops sharply, and starts to decline slowly when the number of epochs is 25. When the epoch is 150, the model loss is less than 0.1. From Figure 13(B), it can be concluded that the learning rate decreases from 0.01 to 0.001 when the epoch number is around 100. Finally, when the epoch number is 150, the learning rate is 1E-07.

Prediction of microbe–disease associations based on DSAE_RF model

In the paper, we present a new model named DSAE_RF for predicting underlying microbe–disease correlations. As shown in Figure 12, we compute four similarities between microbes and diseases, which are then used as characteristics of disease-microbe pairs. Furthermore, the negative samples with an equal number of positive samples are selected by *k*-means clustering, and, subsequently, the positive and negative samples are input into DSAE as training samples to learn the effective features of the disease-microbe pairs. After feature extraction by DSAE, we choose RF as the final classifier for predicting the underlying microbe–disease correlations. Ensemble learning algorithms have attracted increasing attention in recent years because they are more accurate. Ensemble learning, also called classifier ensemble, completes learning tasks by building and combining multiple learners. The structure is generally: using a certain strategy to combine a group of 'individual learners' (or 'basic learners', or 'weak classifiers'). Among them, the combination strategy mainly includes the average method, voting method and learning method. The idea of ensemble learning is to create a strong learner (or ensemble model) by combining the bias sum or variances of these 'base learners' and thus achieve better performance.

Among them, RF is an important comprehensive learning method based on bagging. Multiple different data sets are generated by resampling the data set bootstrap, and a classification tree is trained on each data set, and finally the prediction results of each classification tree are combined as the prediction results of RF. RF has the following advantages: RF can handle high-dimensional data sets with higher accuracy and generalization capabilities. In addition, RF can also effectively reduce the risk of overfitting and underfitting. Therefore, an RF classifier is chosen to process the learned features, and the experimental results in this work also confirm the reliability and effectiveness of choosing RF as the final classifier. We combine random search and halving grid search algorithms to determine the hyperparameters of RF, including n_estimators, max_depth, min_samples_leaf, min_samples_split, criterion and max_feature set to 1100, 23, 5, 6, 'entropy' and 'sqrt', respectively. Details of parameter selection can be found in the Supplementary Document.

Key Points

- DSAE_RF integrates four different similarity networks to better characterize disease-microbe pairs.
- DSAE_RF selects neural networks to extract effective features of disease–microbe pairs, which helps to improve the efficiency and accuracy of the model.
- DSAE_RF contributes to the stability of the model by selecting reliable negative samples, and in addition, our model can predict microbes associated with new diseases.
- A series of experiments have demonstrated that DSAE_RF can effectively predict microbe–disease associations by selecting random forest as the final classifier.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

FUNDING

The work is supported by the National Natural Science Foundation of China (Grant No. 12271216, 11831015).

COMPETING INTERESTS

The authors declare that they have no competing financial interests.

DATA AND CODE AVAILABILITY

The DSAE_RF model is implemented to work under Python 3.8, and the main libraries used are as follows: 'torch', 'numpy', 'pandas', 'sklearn', 'math'. The original datasets and codes are provided in the GitHub repository: https://github.com/wang-124/DSAE_RF.git. In addition, the datasets analyzed in this study can be found in the HMDAD (http://www.cuilab.cn/hmdad), Disbiome (https://disbiome.ugent.be/), and Peryton (https://dianalab.e-ce.uth.gr/peryton/) databases.

References

1. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;**486**: 207–14.
2. Proal AD, Lindseth IA, Marshall TG. Microbe-microbe and host-microbe interactions drive microbiome dysbiosis and inflammatory processes. *Discov Med* 2017;**23**:51–60.
3. Shahi SK, Freedman SN, Mangalam AK. Gut microbiome in multiple sclerosis: the players involved and the roles they play. *Gut Microbes* 2017;**8**:607–15.
4. Dono A, Nickles J, Rodriguez-Armendariz AG, *et al.* Glioma and the gut-brain axis: opportunities and future perspectives. *Neurooncol Adv* 2022;**4**:vdac054.
5. Arweiler NB, Netuschil L. The oral microbiota. *Adv Exp Med Biol* 2016;**902**:45–60.
6. Pozhitkov AE, Leroux BG, Randolph TW, *et al.* Towards microbiome transplant as a therapy for periodontitis: an exploratory study of periodontitis microbial signature contrasted by oral health, caries and edentulism. *BMC Oral Health* 2015;**15**: 125.
7. Ibáñez L, de Mendoza I, Maritxalar Mendia X, *et al.* Role of *Porphyromonas gingivalis* in oral squamous cell carcinoma development: a systematic review. *J Periodontal Res* 2020;**55**:13–22.
8. Zou S, Zhang J, Zhang Z. A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network. *PloS One* 2017;**12**:e0184394.
9. Peng W, Liu M, Dai W, *et al.* Multi-view feature aggregation for predicting microbe-disease association. *IEEE/ACM Trans Comput Biol Bioinform* 2021;1. https://doi.org/10.1109/TCBB.2021.3132611.
10. Wang Y, Lei X, Lu C, *et al.* Predicting microbe-disease association based on multiple similarities and LINE algorithm. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**19**:2399–408.
11. Chen X, Huang YA, You ZH, *et al.* A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 2018;**34**:1440.
12. Li S, Xie M, Liu X. A novel approach based on bipartite network recommendation and KATZ model to predict potential micro-disease associations. *Front Genet* 2019;**10**:1147.
13. Yin MM, Liu JX, Gao YL, *et al.* NCPLP: a novel approach for predicting microbe-associated diseases with network consistency projection and label propagation. *IEEE Trans Cybern* 2022;**52**: 5079–87.
14. Yang X, Kuang L, Chen Z, *et al.* Multi-similarities bilinear matrix factorization-based method for predicting human microbe-disease associations. *Front Genet* 2021;**12**:754425.
15. Xu D, Xu H, Zhang Y, *et al.* Novel collaborative weighted non-negative matrix factorization improves prediction of disease-associated human microbes. *Front Microbiol* 2022;**13**:834982.
16. Wang F, Huang ZA, Chen X, *et al.* LRLSHMDA: Laplacian regularized least squares for human microbe-disease association prediction. *Sci Rep* 2017;**7**:7601.
17. Peng LH, Yin J, Zhou L, *et al.* Human microbe-disease association prediction based on adaptive boosting. *Front Microbiol* 2018;**9**:2440.
18. Li H, Wang Y, Zhang Z, *et al.* Identifying microbe-disease association based on a novel back-propagation neural network model. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**18**:2502–13.
19. Bukhari SNH, Jain A, Haq E, *et al.* Ensemble machine learning model to predict SARS-CoV-2 T-cell epitopes as potential vaccine targets. *Diagnostics* 2021;**11**:1990.
20. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
21. Naseer S, Hussain W, Khan YD, *et al.* Optimization of serine phosphorylation prediction in proteins by comparing human engineered features and deep representations. *Anal Biochem* 2021;**615**:114069.
22. Bukhari SNH, Webber J, Mehbodniya A. Decision Tree based ensemble machine learning model for the prediction of Zika virus T-cell epitopes as potential vaccine candidates. *Sci Rep* 2022;**12**:7810.
23. Bao W, Jiang Z, Huang DS. Novel human microbe-disease association prediction using network consistency projection. *BMC Bioinform* 2017;**18**:543.
24. Luo J, Long Y. NTSHMDA: prediction of human microbe-disease association based on random walk by integrating network topological similarity. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**17**: 1341–51.
25. Wang L, Wang Y, Li H, *et al.* A bidirectional label propagation based computational model for potential microbe-disease association prediction. *Front Microbiol* 2019;**10**:684.
26. Wang L, Li H, Wang Y, *et al.* MDADP: a webserver integrating database and prediction tools for microbe-disease associations. *IEEE J Biomed Health Inform* 2022;**26**:3427–34.
27. Dinakaran D, Manjunatha N, Naveen Kumar C, *et al.* Neuropsychiatric aspects of COVID-19 pandemic: a selective review. *Asian J Psychiatr* 2020;**53**:102188.
28. Salian VS, Wright JA, Vedell PT, *et al.* COVID-19 transmission, current treatment, and future therapeutic strategies. *Mol Pharm* 2021;**18**:754–71.
29. Taglialatela-Scafati O. New hopes for drugs against COVID-19 come from the sea. *Mar Drugs* 2021;**19**:104.
30. Zhang L, Xu Z, Mak J, *et al.* Gut microbiota-derived synbiotic formula (SIM01) as a novel adjuvant therapy for COVID-19: an open-label pilot study. *J Gastroenterol Hepatol* 2022;**37**:823–31.
31. Chen J, Liu X, Liu W, *et al.* Comparison of the respiratory tract microbiome in hospitalized COVID-19 patients with different disease severity. *J Med Virol* 2022;**94**:5284–93.
32. Siegel RL, Miller KD, Goding Sauer A, *et al.* Colorectal cancer statistics, 2020. *CA Cancer J Clin* 2020;**70**:145–64.
33. Gao ZY, Cui Z, Yan YQ, *et al.* Microbe-based management for colorectal cancer. *Chin Med J Engl* 2021;**134**:2922–30.
34. Liu QQ, Li CM, Fu LN, *et al.* Enterotoxigenic *Bacteroides fragilis* induces the stemness in colorectal cancer via upregulating histone demethylase JMJD2B. *Gut Microbes* 2020; **12**:1788900.
35. Chen S, Su T, Zhang Y, *et al.* *Fusobacterium nucleatum* promotes colorectal cancer metastasis by modulating KRT7-AS/KRT7. *Gut Microbe* 2020;**11**:511–25.
36. Ahmadi Badi S, Malek A, Paolini A, *et al.* Downregulation of ACE, AGTR1, and ACE2 genes mediating SARS-CoV-2 pathogenesis by gut microbiota members and their postbiotics on Caco-2 cells. *Microb Pathog* 2022;**173**:105798.
37. Xu Z, Lv Z, Chen F, *et al.* Dysbiosis of human tumor microbiome and aberrant residence of Actinomyces in tumor-associated

fibroblasts in young-onset colorectal cancer. *Front Immunol* 2022;**13**:1008975.

38. Ma W, Zhang L, Zeng P, *et al.* An analysis of human microbe-disease associations. *Brief Bioinform* 2017;**18**:85–97.
39. Janssens Y, Nielandt J, Bronselaer A, *et al.* Disbiome database: linking the microbiome to disease. *BMC Microbiol* 2018;**18**:50.
40. Skoufos G, Kardaras FS, Alexiou A, *et al.* Peryton: a manual collection of experimentally supported microbe-disease associations. *Nucleic Acids Res* 2021;**49**:D1328–33.
41. Cheng H, Wang Z, Wei Z, *et al.* On adaptive learning framework for deep weighted sparse autoencoder: a multiobjective evolutionary algorithm. *IEEE Trans Cybern* 2022;**52**:3221–31.
42. Lu Y, Cheung YM, Tang YY. Self-adaptive multiprototype-based competitive learning approach: a k-means-type algorithm for imbalanced data clustering. *IEEE Trans Cybern* 2021;**51**: 1598–612.
43. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat* 1951;**22**:79–86.