Original Research

Prediction of diabetic protein markers based on an ensemble method

Kaiyang Qu[1], Quan Zou[2,3], Hua Shi[4,*]

[1]*School of Computer and Software, Nanyang Institute of Technology, 473004 Nanyang, Henan, China,* [2]*Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, 610054 Chengdu, Sichuan, China,* [3]*Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, 324000 Quzhou, Zhejiang, China,* [4]*School of Opto-electronic and Communication Engineering, Xiamen University of Technology, 361024 Xiamen, Fujian, China*

TABLE OF CONTENTS

1. Abstract

Introduction: A diabetic protein marker is a type of protein that is closely related to diabetes. This kind of protein plays an important role in the prevention and diagnosis of diabetes. Therefore, it is necessary to identify an effective method for predicting diabetic protein markers. In this study, we propose using ensemble methods to predict diabetic protein markers. **Methodological issues**: The ensemble method consists of two aspects. First, we combine a feature extraction method to obtain mixed features. Next, we classify the protein using ensemble classifiers. We use three feature extraction methods in the ensemble method, including composition and physicochemical features (abbreviated as 188D), adaptive skip gram features (abbreviated as 400D) and g-gap (abbreviated as 670D). There are six traditional classifiers in this study: decision tree, Naive Bayes, logistic regression, part, k-nearest neighbor, and kernel logistic regression. The ensemble classifiers are random forest and vote. First, we used feature extraction methods and traditional classifiers to classify protein sequences. Then, we compared the combined feature extraction methods with single methods. Next, we compared ensemble classifiers to traditional classifiers. Finally, we used ensemble classifiers and combined feature extraction methods to predict samples. **Results**: The results indicated that ensemble methods outperform single methods with respect to either ensemble classifiers or combined feature extraction methods. When the classifier is a random forest and the feature extraction method is 588D (combined 188D and 400D), the performance is best among all methods. The second best ensemble feature extraction method is

Submitted: 26 May 2021 Revised: 10 June 2021 Accepted: 25 June 2021 Published: 30 July 2021

1285D (combining the three methods) with random forest. The best single feature extraction method is 188D, and the worst one is g-gap. **Conclusion**: According to the results, the ensemble method, either the combined feature extraction method or the ensemble classifier, was better than the single method. We anticipate that ensemble methods will be a useful tool for identifying diabetic protein markers in a cost-effective manner.

2. Introduction

Due to continuous improvements and changes in people's lifestyles, an increasing number of people are suffering from diabetes mellitus [1]. At present, diabetes is one of the most prevalent diseases in many countries. According to clinical diagnosis, people who suffer from diabetes to be younger, and the incidence of diabetics is rising [2]. Therefore, improving the diagnostic efficiency of diabetes and identifying diabetic protein markers for use are currently hot topics. The continuous development of machine learning has resulted in its increasing use for disease prediction [3–12]. Machine learning methods to predict diabetes mellitus (DM) have been around for some time, eliciting sufficient controversy.

With the development of sequencing technology, the protein's function has gradually been found. Thus, proteomics has become a popular research hotspot. The essence of proteomics is to study proteins on a large-scale level, including protein expression, post-translational modification, protein-protein interactions [13], etc. Proteome research can not only provide a material basis for the laws of life activities, but also provide theoretical basis and solutions for the elucidation and conquering of many kinds of disease mechanisms [14]. By comparing and analyzing the proteome between normal individuals and pathological individuals, we can find certain "disease-specific protein molecules", which can become molecular targets for new drug design, or provide molecular markers for early diagnosis of diseases. Therefore, proteomics research is not only a necessary work to explore the mysteries of life, but also can bring huge benefits to human health. Proteomics research is a symbol which indicates that life science has entered the post-gene era [15]. Currently, we can make relevant predictions based on machine learning and protein markers. For example, machine learning has also been used for age prediction using protein markers [16] as well as the detection of other prevalent age-related diseases. Fleischer *et al*. [16] studied a computational method that can use ensemble machine learning methods to predict biological age from gene expression data of skin fibroblasts. Reboucas *et al*. [17] used biomarkers to detect the potential for recurrence of lung adenocarcinoma after surgical resection. These studies illustrate the importance of predicting protein markers. So, in this study, we use machine learning method to predict diabetic protein markers.

Protein is the material basis of all life, an important part of the composition of cells, and the primary raw material for the regeneration and repair of human tissue. Changes in protein morphology and quantity may lead to a variety of diseases, and some diseases may affect protein synthesis. Diabetic marker proteins are linked to diabetes, and these proteins directly or indirectly affect the diagnosis of diabetes [18]. Huth *et al*. [19] used proteomics to predict T2D. In this experiment, the authors selected 892 people who were 42 to 81 years old. Through the experiments, the authors found that the level of mannan-binding lectin serine peptidase (MASP) was positively correlated with T2D and prediabetes. However, adiponectin is negatively correlated with the T2D. MASP, adiponectin, apolipoprotein A-IV, apolipoprotein C-II and C-reactive protein are related to prognosis. The results show that diabetes can be predicted by protein levels in the body. As we all know, diabetes can cause a series of complications, such as diabetic nephropathy (DN). Hirao *et al*. [20] conducted a comprehensive analysis of the diabetic patients and healthy people, using label-free semi-quantitative methods. Protein identification analysis showed that there are 327 proteins unique to healthy people and 30 unique proteins to diabetic patients. There are a total of 615 proteins in the two groups. Gestational diabetes mellitus (GDM) refers to abnormal blood sugar that occurs during pregnancy. It is one of the common pregnancy complications in obstetrics, and it has serious adverse effects on the health of mothers and babies [19]. Through experiments, Kim *et al*. [21] proved that the level of apolipoprotein C III in women with GDM was significantly increased. According to experimental data, it can be found that there are biomarkers in patients with gestational diabetes at 16–20 weeks of pregnancy. Therefore, it is completely feasible to determine protein biomarkers and predict the later development of GDM.

As above, there are many proteins related to diabetes. Their presence or level of presence can usually be used as a criterion for judging diabetes. So, it is important to identify the kind of proteins which are associated to diabetes. Establishing a good protein classification model and identifying diabetic marker proteins are important steps for understanding and predicting diabetes. The main methods currently used to study proteomics include two-dimensional gel electrophoresis (2-DE), time-of-flight mass spectrometry (TOFMS), semi-quantitative multiple reaction monitoring (SQMRM) and bioinformatics technology, etc. [22]. These methods mainly use biological methods to analyze proteins. Biological methods can accurately perform qualitative analysis, but these methods produce a series of costs. Moreover, when biological methods face an unknown protein, they cannot rapidly judge the protein's function based on the structural characteristics. Therefore, we hope to use machine learning methods to predict diabetes protein markers.

Table 1. Using machine learning methods to predict protein.

Authors	Feature extraction method	Classifier
Feng *et al.* [23]	amino acid composition and frequency of occurrence of each dipeptide	Naive Bayes
Ding *et al.* [24]	g-gap	SVM
Song *et al.* [25]	188-dimensional features	Ensemble classifier
Yuan *et al.* [26]	the structural information between amino acids	RBF network
Chou *et al.* [27]	pseudo-amino acid information	augmented covariant-discriminant algorithm
Liu *et al.* [28]	physicochemical distance transformation (PDT)	SVM
Zhou *et al.* [29] and Tian *et al.* [30]	fusing Pse-ACC with the dipeptide composition and auto-covariance	SMOTE
Han *et al.* [37]	Physicochemical Properties	two-stage multiclass support vector machine
Bahri *et al.* [38]	-	Greedy-Boost
Chen *et al.* [39]	188-dimensional features	Mulit-classifiers and k-means
Wang *et al.* [40]	PSSM-SPF and RER features	ensemble random forests

Machine learning has been widely used in protein classification. Machine learning methods can build models based on known proteins, which can make function predictions for unknown proteins faster. For example, Feng *et al.* [23] used the amino acid composition and frequency of occurrence of each dipeptide to extract certain features and used Naive Bayes as the classifier to predict samples. Ding *et al.* [24] used the g-gap method to extract features from protein sequences. They used the support vector machine (SVM) to classify protein sequences. Song *et al.* [25] converted protein sequences into 188-dimensional features according to their composition, physical and chemical properties and distribution. Yuan *et al.* [26] extracted features according to peptides. This method considered the structural information between amino acids and obtained comprehensive and informative characteristics. The pseudo-amino acid information proposed by Chou *et al.* [27] contained the sequence information of two amino acids separated by one or more amino acid residues, and this method obtained good accuracy. Liu *et al.* [28] proposed an enhanced method called pseudo amino acid composition (Pse-ACC). In this method, the authors reduced the amino acid alphabet profile, proposing physicochemical distance transformation (PDT), which is similar to Pse-ACC. Zhou *et al.* [29] and Tian *et al.* [30] enhanced feature extraction of protein sequences by fusing Pse-ACC with the dipeptide composition and auto-covariance, encoding proteins based on their grouped weight. In addition to research on feature extraction methods, reconstruction of classifiers is also currently a research hotspot, particularly research on ensemble learning. Several previous studies established a series of random forest models based on different features [31–36]. They ultimately obtained results by voting on the results of each classification. Han *et al.* [37] constructed a two-layer multi-classification support vector machine model to predict subcellular localization. The output of the first layer is the input of the second layer. Bahri *et al.* [38] built an ensemble method called Greedy-Boost. This method not only improves stability but also enhances the speed of classification. Lin *et al.* [39] used 18 classifiers to predict protein sequences, using K-means to cluster the results. Wang *et al.* [40] predicted protein-protein interaction sites using an ensemble random forests with synthetic minority oversampling technique. We use Table 1 (Ref. [23–30, 37–40]) to summarize the above methods.

We are committed to building a predictive model for diabetes protein markers. In this way, it is possible to discover the diabetes protein markers whether contained in the human body or not, and we can label the function of unknown proteins. In this study, we focused on building a diabetes protein markers predictive model, which is based on machine learning.

3. Materials and methods

To build an efficient classification model, the following steps are required. First, the protein sequences must be converted into vectors. Then, the dimensionality of the feature vectors are reduced, if necessary. Finally, the classification model is obtained by training the classifier. In this study, we developed a feature extraction method and a classifier. We used an ensemble method to predict diabetic protein markers. First, we obtained positive data from Uniprot, and then, we obtained negative data from the positive data. Next, we used three methods to extract the features and six classifiers to predict proteins. Then, we obtained four new feature sets by combining three feature extraction methods. Finally, we used a dimensionality reduction method to reduce the features that were obtained in the previous step. In the classification experiment for each step, we used ensemble classifiers and traditional classifiers. The process flow chart is shown in Fig. 1.

3.1 Dataset

Due to the low number of diabetic marker proteins currently available, it is very important to build a representative and non-redundant negative dataset. In this study, we used the protein family database (PFAM) based on struc-

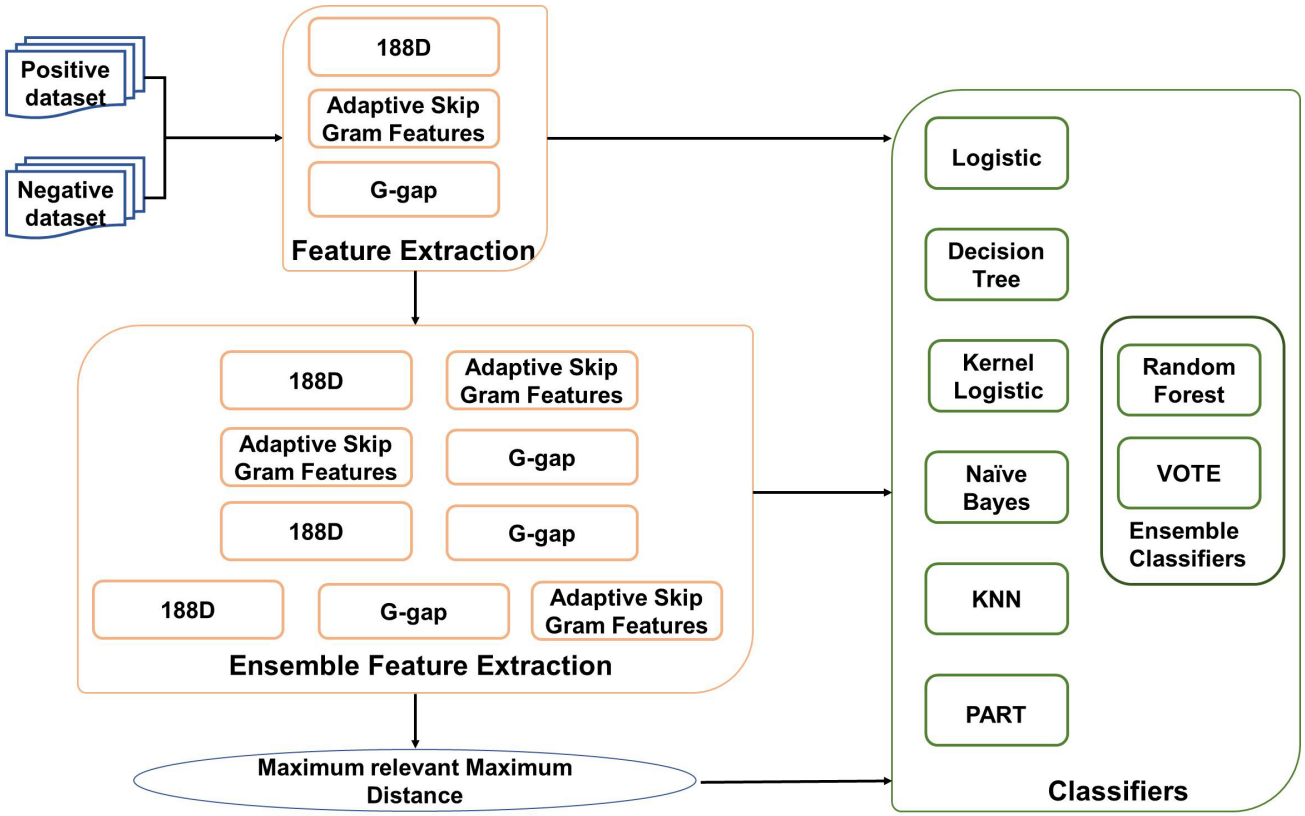

Fig. 1. Overall process of the method described in this paper.

tural information to build a negative dataset according to two principles [41, 42]: (1) extract all positive PFAM information and then choose the longest sequence as the negative sequence among the rest of all positive PFAM members; (2) the positive sequence is derived from the Uniprot Database.

Using 'diabetes' as the key word, 574 sequences were extracted from UniProt (Universal Protein, http://www.uniprot.org/uniprot/), containing human, mouse, cow and other species' protein data. After screening, we obtained 310 human protein sequences, and then, we used CD-HIT [43] to reduce redundant data and removed sequences that contained illegal letters. We were left with 309 diabetic protein markers and 9695 negative protein sequences. Because the data set was unbalanced, we randomly selected a negative dataset according to the positive samples' length and proportion. We randomly selected 5 sets of negative samples and averaged the results of 5 experiments using these 5 sets.

3.2 Feature extraction method

Since a computer cannot directly recognize a protein sequence, it needs to convert the sequence into a set of vectors to recognize the information, which is called feature extraction [44–52]. A good feature extraction method will comprehensively consider the information contained in the protein sequence. As we know, protein is composed of 20 amino acids arranged in combination, so we reflect the properties of proteins through the position of amino acids, physical and chemical properties, etc. At present, feature extraction methods mainly take into account the following aspects: (1) amino acid composition, (2) amino acid physicochemical information, (3) intrinsic correlation information of amino acid sequences and (4) structural information of proteins. Feature extraction methods have a strong influence on experimental results. Therefore, how to enhance the feature extraction method is a problem worth studying.

3.2.1 Composition and physicochemical features

Composition and physicochemical features (188D) can extract 188 features containing composition, transform and distribution information [53]. This method includes three parts: composition, transform and distribution [54, 55]. The first section is the amino acid (AA) composition. By calculating the frequencies of amino acids in a protein sequence, sequences are converted into 20D vectors [53]:

$$(v_1, v_2, v_3, \cdots, v_{20})^T = \left(\frac{n_1}{L}, \frac{n_2}{L}, \cdots, \frac{n_{20}}{L}\right)$$

where n_i represents the quantity of an AA in the protein sequence.

The second section is transform. According to the physicochemical properties of a protein, 20 AAs can be divided into 3 different groups. The proportion of each group

in the protein sequences is calculated. We use the secondary structure as an example:

$$C_j = \frac{\text{count}_{D_i}}{L}(i = 1, 2, 3; \text{j} = 1, 2, \cdots, 8)$$

where D_i represents the number of each kind of amino acids and L is the length of the sequence. In this study, we considered [56] eight physicochemical properties. For each physicochemical property, there are three types of amino acids. Therefore, we can obtain 24D features.

The third section is distribution [55, 57]. We calculate the distribution at five positions: at the beginning, 25%, 50%, 75% and the end. We obtained 120D features in this section. Then, we considered the number of different amino acid dipeptides. According to this step, we obtained the 24D features. The formulas are as follows.

$$T_{i,j} = \frac{D_i D_j \text{ or } D_j D_i}{L - 1}$$
$$i, j \in \{(i = 1, j = 2), (i = 2, j = 3), (i = 3, j = 1)\}$$

$$\text{D} = \frac{H_{ij}}{L}$$
$$(j = \text{ begining }, 25\%, 50\%, 75\%, \text{ ending }; i = 1, 2, 3)$$

where the chain length is measured as H_{ij} at the beginning, 25%, 50%, 75% and end of AAs at which a particular property is located.

This method divides amino acids into three types according to different physical and chemical properties, and considers the position information of three different types of amino acids under different physical and chemical properties. This method comprehensively considers the location information and physical and chemical properties of amino acids. The method is simple and easy to understand. Although the method uses physical and chemical properties, it still mainly focuses on the position information of amino acids, and the physical and chemical properties are not further reflected.

3.2.2 G-gap

G-gap dipeptide composition is a method used to describe the information about the composition of dipeptides in protein sequences. In this study, we used an enhanced method proposed by Huan *et al.* [56], who added a pseudo amino acid composition to the g-gap. Thus, each protein sequence is converted into $400 + n\lambda$ vectors [56, 58, 59].

$$FV_{g-gap} = [x_1, x_2, \cdots, x_{400}, x_{400+1}, \cdots, x_{400+n\lambda}]^T$$

$$x_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{400} f_i + \omega \sum_{j=1}^{n\lambda} \tau_i} (1 \le u \le 400) \\ \frac{\omega \tau_u}{\sum_{i=1}^{400} f_i + \omega \sum_{j=1}^{n\lambda} \tau_i} (400 + 1 \le u \le 400 + n\lambda) \end{cases}$$

$$f_u = \frac{n_i^g}{L - g - 1}$$

where the number of occurrences of *i*-th dipeptide appearances is denoted by n_i^g and ω is the weight. λ is the number of total counted ranks or tires of the correlations along a protein sequence, and *n* is the number of physicochemical properties used in this study. τ_i is the *i*-th tier correlation factor, which reflects the sequence-order correlation between all the *i*-th most contiguous dipeptides along a protein sequence. In this study, *n* is 9 and *g* is 2. Finally, we obtained the 670D features.

This feature extraction method is based on two amino acids. This method first considers the position of each amino acid in the sequence, and fuses the physical and chemical properties of the amino acid through the pseudo amino acid composition.

3.2.3 Adaptive Skip Gram Features

Adaptive Skip Gram Features (hereafter referred to as 400D) can extract 400 features. This method extracts features according to the distance between amino acids [55, 60, 61]. We assume a given protein sequence *P*. *P* is expressed as follows.

$$A_1 A_2 A_3 \cdots A_n$$

$\text{DT}(A_i, A_j)$ is the distance between amino acids.

$$\text{DT}(A_i, A_j) = j - i - 1$$

where *i* and *j* represent the position of amino acids. According to the definition of amino acid distance, if two amino acids are adjacent, the distance is 0. The maximum distance between amino acids is *L*-2. The k-skip-n-gram algorithm counts the frequency of occurrence of any *n* amino acid sequences in the sequence, expressed as follows:

$$FV_{\text{skipgram}} = \left\{ \frac{N(a_{m_1} a_{m_2} \cdots a_{m_n})}{N(T_{\text{skipgram}})} \middle| 1 \le m_1 \le 20, \cdots, 1 \le m_n \le 20 \right\}$$

$$T_{\text{skipgram}} = \left\{ \bigcup_{a=0}^{k} Skip(DT = a) | a = 0, 1, 2, \cdots, k; k \le \frac{L^{\min}}{n - 1} \right.$$

T_{skipgram} represents subsequences, which are composed of *n* amino acids in the sequence. $N(T_{skipgram})$ represents the number of elements in $T_{skipgram}$. $N(a_{m_1} a_{m_2} \cdots a_{m_n})$ represents the number of occurrences of all *n* amino acid component sequences in $T_{skipgram}$. The number of $FV_{skipgram}$ is 20^n. Due to the

differential value of *k*, Wei *et al.* [61] proposed the adaptive skip-n-gram model. This model cancels the limitation of *k*. The values of *k* are adaptive according to the length of the sample sequence, which makes the features contain more distance information and makes the k-skip-n-gram model have no parameters, avoiding the overfitting problem. In this study, *n* is 2, so we obtained 400D features using this method.

This method is derived from the n-gram model in natural language processing. It mainly considers whether each amino acid or each polypeptide appears in the sequence and how often it appears. So, this method take into account the distance between amino acids.

3.3 Classifier

3.3.1 Logistic regression

Logistic regression [62, 63] is a logarithmic model, and its form is a parametric logistic distribution represented by the conditional probability distribution $P(Y|X)$.

$$\mathrm{P}(Y=1 \mid x)=\frac{\exp(\omega \cdot x)}{1+\exp(\omega \cdot x)}$$

$$\mathrm{P}(Y=0 \mid x)=\frac{1}{1+\exp(\omega \cdot x)}$$

where $x \in R^n$ is the feature, $Y \in \{0,1\}$ is the class, and ω is the weight vector. For a given sample, we should calculate both probabilities.

In this study, we used two kinds of logistic regression models. One is shown above. The other is kernel logistic regression. LR is a linear classifier. Therefore, kernel logistic regression (KLR) [64] was proposed, which can be used to classify nonlinear data. KLR uses kernel functions to project the features into high-dimensional space.

3.3.2 Decision tree

Two kinds of decision trees are used in this study. One is C4.5 [12, 65], and the other is PART [66], which is based on C4.5.

(1) C4.5. This model uses the tree structure to describe the process of classification. C4.5 contains nodes and directed edges. There are two kinds of nodes, internal nodes and leaf nodes. Internal nodes indicate attributes, which are the conditional basis of classification. Leaf nodes are classes, which are the labels of samples. The process of C4.5 classifying instances is that C4.5 arranges samples from the root node to a leaf node. Feature selection is based on the information gain ratio.

(2) PART. PART is based on C4.5 and the partial decision tree, proposed by Eibe Frank *et al.* [67] in 1998. PART extracts rules from a dataset according to an incomplete decision tee. The original principle of the algorithm comes from the separate-and-conquer strategy. This method creates a rule and then removes the instance that is covered by the rule. The method builds rules for the remaining instances until there are no existing instances, recursively.

3.3.3 Naïve Bayes

Naïve Bayes [23, 53] is a classifier based on Bayes' theorem and features condition independent hypotheses. In this algorithm, first, the input's and output's joint probability distribution are studied, which is based on the independent hypothesis of feature conditions; then, for a given input *x*, we used Bayes' theorem to calculate the maximum posterior probability of output *y*. Naive Bayes is a common classification method, which is easy to implement [23, 66].

3.3.4 K-Nearest Neighbor

K-Nearest Neighbor (KNN) [68, 69] is a basic classification and regression method. In this study, we only discuss the application of KNN in classification problems. KNN contains three important factors: the selection of *k* values, distance functions and decision rules. The algorithm for KNN is as follows:

(1) According to the determined distance function, we can derive *k* points that are closed to instance *x* in the training set. The neighborhood of *x* that covers these *k* points is called $N_k(x)$;

(2) In $N_k(x)$, the class of *x* is determined according to the classification decision rule.

$$\mathrm{y}=\arg\max_{c_j} \sum_{x_i \in N_k(x)} I\left(y_i=c_j\right),$$
$$i=1,2,3,\ldots,N; j=1,2,3,\ldots,K$$

where x_i is the feature vector and $y_i=\{c_1,c_2,\ldots,c_K\}$ is the label. $I(*)$ is the indicator function.

3.4 Ensemble method

The ensemble method uses many kinds of methods to process a dataset, which may obtain superior results [7, 70–77]. Different methods have different emphases on data processing. We combined different methods to improve the classification efficiency. In this study, we mixed feature extraction methods and used ensemble classifiers to improve their performances.

3.4.1 Ensemble feature extraction methods

We used the 188D, g-gap and 400D to extract features from sequences. These methods have different emphases. 188D contains the amino acid composition and physical and chemical properties. G-gap contains the dipeptide composition, which can indicate the importance of the peptide chain. Since proteins are produced by the distortion and folding of peptide chains, dipeptides can better recognize proteins. G-gap adds nine kinds of physical

and chemical properties to improve its accuracy. 400D features take into account the distance between amino acids. Therefore, it is meaningful to combine these three methods to more comprehensively extract features and improve accuracy.

In this study, we used four combination methods. First, we combined 188D and g-gap and obtained 858D features. 188D divides amino acids into three groups and studies the physical and chemical properties of each type of amino acid. However, g-gap considers the properties of each amino acid.

Second, we combined 188D and 400D and obtained 588D features. Since 400D does not consider the physicochemical properties of amino acids, the combination of 188D and 400D extracts features based on the AA composition and physicochemical properties.

Third, we combined g-gap and 400D and obtained 1070D features. 400D is different from g-gap. The dipeptides in g-gap are adjacent, while in 400D, two amino acids can be separated by several amino acids, that is, the two amino acids considered are not adjacent. G-gap focuses on the composition of dipeptides in the protein sequence, and 400D focuses on the relative position of the amino acids.

Fourth, we combined 188D, g-gap and 400D and obtained 1258D features. 188D considers the frequency of occurrence of a single amino acid and the amino acid's position according to its physical and chemical properties, which are different from the other two methods. G-gap focuses on the composition of dipeptides in the protein sequence, which is different from the other two methods. 400D focuses on the distance between two amino acids, which is different from the other methods.

Ensemble feature extraction methods can make up for the shortcomings between different methods, and then can construct a set of comprehensive features. By combining feature extraction methods, we can get four different sets of feature vectors.

3.4.2 Ensemble classifiers

Ensemble classifiers can organically combine multiple traditional learning models to obtain more stable and accurate results. There are three common ensemble learning algorithms, including bagging, boosting and stacking.

In this study, we used random forest and vote as the ensemble learning methods to classify proteins.

(1) Random Forest (RF) [78–82]. Random forest is an extension of bagging. Leo Breiman proposed RF. RF is composed of many decision trees, with no correlation between different decision trees. When we need to classify a sample, each decision tree in the forest makes a judgment and classification. The final result is the class with the highest number of votes.

(2) Vote. This method classifies the sample according to the seven classifiers. First, we use LR, KLR, NB, DT, PART, RF and KNN to classify protein sequences. We obtain seven results, and we use majority voting to obtain the final result. If a label receives more than half the votes, the prediction is that label; otherwise, the prediction is rejected.

$$\mathrm{H}(\mathbf{x}) = \begin{cases} c_j, if & \sum_{i=1}^{T} h_i^j(\boldsymbol{x}) > 0.5 \sum_{k=1}^{N} \sum_{i=1}^{T} h_i^k(\boldsymbol{x}) \\ reject, & otherwise \end{cases}$$

where $h_i^j(x)$ is the output of classifier h_i on the label c_j. T represents the number of base classifiers, and N represents the number of labels.

3.5 Measurement

In this study, we used accuracy (ACC), the Matthews correlation coefficient (MCC), F-Measure and the area under the receiver operating characteristic curve (AUC) to measure classifier efficacy [83, 84]. The formulas are as follows:

$$\mathrm{ACC} = \frac{TP + TN}{TN + TP + FN + FP}$$

$$\mathrm{MCC} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

$$\mathrm{F1} = \frac{2 \times \frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}}$$

where TP represents the number of correct classifications in the positive dataset. TN is the number of correct classifications in the negative dataset. FN is the number of false negatives. FP is the number of false positives.

4. Result and discussion

Due to the imbalanced dataset, we randomly extracted 5 sets of negative samples and averaged the results of 5 experiments using these 5 sets. Each experiment was subjected to 10-fold cross-validation. The dataset was divided into 10 sections. Nine groups were used to train the model, and the remaining group was used to test the model.

4.1 Using the single feature extraction method and a single classifier

To evaluate the ensemble methods, first, we used the single feature extraction method and traditional classifier to predict proteins. When we used 188D to extract features from the protein sequences, the performances of the six classifiers had negligible differences. The best AUC from KLR was 0.81, and the worst AUC from KNN was 0.70. When 400D was used to extract features, the best clas-

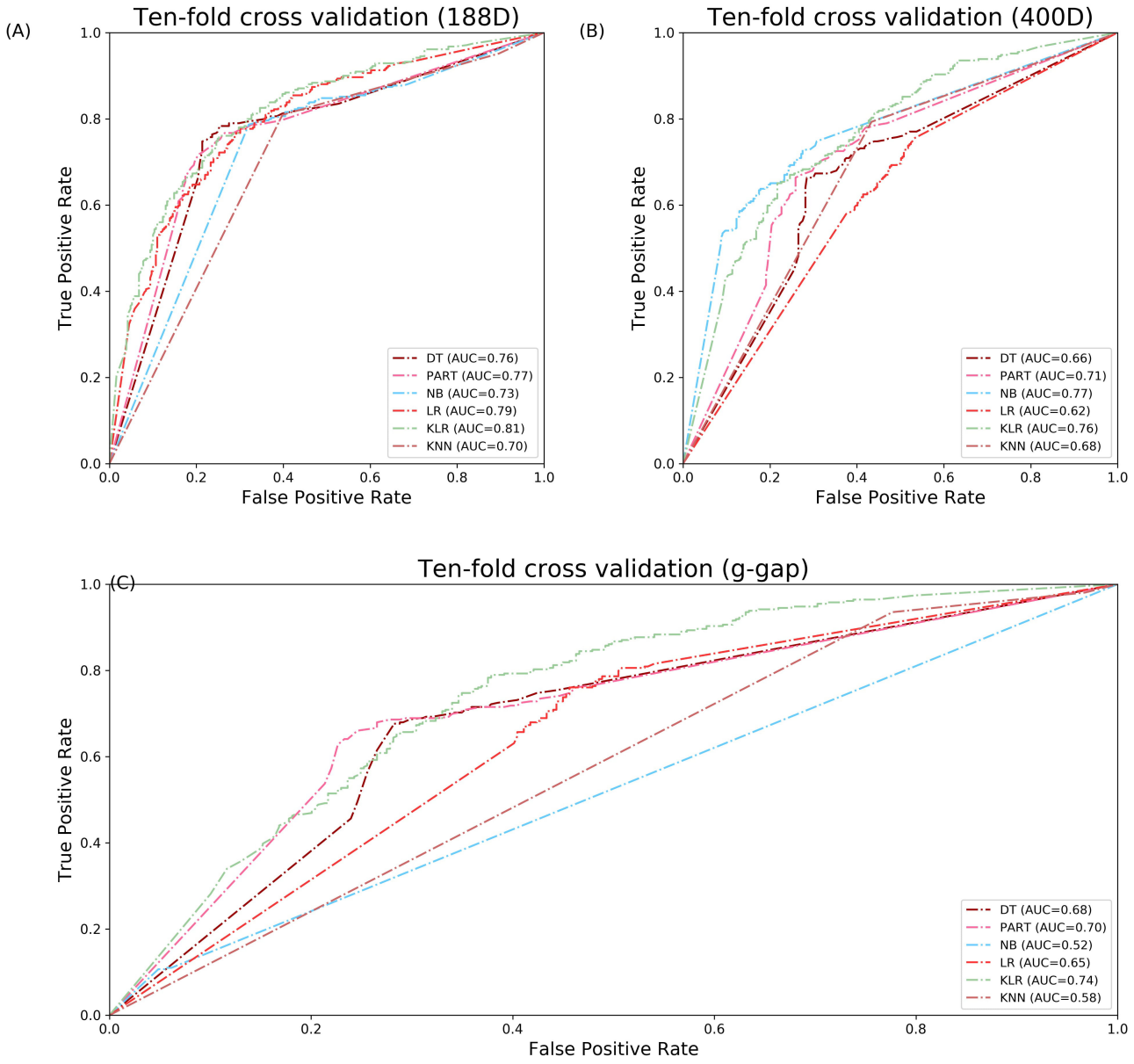

Fig. 2. Performances comparison of single feature extraction method. (A) The ROC curve of 188D with six classifiers. (B) The ROC curve of g-gap with six classifiers. (C) The ROC curve of adaptive skip gram feature with six classifiers.

sifier was NB, the AUC of which was 0.77, and the worst AUC was 0.66 from DT. The best AUC of g-gap method was 0.74 from KLR, and the worst AUC was 0.52 from NB. The detailed classification results are shown in Fig. 2. The overall effect of KLR was the best. KLR can map nonlinear features to high-dimensional space by adding kernel functions, which solves nonlinear problems. NB, which is the best classifier for 400D, but the worst for g-gap, assumes that each feature is independent with an identical distribution, so the classification effect of the different feature extraction methods fluctuates greatly.

In the previous section, we evaluated the classifiers according to the ROC curve and AUC. Next, we evaluated the feature extraction methods, as shown in Fig. 3, which are more vivid. According to Fig. 3, 188D has the best performance among the five classification results, except for NB. 400D has the best performance among the three feature extraction methods when the classifier is NB. The performance of NB indicates that features obtained using 400D have the highest independence among the three feature extraction methods. According to the DT and PART results, 188D has the best performance, which may indicate that features extracted from 188D contain more effective information for classifying diabetic protein markers. All of the experimental results using the single methods are shown in Appendix Table 5.

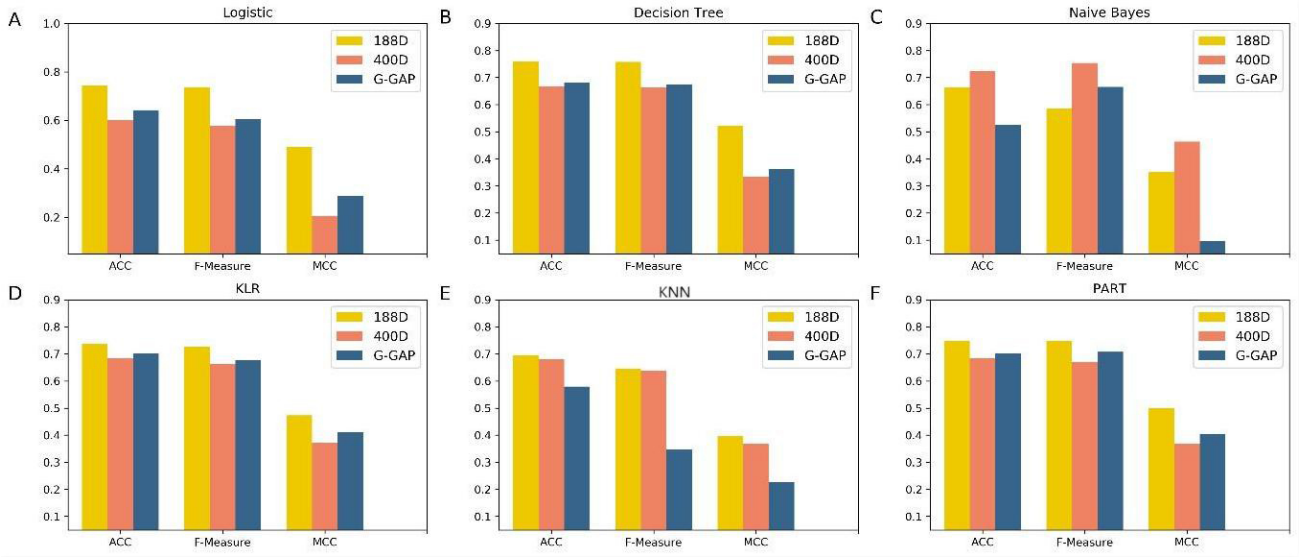

Fig. 3. The results of using single feature extraction method and traditional classifier. Compare the feature extraction methods by controlling the classifier.

4.2 Comparison of the ensemble methods with single methods

4.2.1 Ensemble feature extraction methods outperform the single methods

In this section, we compare the joint features with single ones. We combined the three feature extraction methods, obtaining four joint features: 588D (combining 188D with 400D), 858D (combining 188D with g-gap), 1070D (combining 400D and g-gap) and 1258D (combining 188D with 1070D). In this section, we also use six classifiers for prediction.

First, we conducted the classification experiment on 588D. To make the comparison of the results clearer, we used the DT, NB and PART experimental results to create a histogram, which is shown in Fig. 4. When the classifier is DT, 188D has the best result. NB and PART were selected for similar reasons. 400D had the best result with NB, and 588D had the best result with PART. The experimental results of the other classifiers are shown in Appendix Table 6. According to Fig. 4, 588D is the best feature extraction method among the three methods, except for DT. 588D improves accuracy most of the time, but it is slightly worse than 188D when the classifiers are DT and LR.

Similar to the above method, we created histograms of the remaining three combined methods, which are shown in Fig. 5. According to Fig. 5, we found the accuracy was generally improved, but the improvement rate was not large. Specifically, 1258D had little improvement compared to 588D and 1070D. However, when the classifier was NB, the performance of 1070D was worse than 400D, potentially because the ensemble method increases the correlation between features and reduces feature independence. We used Max-revelation-Max-Distance (MRMD) to reduce the dimensionality, which may improve the accuracy. This method used the Pearson correlation coefficient (PCC) to calculate the relevance and the Euclidean distance to identify instances of redundancy. The results are shown in Appendix Table 7. Compared to the results without dimensionality reduction, the effect of the classifiers increased and decreased. When the classifiers are LR, DT and NB, the results improved. Compared to the single method, overall accuracy was improved. Therefore, using the ensemble feature method is better than the single feature extraction method.

4.2.2 Ensemble classifiers outperform single classifiers

In this section, we compare the ensemble classifiers with traditional classifiers. We use RF and VOTE as the ensemble classifiers. RF is an ensemble classifier based on the bagging algorithm, using DT as the base learner. RF combines all the classification results for voting and designates the label with the most votes as the final result. We proposed the vote method. In this method, we used seven classifiers to classify the samples, obtaining the final result according to majority voting. The seven classifiers are DT, PART, KLR, LR, KNN, RF and NB, which were used in the previous sections. The results are shown in Fig. 6. There are three ROC curves in each subgraph. Two of them are RF and VOTE, and the other is the traditional classifier with the best performance. From Fig. 6, we observe that the results using ensemble classifiers are better than the results using traditional classifiers. The best AUC was 0.90, for which the classifier was RF and the feature extraction method was 188D. Moreover, the worst AUC in the section was 0.70, for which the classifier was PART and the feature extraction method was g-gap. According to Fig. 6, RF is superior to VOTE. The reason the ensemble method is better

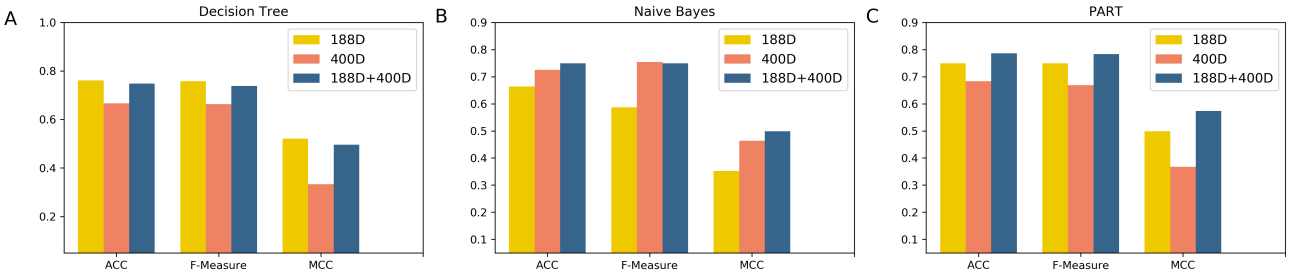

Fig. 4. The results of combining 188D and 400D. Three classifiers are selected for comparison. (A) When classifier is DT, 188D has the best performance. (B) When feature extraction method is 400D, using NB can have the best performance, but performance of 588D is better than 400D. (C) When classifier is PART, 588D has the best performance.

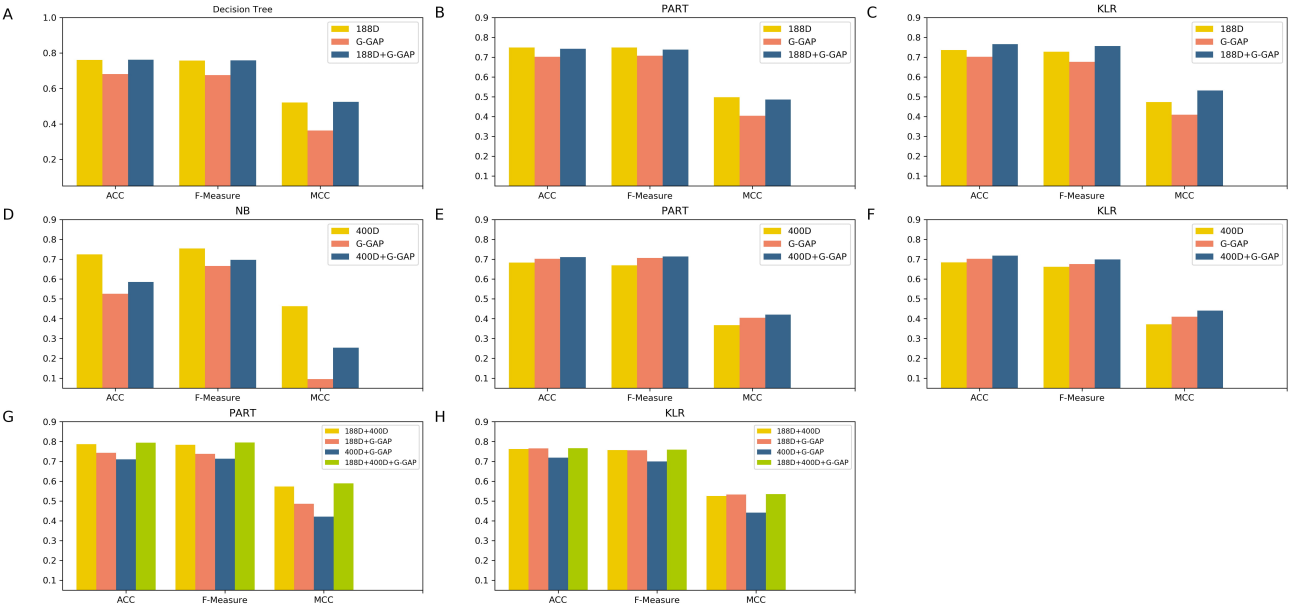

Fig. 5. Results of using ensemble feature extraction method. In this section, we used traditional classifiers. (A), (B), (C) are the results of 188D, g-gap and 858D. (D), (E), (F) are the results of 400D, g-gap and 1070D. (G), (H) are the results of 588D, 858D, 1070D, and 1258D.

Table 2. The results of using ensemble classifiers with single feature extraction method.

Method	Classifier	ACC	F_measure	MCC
188D	RF	0.8139	0.8255	0.6334
188D	VOTE	0.8042	0.8013	0.6087
400D	RF	0.7718	0.7800	0.5452
400D	VOTE	0.7573	0.7541	0.5147
g-gap	RF	0.7977	0.8109	0.6013
g-gap	VOTE	0.7686	0.7682	0.5372

than the traditional classification method is that the ensemble method avoids the accidental errors of single methods by comprehensively considering multiple classifiers. All experimental results are shown in Table 2.

According to the results, we found ensemble classifiers are better than single classifier. Ensemble classifiers are beneficial in three ways. First, since the learning task has a large hypothesis space, there may be many hypotheses in the training set to achieve the same performance. Therefore, the single classifier may choose the hypothesis space by mistake, resulting in poor generalization. Ensemble classifiers can reduce this risk. Second, ensemble classifiers can reduce the risk of falling into a terrible local minimum. Third, by combining multiple classifiers, the corresponding hypothesis space will expand, making it possible to learn the best approximation.

4.3 Ensemble classifiers with a combined feature extraction method have the best performance

According to the above results, we know that when the classifier is traditional, the ensemble extraction method is better than a single method, and ensemble classifiers are better than traditional classifiers when a single feature extraction method is used. Therefore, in this section, we discuss the performance, which used ensemble classifiers and ensemble extraction methods.

In this section, we created a histogram according to the RF, VOTE and traditional classifiers results, as

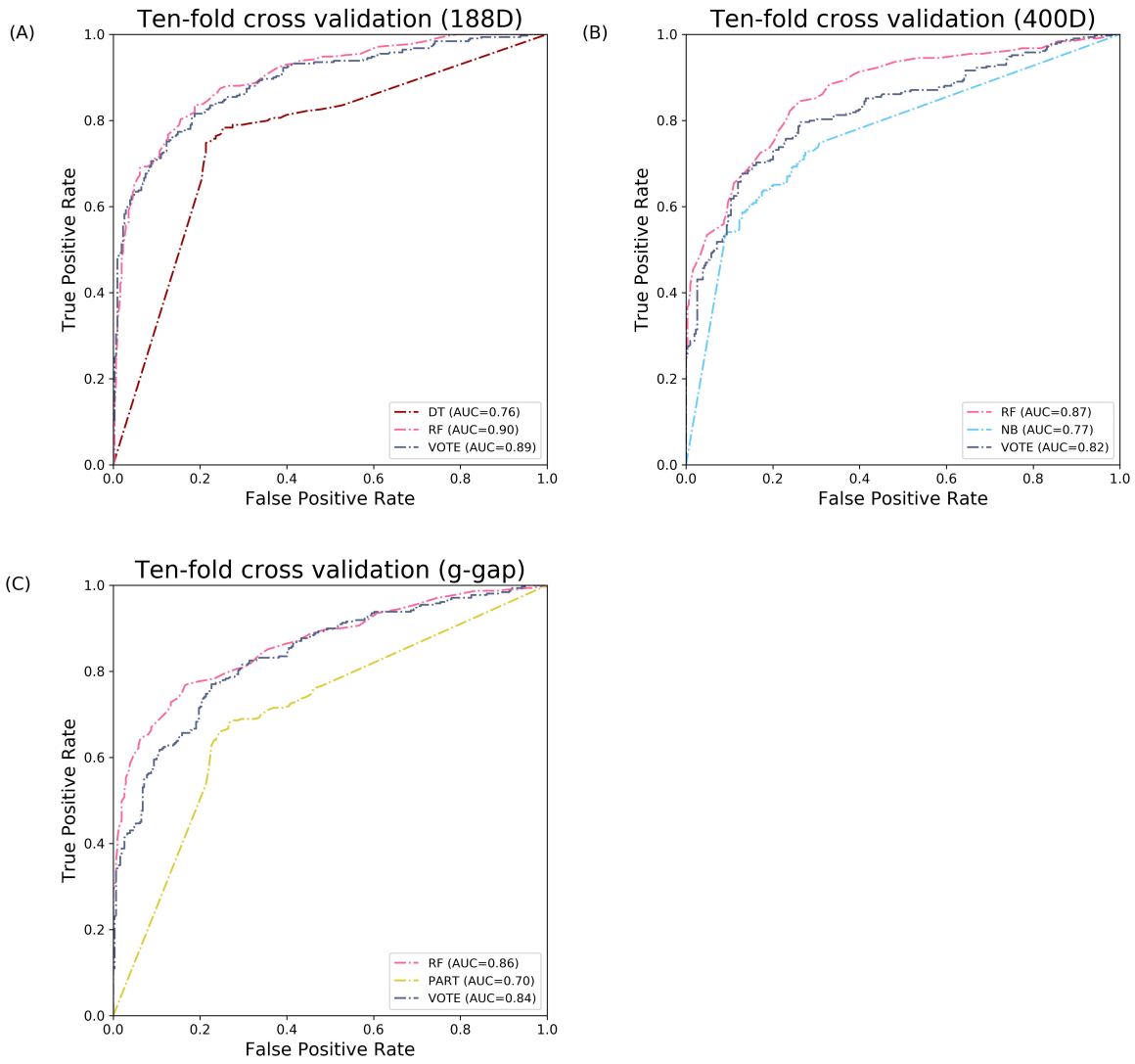

Fig. 6. Performance comparison of ensemble classifiers and traditional classifier. (A) The ROC curve of the ensemble classifiers and DT on 188D. (B) The ROC curve of the ensemble classifiers and NB on g-gap. (C) The ROC curve of the ensemble classifiers and PART on 400D.

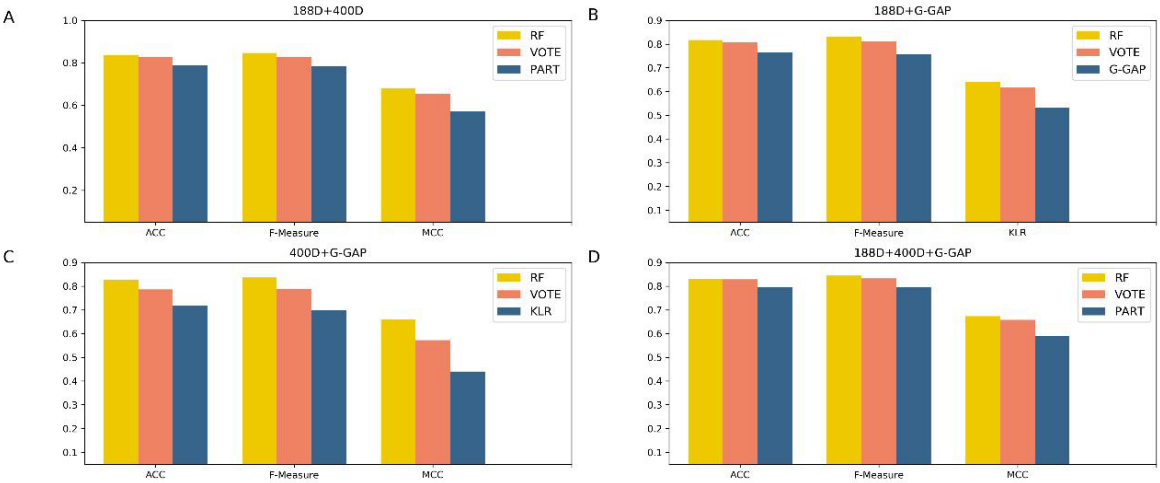

Fig. 7. Performance comparison of ensemble classifiers and traditional classifier. In this section, we used combined feature extraction methods.

Table 3. The results of using ensemble classifiers with combined feature extraction method.

Method	Classifier	ACC	F_measure	MCC
188 + 400D	RF	0.8366	0.8463	0.6786
188 + 400D	VOTE	0.8269	0.8277	0.6538
188 + g-gap	RF	0.8172	0.8296	0.6411
188 + g-gap	VOTE	0.8074	0.8120	0.6156
400 + g-gap	RF	0.8269	0.8366	0.6585
400 + g-gap	VOTE	0.7864	0.7871	0.5728
188 + 400 + g-gap	RF	0.8317	0.8448	0.6730
188 + 400 + g-gap	VOTE	0.8285	0.8323	0.6576

Table 4. The results of using ensemble classifiers with combined feature extraction method after dimensionality reduction.

Method	Classifier	ACC	F_measure	MCC
188 + 400D (reduced)	RF	0.8285	0.8374	0.6610
188 + 400D (reduced)	VOTE	0.8123	0.8135	0.6246
188 + g-gap (reduced)	RF	0.8155	0.8262	0.6359
188 + g-gap (reduced)	VOTE	0.7961	0.7886	0.5937
400 + g-gap (reduced)	RF	0.8139	0.8250	0.6329
400 + g-gap (reduced)	VOTE	0.7913	0.7923	0.5826
188 + 400 + g-gap (reduced)	RF	0.8333	0.8451	0.6745
188 + 400 + g-gap (reduced)	VOTE	0.8220	0.8308	0.6475

shown in Fig. 7 and Table 3. The selected traditional classifier had the best performance among the six classifiers. According to Fig. 7, we found the ensemble classifiers were better than traditional classifiers when we use combined feature extraction methods. In this section, RF was better than VOTE. This conclusion is the same as for Section 4.2.1. When we used MRMD, the effect was not improved. The results are shown in Table 4.

Table 5. The results of using three feature extraction methods.

Method	Classifier	ACC	F_measure	MCC
188D	DT	0.7605	0.7574	0.5212
188D	NB	0.6634	0.5873	0.3517
188D	LR	0.7443	0.7367	0.4895
188D	KLR	0.7362	0.7279	0.4734
188D	KNN	0.6957	0.6459	0.3955
188D	PART	0.7492	0.7488	0.4984
400D	LR	0.6019	0.5759	0.2054
400D	NB	0.7249	0.7543	0.4633
400D	DT	0.6667	0.6634	0.3334
400D	KLR	0.6845	0.6620	0.3722
400D	PART	0.6828	0.6689	0.3670
400D	KNN	0.6796	0.6387	0.3688
g-gap	DT	0.6812	0.6755	0.3627
g-gap	LR	0.6424	0.6061	0.2898
g-gap	NB	0.5259	0.6659	0.0949
g-gap	KNN	0.5793	0.3467	0.2258
g-gap	PART	0.7023	0.7070	0.4047
g-gap	KLR	0.7023	0.6761	0.4099

Table 6. The results of using ensemble feature extraction methods.

Method	Classifier	ACC	F_measure	MCC
188 + 400D	NB	0.7492	0.7496	0.4984
188 + 400D	LR	0.6440	0.6393	0.2881
188 + 400D	DT	0.7476	0.7383	0.4964
188 + 400D	KLR	0.7621	0.7570	0.5247
188 + 400D	PART	0.7864	0.7836	0.5730
188 + 400D	KNN	0.7136	0.6788	0.4376
188 + g-gap	LR	0.6197	0.6010	0.2405
188 + g-gap	NB	0.5340	0.6705	0.1214
188 + g-gap	KNN	0.6327	0.4829	0.3256
188 + g-gap	PART	0.7427	0.7381	0.4857
188 + g-gap	KLR	0.7654	0.7563	0.5322
188 + g-gap	DT	0.7621	0.7586	0.5245
400 + g-gap	NB	0.5858	0.6974	0.2541
400 + g-gap	LR	0.6068	0.5714	0.2166
400 + g-gap	DT	0.7071	0.6917	0.4163
400 + g-gap	KLR	0.7184	0.6990	0.4406
400 + g-gap	KNN	0.6489	0.5373	0.3399
400 + g-gap	PART	0.7104	0.7136	0.4208
188 + 400 + g-gap	DT	0.7654	0.7619	0.5310
188 + 400 + g-gap	LR	0.6036	0.5769	0.2088
188 + 400 + g-gap	NB	0.6133	0.7131	0.3154
188 + 400 + g-gap	PART	0.7945	0.7955	0.5890
188 + 400 + g-gap	KLR	0.7670	0.7592	0.5351
188 + 400 + g-gap	KNN	0.6570	0.5583	0.3508

According to performance, the ensemble method is better than the single method. When the classifier is RF and the feature extraction is 588D, the performance is the best among all the methods. The second best ensemble method was 1285D with RF. Therefore, we can use 588D and RF to build the prediction model. All of the experimental results are shown in Appendix Table 6.

5. Conclusions

Diabetes is a common chronic disease. If diabetes is not detected and treated in time, it can lead to serious complications. In this study, we conducted research on diabetic protein markers. By classifying proteins, we can determine whether there are diabetic protein markers in the human body that can be used to better diagnose diabetes.

In this study, we proposed using ensemble methods to predict diabetes protein markers, including ensemble feature extraction methods and ensemble classifiers. We used three feature extraction methods and six traditional classifiers. We combined three methods, obtaining four combined methods. We used seven classifiers to form an ensemble learning method. To validate the performance of our ensemble classifier, we evaluated and compared it with the traditional classifier using 10-fold cross validation.

According to the results, ensemble method is better than single method. We compared the combined features with existing features. The performance revealed that

Table 7. The results of using MRMD.

Method	Classifier	ACC	F_measure	MCC
188 + 400D (reduced)	NB	0.7573	0.7440	0.5173
188 + 400D (reduced)	LR	0.6780	0.6700	0.3564
188 + 400D (reduced)	KLR	0.7540	0.7556	0.5081
188 + 400D (reduced)	PART	0.7751	0.7702	0.5506
188 + 400D (reduced)	KNN	0.7136	0.6811	0.4363
188 + 400D (reduced)	DT	0.7621	0.7648	0.5244
188 + g-gap (reduced)	DT	0.7346	0.7320	0.4693
188 + g-gap (reduced)	LR	0.7282	0.7143	0.4585
188 + g-gap (reduced)	NB	0.6586	0.5772	0.3437
188 + g-gap (reduced)	KNN	0.6926	0.6520	0.3960
188 + g-gap (reduced)	PART	0.7249	0.7231	0.4499
188 + g-gap (reduced)	KLR	0.7330	0.7236	0.4671
400 + g-gap (reduced)	NB	0.5922	0.7000	0.2652
400 + g-gap (reduced)	LR	0.6133	0.5770	0.2299
400 + g-gap (reduced)	DT	0.7282	0.7191	0.4573
400 + g-gap (reduced)	KLR	0.7152	0.6966	0.4337
400 + g-gap (reduced)	KNN	0.6505	0.5365	0.3457
400 + g-gap (reduced)	PART	0.7071	0.7076	0.4142
188 + 400 + g-gap (reduced)	DT	0.7605	0.7613	0.5210
188 + 400 + g-gap (reduced)	LR	0.6683	0.6623	0.3368
188 + 400 + g-gap (reduced)	NB	0.7282	0.7742	0.4997
188 + 400 + g-gap (reduced)	PART	0.7913	0.7875	0.5829
188 + 400 + g-gap (reduced)	KLR	0.7443	0.7492	0.4890
188 + 400 + g-gap (reduced)	KNN	0.7168	0.6858	0.4424

the combined feature extraction method was more effective. Especially when the feature dimension is 588D and the classification method is random forest, the effect is best. Therefore, 588D features and random forest can be used to construct a model for predicting diabetes protein markers. Using machine learning methods can quickly predict protein function. 188D divides amino acids into three groups and studies the physical and chemical properties of each type of amino acid. Since 400D does not consider the physicochemical properties of amino acids, the combination of 188D and 400D extracts features based on the AA composition and physicochemical properties. Combining these two feature extraction methods, protein sequences can be analyzed in terms of physical and chemical properties, amino acid positions, amino acid fragments, etc. The ensemble feature extraction methods can analyze the sequence comprehensively, and the ensemble machine learning method can avoid many problems, e.g., poor generalization ability. According to the results, the ensemble method, either the combined feature extraction method or the ensemble classifier, was better than the single method. We anticipate that ensemble methods will be a useful tool for identifying diabetic protein markers in a cost-effective manner.

In this study, we only made predictions for diabetes marker proteins. We can use the model to predict protein. Due to lack of relevant data, we are temporarily unable to predict diabetes. Currently, the obtained diabetes marker proteins are not used for diabetes prediction. Therefore, in the next step of research, we will focus on using diabetes marker proteins to predict diabetes, which is more valuable in clinical applications.

6. Author contributions

KQ and HS implemented the experiments and drafted the manuscript. QZ and KQ initiated the idea, conceived the whole process, and finalized the paper. HS and QZ helped with data analysis and revised the manuscript. All authors have read and approved the final manuscript.

7. Ethics approval and consent to participate

Not applicable.

8. Acknowledgment

Not applicable.

9. Funding

The work was supported by the National Natural Science Foundation of China (No. 61922020, No. 61771331), and the Special Science Foundation of Quzhou (2020D003).

10. Conflict of interest

The authors declare no conflict of interest.

11. Appendix

See Tables 5,6,7.

12. References

[1] Gupta A, Behl T, Sehgal A, Sharma S, Singh S, Sharma N, *et al*. Unmasking the therapeutic potential of biomarkers in type-1 diabetes mellitus. Biointerface Research in Applied Chemistry. 2021. 11: 13187–13201.

[2] Giglio RV, Stoian AP, Haluzik M, Pafili K, Patti AM, Rizvi AA, *et al*. Novel molecular markers of cardiovascular disease risk in type 2 diabetes mellitus. Biochimica et Biophysica Acta (BBA). Molecular Basis of Disease. 2021; 1867: 166148.

[3] Shi H, Liu S, Chen J, Li X, Ma Q, Yu B. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. Genomics. 2019; 111: 1839–1852.

[4] Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Research. 2019; 47: e127.

[5] Xu Q, Xiong Y, Dai H, Kumari KM, Xu Q, Ou H, *et al*. PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm. Journal of Theoretical Biology. 2017; 417: 1–7.

[6] Zou Q, Li J, Song L, Zeng X, Wang G. Similarity computation strategies in the microRNA-disease network: a survey. Briefings in Functional Genomics. 2016; 15: 55–64.

[7] Xu L, Liang G, Liao C, Chen GD, Chang CC. k-Skip-n-Gram-RF: A Random Forest Based Method for Alzheimer's Disease Protein Identification. Frontiers in Genetics. 2019; 10: 33.

[8] Xu L, Liang G, Liao C, Chen G, Chang C. An Efficient Classifier for Alzheimer's Disease Genes Identification. Molecules. 2019; 23: 3140

[9] Cheng L, Jiang Y, Ju H, Sun J, Peng J, Zhou M, *et al*. InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. BMC Genomics. 2018; 19: 919.

[10] Cheng L, Hu Y, Sun J, Zhou M, Jiang Q. DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. Bioinformatics. 2018; 34: 1953–1956.

[11] Cheng L, Zhuang H, Yang S, Jiang H, Wang S, Zhang J. Exposing the Causal Effect of C-Reactive Protein on the Risk of Type 2 Diabetes Mellitus: A Mendelian Randomization Study. Frontiers in Genetics. 2018; 9: 657.

[12] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus with Machine Learning Techniques. Frontiers in Genetics. 2018; 9: 515.

[13] Mauvoisin D. Circadian rhythms and proteomics: it's all about posttranslational modifications! Wiley Interdisciplinary Reviews: Systems Biology and Medicine. 2019; 11: e1450.

[14] Vaudel M, Barsnes H, Ræder H, Berven FS. Using Proteomics Bioinformatics Tools and Resources in Proteogenomic Studies. Advances in Experimental Medicine and Biology. 2016; 422: 65–75.

[15] Puentes-Osorio Y, Amariles P, Calleja M, Merino V, Díaz-Coronado JC, Taborda D. Potential clinical biomarkers in rheumatoid arthritis with an omic approach. Autoimmunity Highlights. 2021; 12: 9.

[16] Fleischer JG, Schulte R, Tsai HH, Tyagi S, Ibarra A, Shokhirev MN, *et al*. Predicting age from the transcriptome of human dermal fibroblasts. Genome Biology. 2018; 19: 221.

[17] Rebouças DB, Sartori JM, Librenza-Garcia D, Rabelo-da-Ponte FD, Massuda R, Czepielewski LS, *et al*. Accelerated aging signatures in subjects with schizophrenia and their unaffected siblings. Journal of Psychiatric Research. 2021; 139: 30–37.

[18] Cheng L, Wang P, Tian R, Wang S, Guo Q, Luo M, et al. LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Research. 2019; 47: D140–D144

[19] Huth C, von Toerne C, Schederecker F, de Las Heras Gala T, Herder C, Kronenberg F, *et al*. Protein markers and risk of type 2 diabetes and prediabetes: a targeted proteomics approach in the KORA F4/FF4 study. European Journal of Epidemiology. 2019; 34: 409–422.

[20] Hirao Y, Saito S, Fujinaka H, Miyazaki S, Xu B, Quadery AF, *et al*. Proteome Profiling of Diabetic Mellitus Patient Urine for Discovery of Biomarkers by Comprehensive MS-Based Proteomics. Proteomes. 2018; 6: 9.

[21] Kim SM, Park JS, Norwitz ER, Lee SM, Kim BJ, Park C, *et al*. Identification of proteomic biomarkers in maternal plasma in the early second trimester that predict the subsequent development of gestational diabetes. Reproductive Sciences. 2012; 19: 202–209.

[22] Tao Y, Wu J, Chang L. Application of proteomics in diabetes and its complications. Journal of China Pharmaceutical University. 2020; 51: 368–373.

[23] Feng PM, Ding H, Chen W, Lin H. Naïve Bayes classifier with feature selection to identify phage virion proteins. Computational and Mathematical Methods in Medicine. 2013; 2013: 530696.

[24] Ding H, Feng P, Chen W, Lin H. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Molecular Omics. 2014; 10: 2229–2235.

[25] Song L, Li D, Zeng X, Wu Y, Guo L, Zou Q. NDNA-Prot: identification of DNA-binding proteins based on unbalanced classification. BMC Bioinformatics. 2014; 15: 298.

[26] Yuan L, Ding C, Guo S, Ding H, Chen W, Lin H. Prediction of the types of ion channel-targeted conotoxins based on radial basis function network. Toxicology in Vitro. 2013; 27: 852–856.

[27] Chou KC. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochemical and Biophysical Research Communications. 2000; 278: 477–483.

[28] Liu B, Xu J, Lan X, Xu R, Zhou J, Wang X, Chou KC. iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS ONE. 2014; 9: e106691.

[29] Zhou H, Chen C, Wang M, Ma Q, Yu B. Predicting Golgi-Resident Protein Types Using Conditional Covariance Minimization with XGBoost Based on Multiple Features Fusion. IEEE Access. 2019; 7: 144154–144164.

[30] Tian B, Wu X, Chen C, Qiu W, Ma Q, Yu B. Predicting protein–protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach. Journal of Theoretical Biology. 2019; 462: 329–346.

[31] Yang R, Zhang C, Gao R, Zhang L. An ensemble method with hybrid features to identify extracellular matrix proteins. PLoS ONE. 2015; 10: e0117804.

[32] Sun J, Shi H, Wang Z, Zhang C, Liu L, Wang L, *et al*. Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. Molecular BioSystems. 2014; 10: 2074–2081.

[33] Zhou M, Sun Y, Sun Y, Xu W, Zhang Z, Zhao H, *et al*. Comprehensive analysis of lncRNA expression profiles reveals a novel lncRNA signature to discriminate nonequivalent outcomes in patients with ovarian cancer. Oncotarget. 2016; 7: 32433–32448.

[34] Zhou M, Wang X, Li J, Hao D, Wang Z, Shi H, *et al*. Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. Molecular BioSystems. 2015; 11: 760–769.

[35] Zhou M, Zhao H, Wang X, Sun J, Su J. Analysis of long noncoding RNAs highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. Briefings in Bioinformatics. 2019; 20: 598–608.

[36] Zhou M, Zhao H, Wang Z, Cheng L, Yang L, Shi H, *et al*. Identification and validation of potential prognostic lncRNA biomarkers for predicting survival in patients with multiple myeloma. Journal of Experimental & Clinical Cancer Research. 2015; 34: 102.

[37] Han GS, Yu ZG, Anh V, Krishnajith APD, Tian Y. An ensemble method for predicting subnuclear localizations from primary protein structures. PLoS ONE. 2013; 8: e57225.

[38] Bahri E, Harbi N, Huu HN. Approach Based Ensemble Methods for Better and Faster Intrusion Detection. Computational Intelligence in Security for Information Systems. 2011; 53: 17–24.

[39] Lin C, Zou Y, Qin J, Liu X, Jiang Y, Ke C, *et al*. Hierarchical classification of protein folds using a novel ensemble classifier. PLoS ONE. 2013; 8: e56499.

[40] Wang X, Yu B, Ma A, Chen C, Liu B, Ma Q. Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. Bioinformatics. 2018; 35: 2395–2402.

[41] Zou Q, Wan S, Ju Y, Tang J, Zeng X. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. BMC Systems Biology. 2016; 10: 114.

[42] Zou Q, Wang Z, Guan X, Liu B, Wu Y, Lin Z. An approach for identifying cytokines based on a novel ensemble classifier. BioMed Research International. 2013; 2013: 686090.

[43] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28: 3150–3152.

[44] Liu B. BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. Briefings in Bioinformatics. 2019; 20: 1280–1294.

[45] Liu B, Li C, Yan K. DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. Briefings

in Bioinformatics. 2020; 21: 1733–1741.
[46] Qiao Y, Xiong Y, Gao H, Zhu X, Chen P. Protein-protein interface hot spots prediction based on a hybrid feature selection strategy. BMC Bioinformatics. 2018; 19: 14.
[47] Zhang X, Zou Q, Rodriguez-Paton A, Zeng X. Meta-Path Methods for Prioritizing Candidate Disease miRNAs. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019; 16: 283–291.
[48] Cabarle FGC, Adorna HN, Jiang M, Zeng X. Spiking Neural P Systems with Scheduled Synapses. IEEE Transactions on Nanobioscience. 2017; 16: 792–801.
[49] Shen Y, Tang J, Guo F. Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. Journal of Theoretical Biology. 2019; 462: 230–239.
[50] Ding Y, Tang J, Guo F. Identification of drug-target interactions via multiple information integration. Information Sciences. 2017; 418–419: 546–560.
[51] Zhang M, Li F, Marquez-Lago TT, Leier A, Fan C, Kwoh CK, *et al*. MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters. Bioinformatics. 2019; 35: 2957–2965.
[52] Xu L, Liang G, Shi S, Liao C. SeqSVM: a Sequence-Based Support Vector Machine Method for Identifying Antioxidant Proteins. International Journal of Molecular Sciences. 2018; 19: 1773.
[53] Qu K, Wei L, Yu J, Wang C. Identifying Plant Pentatricopeptide Repeat Coding Gene/Protein Using Mixed Feature Extraction Methods. Frontiers in Plant Science. 2019; 9: 1961.
[54] Zhang W, Liu J, Zhao M, Li Q. Predicting linear B-cell epitopes by using sequence-derived structural and physicochemical features. International Journal of Data Mining and Bioinformatics. 2012; 6: 557–569.
[55] Qu K, Han K, Wu S, Wang G, Wei L. Identification of DNA-Binding Proteins Using Mixed Feature Representation Methods. Molecules. 2017; 22: 1602.
[56] Yang H, Tang H, Chen X, Zhang C, Zhu P, Ding H, *et al*. Identification of Secretory Proteins in Mycobacterium tuberculosis Using Pseudo Amino Acid Composition. BioMed Research International. 2016; 2016: 5413903.
[57] Cheng J, Yang H, Liu M, Su W, Feng P, Ding H, et al. Prediction of bacteriophage proteins located in the host cell using hybrid features. Chemometrics and Intelligent Laboratory Systems. 2018; 180: 64–69.
[58] Tang H, Chen W, Lin H. Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. Molecular BioSystems. 2016; 12: 1269–1275.
[59] Chen X, Tang H, Li W, Wu H, Chen W, Ding H, *et al*. Identification of Bacterial Cell Wall Lyases via Pseudo Amino Acid Composition. BioMed Research International. 2016; 2016: 1654623.
[60] Liu B, Xu J, Zou Q, Xu R, Wang X, Chen Q. Using distances between top-n-gram and residue pairs for protein remote homology detection. BMC Bioinformatics. 2014; 15: S3.
[61] Wei L, Tang J, Zou Q. SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. BMC Genomics. 2017; 18: 742.
[62] Gould KA. The Elements of Statistical Learning (2nd edition): Data Mining, Inference, and Prediction. Dimensions of Critical Care Nursing. 2016; 35: 52.
[63] Rymarczyk T, Kozłowski E, Kłosowski G, Niderla K. Logistic Regression for Machine Learning in Process Tomography. Sensors. 2019; 19: 3400.
[64] Lei D, Tang J, Li Z, Wu Y. Using Low-Rank Approximations to Speed up Kernel Logistic Regression Algorithm. IEEE Access. 2019; 7: 84242–84252.
[65] Salzberg SL. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning. 1994; 16: 235–240.
[66] Feng PM, Lin H, Chen W. Identification of antioxidants from sequence information using naïve Bayes. Computational and Mathematical Methods in Medicine. 2013; 2013: 567529.
[67] Frank E, IH Witten. Generating accurate rule sets without global optimization. Proceeding of International Conference on Machine Learning (ICML). Morgan Kaufmann. 1998; 144–151.
[68] Johnson HR, Trinidad DD, Guzman S, Khan Z, Parziale JV, DeBruyn JM, *et al*. A Machine Learning Approach for Using the Postmortem Skin Microbiome to Estimate the Postmortem Interval. PLoS ONE. 2016; 11: e0167370.
[69] Borghesan F, Chioua M, Thornhill NF. Forecasting of process disturbances using k-nearest neighbours, with an application in process control. Computers & Chemical Engineering. 2019; 128: 188–200.
[70] Liu B, Zhu Y. ProtDec-LTR3.0: Protein Remote Homology Detection by Incorporating Profile-Based Features into Learning to Rank. IEEE Access. 2019; 7: 102499–102507.
[71] Liu B, Li K, Huang D, Chou K. IEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. Bioinformatics. 2019; 34: 3835–3842.
[72] Wang X, Wang Y, Xu Z, Xiong Y, Wei D. ATC-NLSP: Prediction of the Classes of Anatomical Therapeutic Chemicals Using a Network-Based Label Space Partition Method. Frontiers in Pharmacology. 2019; 10: 971.
[73] Xiong Y, Wang Q, Yang J, Zhu X, Wei D. PredT4SE-Stack: Prediction of Bacterial Type IV Secreted Effectors from Protein Sequences Using a Stacked Ensemble Method. Frontiers in Microbiology. 2018; 9: 2571.
[74] Zeng X, Wang W, Chen C, Yen GG. A Consensus Community-Based Particle Swarm Optimization for Dynamic Community Detection. IEEE Transactions on Cybernetics. 2020; 50: 2502–2513.
[75] Wang X, Zeng X, Ju Y, Jiang Y, Zhang Z, Chen W. A Classification Method for Microarrays Based on Diversity. Current Bioinformatics. 2016; 11: 590–597.
[76] Zhu H, Du X, Yao Y. ConvsPPIS: Identifying Protein-protein Interaction Sites by an Ensemble Convolutional Neural Network with Feature Graph. Current Bioinformatics. 2020; 15: 368–378.
[77] Sultana N, Sharma N, Sharma KP, Verma S. A Sequential Ensemble Model for Communicable Disease Forecasting. Current Bioinformatics. 2020; 15: 309–317.
[78] Breiman L. Random Forest. Machine Learning, 2001; 45: 5–32.
[79] Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002; 2: 18–22.
[80] Liu B, Yang F, Huang D, Chou K. IPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics. 2018; 34: 33–40.
[81] Ding Y, Tang J, Guo F. Identification of Protein-Ligand Binding Sites by Sequence Information and Ensemble Classifier. Journal of Chemical Information and Modeling. 2017; 57: 3149–3161.
[82] Ding Y, Tang J, Guo F. Predicting protein-protein interactions via multivariate mutual information of protein sequences. BMC Bioinformatics. 2016; 17: 398.
[83] Lv H, Zhang Z, Li S, Tan J, Chen W, Lin H. Evaluation of different computational methods on 5-methylcytosine sites identification. Briefings in Bioinformatics. 2020; 21: 982–995.
[84] Lai H, Zhang Z, Su Z, Su W, Ding H, Chen W, *et al*. IProEP: a Computational Predictor for Predicting Promoter. Molecular Therapy - Nucleic Acids. 2019; 17: 337–346.

Keywords: Diabetic protein marker; Machine learning; Feature extraction method; Ensemble classifiers; Dimensionality reduction

Send correspondence to: Hua Shi, School of Optoelectronic and Communication Engineering, Xiamen University of Technology, 361024 Xiamen, Fujian, China, E-mail: shihua@xmut.edu.cn

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.