IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 8, 2018, accepted October 17, 2018, date of publication November 19, 2018,
date of current version December 27, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2879635

Prediction for Membrane Protein Types Based on Effective Fusion Representation and MIC-GA Feature Selection

LEI GUO[1], SHUNFANG WANG[1], ZHENFENG LEI[2], AND XUEREN WANG[3]
[1]School of Information Science and Engineering, Yunnan University, Kunming 650504, China
[2]School of Information Science and Engineering, Xiamen University, Xiamen 361005, China
[3]School of Mathematics and Statistics, Yunnan University, Kunming 650504, China

Corresponding author: Shunfang Wang (sfwang_66@ynu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grants 11661081 and 11461079, in part by the Natural Science Foundation of Yunnan Province under Grant 2017FA032, and in part by the Training Plan for Young and Middle-Aged Academic Leaders of Yunnan Province under Grant 2018HB031.

ABSTRACT Membrane proteins occupy an important position in the life activities of humans and other species. The elucidation of membrane protein types provides clues for understanding the structure and function of proteins. With the fusion of various protein information including amino acid classification, physicochemical property, and evolutionary information, this paper proposes a system for predicting membrane protein types. In this system, a new feature selection method called MIC-GA is proposed to deal with the curse of high-dimensional features. The findings show that this approach is effective in reducing feature dimensions and improves prediction accuracy. Ensemble method based on stacked generalization is also used to solve the problem of feature heterogeneity. The performance of the present method is evaluated on two benchmark datasets. The overall prediction accuracies of eight types are 89.23% and 93.49% using jackknife test and independent test, respectively. The final experimental results show that our method is more effective than the existing methods for prediction of membrane protein types.

INDEX TERMS Prediction for membrane protein types, fusion representation, MIC-GA feature selection, ensemble method, stacked generalization.

I. INTRODUCTION

Membrane proteins play a vital role in the life activities of humans and other species. In the genome that has been sequenced, the membrane protein accounts for 30% [1]. Membrane proteins participate in important reactions of the cell, including transporting the substance into and out of the cell as a carrier, acting as a specific receptor for the hormone, carrying the recognition function of the cell and being responsible for signal transduction and cell-cell interactions [2]. In addition, membrane proteins are of particular importance in drug therapy as the targets for many drugs [3], [4]. Currently, more than 50% of drugs on the market are exerted by membrane proteins [5]. Because of the closely relation between the type and function of membrane proteins, knowing the type can provide clues for the structure and function of the protein [6]. Many reports proved that nuclear magnetic resonance (NMR) is an effective tool to determine the 3D structures of membrane proteins, however, it is time-consuming and expensive [7], [8]. With the incredibly growing number of protein sequences discovered in the postgenomic era, there is an urgent need for an effective method to predict membrane proteins and the introduction of machine learning methods greatly solve the problems.

According to their functions, membrane proteins can be classified into three classes: integral, peripheral and lipid-anchored. Based on the direct interaction relation between membrane proteins and lipid bilayers, the three classes can be further extended into eight basic types: (1) type I membrane proteins, (2) type II membrane proteins, (3) type III membrane proteins, (4) type IV membrane proteins, (5) multi-pass transmembrane proteins, (6) lipid chain-anchored membrane proteins, (7) GPI-anchored membrane proteins, (8) peripheral membrane proteins. Among them, Types I, II, III, and IV are of single-pass transmembrane proteins and detailed description of their differences are given in [9].

Feature representation is the basis of machine learning algorithms, which plays an important part for accurate prediction and classification tasks in biomedical scenarios [10]. In last few decades, many feature extraction methods had been applied to the prediction of membrane protein types. Amino acid composition (AAC) was firstly used to predict membrane protein types by Chou and Elrod [6]. However, the amino acid composition lost the order information in the sequence. To overcome this drawback, dipeptide composition (DipC) [11], [12], as a powerful feature extraction method, was proposed to improve the prediction accuracy of membrane protein types. Subsequently, considering both amino acid composition information and amphipathic sequence-order information, Chou [13] proposed pseudo-amino acid composition (PseAAC) to improve the performance of prediction. Then, other different forms of PseAAC [14]–[18] were proposed by many researchers to represent protein samples. Besides sequence information, feature extraction methods based on protein database information such as functional domain composition (FunD) [19] and gene ontology (GO) [20] were also applied into the prediction of membrane protein types. Afterwards, position-specific scoring matrix (PSSM) [21] based on evolutionary information and various forms of PSSM [22]–[26] were applied to many fields of bioinformatics. Many researchers had illustrated that the evolutionary information is more informative than the sequence itself [27]. However, the single feature extraction methods were still unable to meet the expectation of researchers so that the fusion representation by combining many features extracted from different methods were widely used [12], [28]. Since feature vectors come from different feature extraction methods, simple feature stitching may cause some problems, many fusion strategies [29], [30] were proposed to solve the problem of feature heterogeneity. Although fusion representation improves the accuracy of prediction, the high dimensionality of features in turn increases the computing time and complexity. To deal with the curse of high-dimensional features, many researchers proposed their own methods: a two-step optimal feature selection process based on minimum redundancy maximum relevance (mRMR) method was used to reduce the dimension of feature and obtained ideal result in prediction of membrane protein types [31]. Supervised dimensionality reduction methods like local linear discriminant analysis reduction (LLDA) [32] and kernel discriminant analysis (KDA) [33] were proposed to reduce the dimension and achieved the desired result. Zou *et al.* [34] proposed a max-relevance-max-distance (MRMD) feature ranking method, which balanced accuracy and stability of feature ranking and prediction task, and it was proved to be effective in image classification and protein-protein interaction prediction.

Furthermore, classification algorithm is also crucial in prediction of membrane protein types. General methods that were applied in predicting membrane protein types are listed here: k-nearest neighbor (KNN) classifier [11], naive nayes (NB) [15], support vector machine (SVM) [35], [36], random forest (RF) [37], neural network with back propagation training (NN) [36], probabilistic neural network (PNN) [38] and multi-label elastic net (EN) classifier [39]. Apart from single classifiers, various ensemble methods like stacking generalization [40], bagged decision tree [36], numerous support vector machines combined by vote rule [41] were applied into the prediction of membrane protein types and they achieved better results than single classifier did.

TABLE 1. The training dataset and independent dataset for eight types of membrane proteins.

Membrane protein types	Training dataset	Independent dataset
Type I	610	444
Type II	312	78
Type III	24	6
Type IV	44	12
Mutipass	1316	3265
Lipid-chain-anchored	151	38
GPI-anchored	182	46
Peripheral	610	444
overall	3249	4333

In this paper, a fusion representation method is used to extract the information from membrane protein samples. To reduce the dimensionality of features in the fusion representation and obtain higher prediction accuracy, we propose a novel feature selection called MIC-GA. This method incorporates maximum information coefficient (MIC) into the general form of genetic algorithm (GA). It can get the best feature subset and optimal classifier parameters for each feature representation simultaneously. After feature selection, to solve the problem of feature heterogeneity, being different from previous researchers who proposed the fusion strategies to combine features, we provide a strategy which combined the outputs of the classifiers trained by different feature extraction methods. By this way, the problem of feature heterogeneity would be transformed into classifier heterogeneity problem which solve the heterogenous feature problem from a new perspective. At last, two benchmark datasets including training dataset and independent dataset are used to evaluate the performance of our method. The overall accuracy of prediction for eight types are respectively 89.23% and 93.49% using the jackknife test and independent test.

II. DATASET AND METHODS

A. DATASET

The training dataset and the independent dataset we adopt are from [42] and they have been used in various papers [9], [12], [15], [32], [41], [43]. The datasets are screened from SWISS-PROT database through three-steps procedure which is presented in [9]. Then we get the training dataset consisting of 3249 samples and the independent dataset consisting of 4333 samples. Detailed distribution of samples is shown in Table 1.

B. FEATURE EXTRACTION

In order to establish an effective membrane protein prediction system, the key point is how to convert an original membrane protein sequence into a feature vector. To capture as much information of protein samples as possible, we apply such feature extraction method as amino acid classification-based methods, physicochemical property-based methods and evolutionary information-based methods in our experiment. Abundant features in this paper contain more information than most previous methods do. Among these features, local amino acid composition (LAAC), local dipeptide composition (LDC) [44] and tripeptide composition (TC) [45] are amino acid classification-based methods; sum of physicochemical property index (SPPI), auto correlation function (ACF) [46] are based on physicochemical property; reduced position-specific score matrix (RPSSM), evolutionary difference position-specific score matrix (EDP) and pseudo position-specific score matrix (PsePSSM) [9], [23], [24] are evolutionary information-based methods. SPPI is used in our study for the first time.

1) AMINO ACID CLASSIFICATION-BASED METHODS

Both the LAAC and LDC emphasize the amino acid classification information in the sequence composition. In the two methods, amino acids are classified into different amino acid groups according to certain classification methods which are listed in Table 2. The original 20 kinds of amino acids are divided into n groups. We use a symbol to represent all the amino acids in each group. According to the grouping of these amino acid classifications approaches, we can obtain 132 and 1302 dimensional feature vectors through LAAC and LDC respectively.

Since TC can reflect amino acid related information from spatial structure of the sequence, we also use tripeptide composition to extract the protein information and construct a prediction model. The tripeptide is a sequence consisting of three adjacent amino acids in the sequence. Then, we obtain 8000 feature vectors by using TC. Compared with other feature expression, less researchers used TC to study protein property due to its high dimensionality. In view of this, this paper tries to do some distinctive work with it. Detailed description will be presented in part II.

TABLE 2. Amino acid classifications approaches.

Method	Amino acid classification	LAAC	LDC
HP [47]	(ALIMFPWV)(DENCQGST YRHK)	2	4
DHP [48]	(ALVIFWMP)(STYCNGQ) (KRH)(DE)	4	16
7-Cat [49]	(AGV)(ILFP)(YMTS)(HNQW) (RK)(DE)C	7	49
20-Cat	AGVILFPYMTSHNQWRKDEC	20	400
ms [50]	(AVLIMC)(WYHF)(TQSN) (RK)(ED)(GP)	6	36
lesk [50]	(AST)(CVILWYMPF)(HQN) (RK)(ED)G	6	36
F-Ic4 [50]	(AWM)(GST)(HPY)(CVIFL) (DNQ)(ER)K	7	49
F-Ic2 [50]	(AWM)(GS)(HPY)(CVI) (FL)(DNQ)(ER)KT	9	81
F-IIIc4 [50]	(ACV)(HPL)(DQ)S(ERGN) F(IMT)(KW)Y	9	81
Murphy8 [51]	(LVMIC)(AG)(ST)P(FYW) (DENQ)(KR) H	8	64
Murphy15 [51]	(LVIM)CAGSTP(FY) WEDNQ(KR)H	15	225
Letter6 [52]	(VIM)(CYFQLTW)(RGD) (HKSNP)AE	6	36
Letter12 [53]	(LVIM)C(AG)(ST)P(FY)W (ED)NQ(KR)H	12	144
Hydrophobicity [54]	(RKEDQN)(GASTPHY) (CLVIMFW)	3	9
Polarity [54]	(LIFWCMVY)(PATGS) (HQRKNED)	3	9
Polarizability [54]	(GASDT)(CPNVEQIL) (KMHFRYW)	3	9
Charge [54]	(KR)(ANCQGHILMFPS TWYV)(DE)	3	9
Secondary structure [54]	(EALMQKRH)(VIYCW FT)(GNPSD)	3	9
Electronic group [38]	(DEPA)(VLI)(KNR)(FM YTQ)(GHWS)C	6	36
Total	—	132	1302

2) PHYSICOCHEMICAL PROPERTY-BASED METHODS

The physicochemical properties index of 20 kinds of amino acids are very different, such as chargeability, hydrophilicity, electron transferability and so on. The physicochemical and biochemical properties of amino acids are also important factors affecting the type of membrane protein. The AAindex database [55] contains the index of physicochemical properties for each amino acid derive from previous published papers. There is a total of 566 physicochemical properties at present. In this study, the physicochemical property index including NA missing value is screened out, and the remaining 537 kinds of amino acid physicochemical properties index are extracted for feature representation. Based on the physicochemical properties of each amino acid, we can get the sum of physicochemical property index for each sequence. For example, the index of hydrophobicity is calculated as the formula (1).

$$SH = \sum_{i=1}^{L} HI^i. \tag{1}$$

where HI^i indicates the hydrophobicity index of i-th amino acid and 537 features can be obtained based on SPPI.

The autocorrelation function based on the physicochemical properties is also an important factor affecting the type of membrane protein. We select five properties of amino acids — (I) Codon diversity; (II) Electrostatic charge; (III) Molecularvolume; (IV) Polarity; and (V) Secondary structure. The autocorrelation function of a protein is defined as:

$$r_\lambda = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} p_i p_{i+\lambda}, \quad \lambda = 1, 2, \cdots, m. \tag{2}$$

where L is the length of the sequence of the membrane protein, λ represents the sequence of molecules, p_i represents the i-th amino acid physicochemical property index. We set λ to 30, then the number of features using ACF is 150.

3) EVOLUTIONARY INFORMATION-BASED METHODS

In this paper, we mainly use position-specific scoring matrix (PSSM) to extract evolutionary information. PSSM is a general feature extraction obtained by searching protein sequences that have the evolutionary relationship with searched sequence in the protein database. In the resulting score matrix, each amino acid in the sequence is given a specific score. It is expressed as follows:

$$PSSM = \begin{bmatrix} M_{1\rightarrow 1} & M_{1\rightarrow 2} & \cdots & M_{1\rightarrow 20} \\ M_{2\rightarrow 1} & M_{2\rightarrow 2} & \cdots & M_{2\rightarrow 20} \\ \vdots & \vdots & \ddots & \vdots \\ M_{L\rightarrow 1} & M_{L\rightarrow 2} & \cdots & M_{L\rightarrow 20} \end{bmatrix} \tag{3}$$

where L is the length of the sequence of the membrane protein and $M_{i\rightarrow j}$ denotes the score of the amino acid which mutates from i-th to j-th position during evolution process. The PSSM in this paper is obtained through PSI-Blast [56] software. The number of iterations used for Blast is 3, and E-value threshold used for Blast is 0.001.

To extract more information from position-specific scoring matrix, we summarize the research of the predecessors, considering RPSSM, EDP and PsePSSM as feature extraction methods to extract evolutionary information. Detail of description of methods above are presented in [9], [23], and [24].

C. MIC-GA FEATURE SELECTION

After deploying feature extraction, all original membrane protein sequences are converted into high dimensional vectors. To reduce the feature dimensionality and obtain better results, we propose an effective feature selection method called MIC-GA. This method incorporates maximum information coefficient into the general form of genetic algorithm which can get the best feature subset and optimal classifier parameters for each feature representation simultaneously.

1) MAXIMUM INFORMATION COEFFICIENT

Maximum information coefficient was proposed in [57] which can measure the linear and nonlinear relationships inside the data. Because the MIC is mainly calculated by mutual information and meshing method, we define the mutual information as follows:

To give a random variable $X = \{x_1, x_2, \cdots, x_n\}$ and a random variable $Y = \{y_1, y_2, \cdots, y_n\}$, where n is the sample size, the mutual information is defined as:

$$MI(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{4}$$

where $p(x,y)$ represents the joint probability density function, $p(x)$ and $p(y)$ are the marginal probability density functions of X and Y, respectively. Then, the calculation process of the MIC as follows:

Step. 1 To Give a finite ordered set $D = \{(x_i, y_i), i = 1, 2, \cdots, n\}$, we divide the range of X into a segments, and divide the range of Y into b segments to form a grid which is defined as G. There are many ways to segment different grids for the same i, j. We define the max value of $MI(X, Y)$ as the mutual information for the set D under the grid G.

$$\hat{MI}(D, i, j) = \max MI(D|G) \tag{5}$$

where $D|G$ indicates that set D is divided by grid G.

Step. 2 In order to fairly compare the mutual information values of the grid G under different division methods, the mutual information should be normalized. After normalization, we combine the $\hat{MI}(D, i, j)$ obtained by different division methods into a feature matrix which is defined as $M(D)_{i,j}$. The calculation formula is as follows:

$$M(D)_{i,j} = \frac{\hat{MI}(D, i, j)}{\log_2 \min(i, j)} \tag{6}$$

Step. 3 Get the MIC value for set D.

$$MIC(D) = \max_{i\times j<B(n)} \left\{M(D)_{x,y}\right\} \tag{7}$$

where $B(n)$ is the upper limit of the number of grids after meshing. The study [57] points out that it is best when $B(n) = n^{0.6}$ in general. Then we can evaluate the quality of features by calculating the maximum information coefficient between features and categories.

2) GENETIC ALGORITHM

Genetic algorithm is built based on natural selection and population genetics. It starts from the original population and experiences the selection, crossover and mutation to form a better population. We apply GA to feature selection in detail as follow:

a: Code of Samples

To obtain the best feature subset and optimal classifier parameters, in our algorithm, individuals in the population are mainly composed of feature code and classifier parameter code. Wherein, the feature code indicates that whether the feature is selected; and the parameter code indicates a specific value of the parameter in the classifier.

b: Selection Strategy

In the genetic algorithm, the quality of an individual is evaluated by the fitness function value. The higher the fitness function value is, the higher the individual quality will be. The feature subset with higher fitness is selected as potential results. The fitness function in our algorithm is 5-fold cross validation on the training dataset, where the classifier in our algorithm is random forests which is widely used in many fields of bioinformatics [58], [59].

c: Crossover Operation

Randomly selecting two individuals from the population as parents during crossover operation. For the feature code of individuals, the common feature shared by parents are selected as advantageous gene and preserved in next generation. The feature selected by only one of the parents is defined as a non-advantageous gene and it is inherited to next generation through stochastic selection. For the

TABLE 3. The specific process of MIC-GA feature selection.

MIC-GA feature selection
Step 1. Calculating the maximum information coefficient between features and categories to obtain effective feature scores.
Step 2. Taking the maximum information coefficient of each feature as the probability that the feature is selected in the process of population initialization of genetic algorithm.
Step 3. Calculating the fitness of each individual using the fitness function.
Step 4. Each individual in the population is selected, crossed and mutated to form a new generation of populations.
Step 5. The algorithm terminates when repeating Step3 and Step4 m times or the optimal value for five consecutive generations no longer changes.

parameter code of individuals, it respectively inherits the first half of paternal gene and the last half of maternal gene.

d: Mutation Operation

Mutation operation unables the genetic algorithm to make local random search ability and maintain population diversity. In this paper, a position in feature code and parameter code is randomly selected and its value is going to change. The mutated individual obtained from this step is added into population.

e: Condition of Termination

The algorithm terminates when genetic manipulation reaches the maximum number of iterations. If continuous generations of optimal individual do not change any more in iteration, the algorithm may terminate in advance.

3) MIC-GA FEATURE SELECTION

Theoretically, genetic algorithm is a stochastic search metaheuristic and is considered to be a unbiased global optimization method [60]. Nevertheless, problems emerge in practical application. For instance, the convergence is too slow to result in high time complexity.

A good initial population is an effective solution to speed up the convergence of the genetic algorithm. MIC-GA incorporates maximum information coefficient into the general form of genetic algorithm. The MIC value of feature is used as the probability that the feature is selected. Then, by this way, we can get a better initial population at the beginning of the algorithm. Two important reasons why MIC can be interpreted as probability as follows: Firstly, the value range of the MIC is between [0,1] for each feature after normalization. Coinciding with the probability in reality. Secondly, as the correlation between features and categories increases, the MIC calculation results also increase. The larger the MIC value is, the higher possibility that the feature is selected. Even the feature with the smallest value still has the probability to be selected. This is consistent with our intention at the beginning.

The specific processes of MIC-GA feature selection are listed in Table 3.

D. ENSEMBLE CLASSIFIER BASED ON STACKED GENERALIZATION

After feature selection, we obtain the best feature subset and optimal classifier parameters for each feature extraction. Since the features come from different feature extraction methods, the problem of feature heterogeneity becomes a big challenge in prediction of membrane protein types. Ensemble method based on stacked generalization can combine the results of the base classifiers and learns two or more times to obtain the final result. If we use one single feature representation as the input to train base classifier, the problem of feature heterogeneity would be transformed into the classifier heterogeneity problem. The ensemble method based on stacked generalization [61] is an effective method to solve the problem of classifier heterogeneity, so we can solve the heterogeneous feature problem from a new perspective. In our study, random forests (RF) and neural networks (NN) are respectively used as base classifier and meta classifier.

1) BASE CLASSIFIER–RANDOM FORESTS

Random forest is the classifier which consists of a series of decision tree. They can be expressed as $\{h(\theta_n, x); n = 1, \cdots\}$, where θ_n is independent identically distributed random vectors. Each decision tree can give its final result to vote. The class probability of random forest is as follows:

$$P_j(x) = \frac{1}{N}\sum_{n=1}^{N} I(h(\theta_n, x) = j) \tag{8}$$

where N is the total number of decision trees. When the output of decision tree $h(\theta_n, x)$ is j, $h(\theta_n, x) = 1$, otherwise, $h(\theta_n, x) = 0$. The research shows that the class probability as the output of the base classifier performer better [62]. Then we use $p(x)$ as the expression of output for the base classifiers, it can be represented as follows:

$$\begin{aligned} P(x) &= (P_1^{\mathrm{T}}, P_2^{\mathrm{T}}, \cdots, P_L^{\mathrm{T}})^{\mathrm{T}} \\ &= (\underbrace{P_1^{1}, \cdots, P_c^{1}}_{Classifier\_C_1}, \underbrace{P_1^{2}, \cdots, P_c^{2}}_{Classifier\_C_2}, \cdots, \underbrace{P_1^{\mathrm{L}}, \cdots, P_c^{\mathrm{L}}}_{Classifier\_C_L}) \end{aligned}$$

where L is number of base classifiers, there is eight kind of feature extraction methods so that we set L to 8 in our works. C is the class number of samples.

Furthermore, due to the unbalanced training dataset, we set the sample weight by giving different misclassification costs to different types of membrane protein samples when training in the random forest. The weights of each class samples W_L are calculated by the formula (9):

$$W_L = M/N_L \tag{9}$$

where M is the number of largest class samples and N_L is the number samples of class L.

2) META CLASSIFIER-NEURAL NETWORKS

Neural network is a computing model which consists of many neurons and the connection of the neurons. It can perform complex nonlinear transformation for the input features

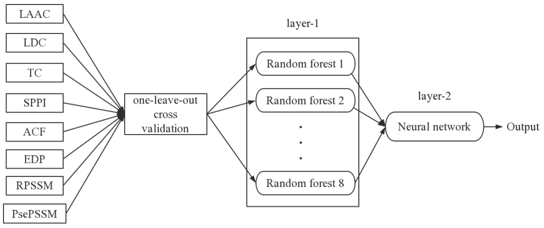


FIGURE 1. A two-layer stacking architecture of ensembles of classifiers.

through training so that its output is infinitely close to the target value. The neural network we use including three basic layers: (1) The input layer, which contains some perception units which would connect network with the external environment. (2) The hidden layer, which transfers the input space into the hidden space with nonlinear. The dimension of hidden layer is often high in most cases. (3) The output layer, which provides final results of neural network.

Then, we can combine the results of the base classifiers together and balance the advantages and disadvantages of each heterogeneous classifier by using neural networks.

3) STACKED GENERALIZATION

At last, we construct an ensemble method based on stacked generalization to predict membrane protein types. The flow chart of ensemble method in this paper is shown in Figure 1. The specific process is presented as follows:

Step 1: Adopting the leave-one-out cross validation method to train the base classifiers in layer-1. For example, a training dataset $\{(x_1, y_1), \cdots, (x_m, y_m)\}$ and base classifier are given. In each iteration, choosing one sample from training dataset and using it as test sample in order, and the remaining $m - 1$ samples are used as training samples by base classifier. We can get eight classification results by using eight feature extraction methods to classify the test samples. The results of repetition of the process m times are regarded as the outputs of the layer-1.

Step 2: The outputs of the layer-1 are combined together as new training dataset to train the meta-classifier. When testing unknown sample, we take the result of the meta-classifier as final output.

The system flow chart we proposed is shown in Figure 2.

E. PERFORMANCE EVALUATION

To evaluate the performance of the model, we adopt sensitivity (Sn), specificity (Sp), overall prediction accuracy (ACC) and Matthew's Correlation Coefficient (MCC) [63] in our work. Their definetions are presented as follows:

$$Sn_i = TP_i/(TP_i + FN_i) \tag{10}$$

$$Sp_i = TN_i/(TN_i + FP_i) \tag{11}$$

$$MCC_i = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i+FN_i)(TP_i + FP_i)(TN_i + FP_i)(TN_i + FN_i)}} \tag{12}$$

TABLE 4. Comparison of the overall prediction accuracy between different feature extraction methods using the jackknife test and independent test.

Method	Jacknife test (%)	Independent test(%)
Method I	79.41	86.45
Method II	79.29	87.47
Method III	85.60	91.71
Combination I	87.57	92.57
Combination II	89.23	93.49

$$ACC = \frac{\sum_i TP_i}{N} \tag{13}$$

where true positive (TP) is the number of positive samples predict correctly; false positive (FP) is the number of negative events that are incorrectly predicted to be positive; true negative (TN) is the number of negative samples predict correctly; false negative (FN) is the number of subjects that are predicted to be negative despite they are positive.

In addition, three validation methods are also used to examine our model for its effectiveness: independent test, sub-sampling test and jackknife test.

III. RESULTS AND DISCUSSION

A. COMPARISON OF FEATURE EXTRACTION METHODS

To assess which information-based method is more effective, we compared the overall prediction accuracy for each method on the training dataset and independent dataset. The best feature subset and the optimal parameters for each feature extraction method are obtained by MIC-GA feature selection on training dataset. Note that all parameters of the system including best feature subsets and the optimal parameters are not re-parameterized to apply on the independent dataset.

Considering single information-based methods, findings show that the most effective method is based on evolutionary information, whose overall prediction accuracy is higher than those of other information-based methods. Amino acid classification-based methods LAAC, LDC and TC are combined together as one method, named method I; physicochemical property-based methods SPPI and AFS are combined as method II, evolutionary information-based methods EDP, RPSSM and PsePSSM are combined as method III. As shown in Table 4, the overall prediction accuracy of method III achieves better results than those of other methods both on the training dataset and independent dataset, which indicates that evolutionary information has higher efficiency.

Furthermore, to show the effectiveness of the proposed SPPI extraction method, we integrated the feature extraction methods LAAC, LDC, TC, ACF, EDP, RPSSM, PsePSSM as combination 1, all of feature extraction methods are integrated as combination 2. Comparison of their prediction performance are made. In Table 4, with SPPI feature extraction method, the overall prediction accuracy increases by 1.66% and 0.92% using the jackknife test and independent test respectively. The experimental results illustrate the assumption that the performance of membrane protein types prediction could be improved by adding SPPI feature extraction.

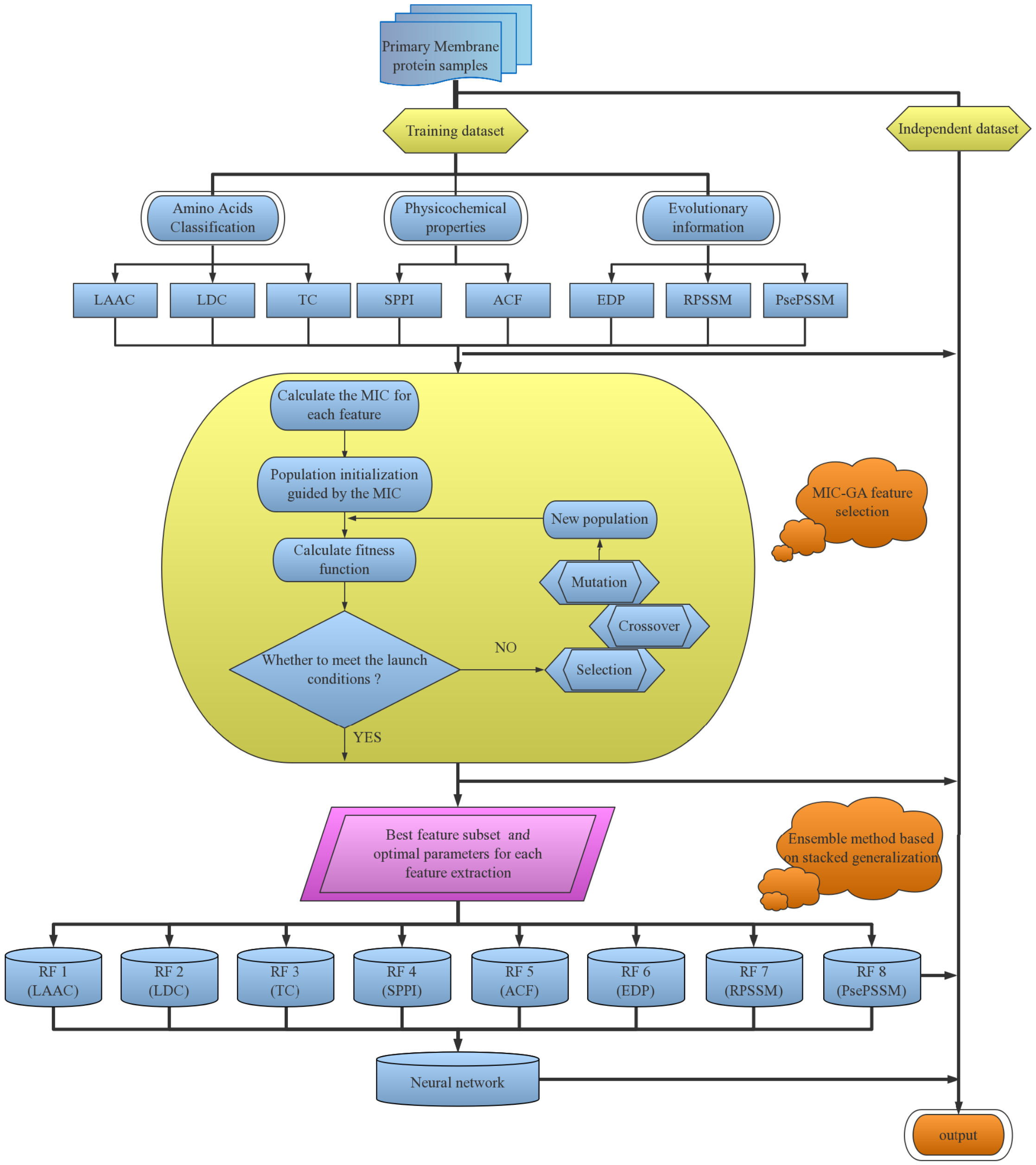

Primary Membrane protein samples
Training dataset
Independent dataset
Amino Acids Classification
Physicochemical properties
Evolutionary information
LAAC
LDC
TC
SPPI
ACF
EDP
RPSSM
PsePSSM
Calculate the MIC for each feature
Population initialization guided by the MIC
New population
Calculate fitness function
Mutation
Crossover
NO
Whether to meet the launch conditions ?
Selection
YES
MIC-GA feature selection
Best feature subset and optimal parameters for each feature extraction
Ensemble method based on stacked generalization
RF 1 (LAAC)
RF 2 (LDC)
RF 3 (TC)
RF 4 (SPPI)
RF 5 (ACF)
RF 6 (EDP)
RF 7 (RPSSM)
RF 8 (PsePSSM)
Neural network
output

FIGURE 2. **The complete flow chart of the proposed method.**

B. EFFICIENCY ANALYSIS OF MIC-GA FEATURE SELECTION

Since the performances of GA strongly rely on their settings, such as population size, crossover rate and mutation rate. Taking LAAC extraction method as an example, the influence of different parameters on MIC-GA convergence is analyzed. The effects of different parameters settings on the optimal results of each generation are shown in Figure 3. Among them, the population size occupies an honorable position in the whole process of GA optimization for it limits the number of individual when searching samples. Smaller population can accelerate the operation of GA but could reduce the diversity of population. In this paper, we adaptively set population size as the individualąŕs gene number for maintaining diversity of population and effectiveness of operation. The first figure in Figure 3 presents the respective accuracy when mutation rate is 0.1 and crossover rates are respectively 0.3, 0.5 and 0.7. Findings show that the convergence of population performs better when crossover rate is 0.5. Too high crossover rate may destroy excellent individuals in population and make negatively affect on evolutionary computation. Too low crossover rate, however, may cause slow production of new individuals, decelerate optimization and

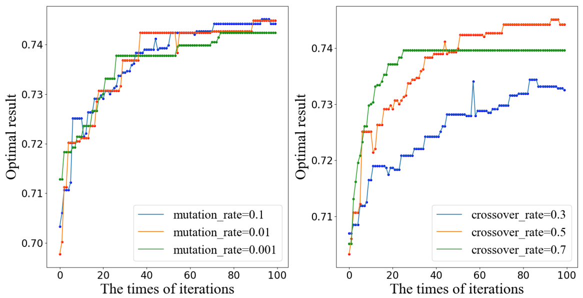


FIGURE 3. Optimal results of each generation with different parameter settings.

reduce the convergence performance. The second figure in Figure 3 demonstrates the accuracy of different mutation rates when crossover rate is 0.5. From the figure, it is clear that GA performs the best when mutation rate is 0.1. Setting a larger mutation probability increases the chance of destroying the better individuals and makes the optimal value of each generation fluctuate greatly in the process of optimization, but it also produces numerous new individuals and increases diversity of population. Taking those into account, we set mutation rate as 0.1 and crossover rate as 0.5.

Now, we intend to perform the MIC-GA feature selection to deal with the curse of high-dimensional features. The effects of the iteration times on the optimal results of each generation for all feature extraction methods are shown in Figure 4, where the ordinate is the optimal result for each generation and the abscissa means the number of iterations. From the figure, we find that there is a certain fluctuation in the optimal result for each generation with the number of iterations increases. The reason is that the MIC-GA is a kind of probability search feature selection method which has certain randomness. With the incremental number of iterations, its prediction accuracy gradually increases and the results are increasingly stable.

To assure whether the MIC-GA feature selection method is effective, we take the feature dimension and the prediction accuracy into account to compare the performance between original features and MIC-GA selected features. Figure 5 is the distribution of the MIC-GA selected features and non-selected features. From the figure, we find that although the features from TC methods have the highest dimension, nearly 94.13% are redundant features, while features based on EDP method retain the largest proportion of original features after MIC-GA feature selection. The experimental results show that MIC-GA can remove the redundant feature, thus greatly reduce the feature dimension and further decrease the computing complexity.

Then, we intend to validate if MIC-GA would improve the prediction accuracy of membrane protein types. The detailed results are shown in Table 5 and Table 6. From the tables, we can find that after MIC-GA feature selection, the overall prediction accuracy of each feature extraction method has been well improved in jackknife test. However, for the independent test, the accuracy of the TC and RPSSM becomes slightly worse after MIC-GA feature selection, but it only decreases by 1% at the most. MIC-GA is still effective since it removes numerous redundant features for these methods.

C. ANALYSIS OF OPTIMAL FEATURE

After MIC-GA feature selection, we obtain the optimal feature subset for each feature representation. To intuitively observe the distribution of optimal features MIC-GA selected, we plot the MIC values for each feature representation which are shown in Figure 6, where the selected optimal feature subset has a key mark. From the figure, we find that

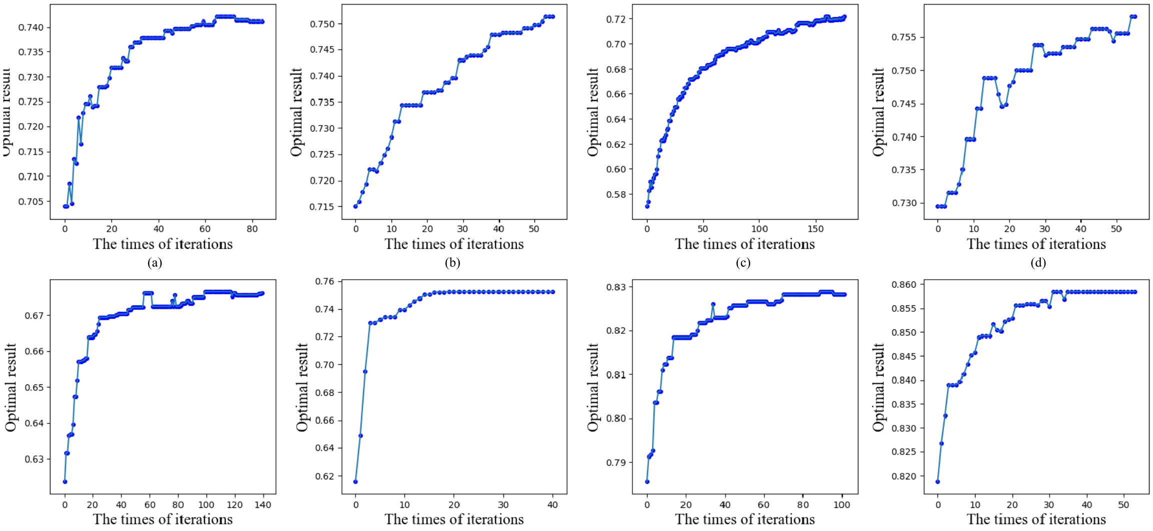


FIGURE 4. Optimal results of each generation with different iteration times for LAAC, LDC, TC, SPPI, ACF, EDP, RPSSM, PsePSSM respectively.

TABLE 5. Comparison of the overall prediction accuracy between original features and MIC-GA selected features using the jackknife test.

Feature extraction method	Original features (%)	MIC-GA selected features(%)
LAAC	71.56	73.90
LDC	72.73	75.36
TC	65.65	71.56
SPPI	73.53	75.13
ACF	65.28	66.61
EDP	74.52	75.50
RPSSM	81.04	82.58
PsePSSM	84.30	84.98

TABLE 6. Comparison of the overall prediction accuracy between original features and MIC-GA selected features using the independent test.

Feature extraction method	Original features (%)	MIC-GA selected features(%)
LAAC	82.69	83.45
LDC	82.89	83.20
TC	82.13	81.51
SPPI	81.86	82.71
ACF	80.15	79.90
EDP	85.10	85.23
RPSSM	89.31	89.38
PsePSSM	90.53	91.16

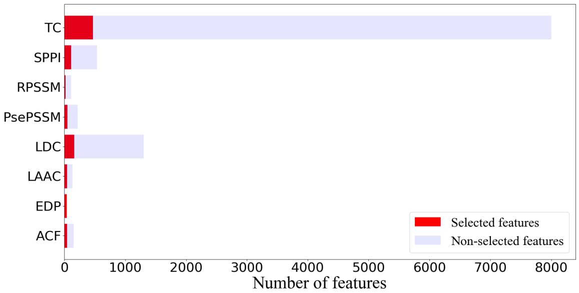


FIGURE 5. Distribution of the selected features and Non-selected features for each extraction method.

the best feature subset MIC-GA selected is not the features which has strong correlation with categories. Even the worst feature can be the part of the subset.

To verify whether the feature subset MIC-GA selected is reliable for classification, we select the top n features in MIC for comparative experiment, where n is set to the same values as the number of features MIC-GA selected. Figure 7 is the results of comparison of the overall prediction accuracy between top n features and MIC-GA selected features using the jackknife test and independent test. The results suggest that the top n features combined would not be the best subset for classification. The reason is that although the top n features have strong correlations with categories, there may be a lot of redundant features. The correlation between features and categories should be considered in the construction, as well as noticing the redundant features. MIC-GA feature selection we proposed works well in this respect.

Furthermore, we also find that the features extracted by SPPI method have a higher correlation with categories. The MIC of features extracted by other methods are generally lower than 0.35, especially the TC method whose MIC is no more than 0.1; while a large number of features whose MIC values exceed 0.35 when the features are extracted by SPPI method. We listed the top ten features are extracted by SPPI: PUNT030101, CIDH920104, PUNT030102, CORJ870107, CORJ870108, QIAN880119, KYTJ820101, CORJ870103, BIOV880102 and WOLS870101. The overall prediction accuracy of the individual feature that extracted by each physicochemical property is listed in Table 7. From the table we find that even if there is only one feature, the overall

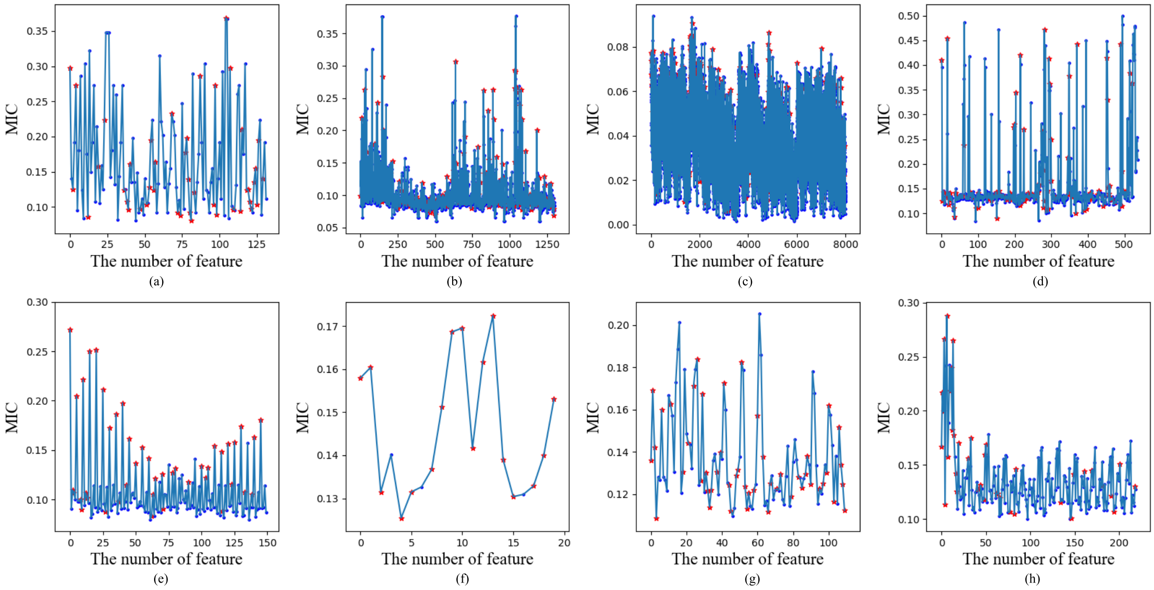


FIGURE 6. Distribution of the optimal features MIC-GA selected for each feature extraction method.

TABLE 7. The overall accuracy of top 10 features extracted by SPPI using the jackknife test and independent test.

AA index	Jacknife test(%)	Independent test(%)
PUNT030101	43.43	63.49
CIDH920104	42.84	62.45
PUNT030102	42.10	62.50
CIDH920104	43.40	60.56
CORJ870107	42.32	62.06
QIAN880119	43.37	63.12
KYTJ820101	42.29	59.59
CORJ870103	43.52	62.77
BIOV880102	43.06	60.79
WOLS870101	42.41	61.80

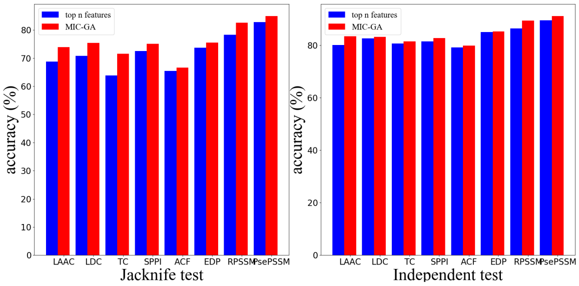


FIGURE 7. Comparison of the overall prediction between top *n* features and MIC-GA selected features using the jackknife test and independent test respectively.

accuracy is still more than 42% and 59% using the jackknife test and independent test. Therefore, we can make a conclusion that the types of membrane protein have a strong correlation with these physicochemical properties. The total content of these amino acid physicochemical property index varies from different membrane protein types.

D. THE EFFICIENCY OF ENSEMBLE METHOD

To illustrate the effectiveness of ensemble method based on stacked generalization, we compare it with base classifiers trained by each feature extraction method using the jackknife test and independent test. The detailed performance of each classifier on training dataset and independent dataset are listed in Table 8 and Table 9. As presented in the two tables, the ensemble classifier in most cases is better than the base classifiers in Sn, Sp and MCC. Although the ensemble method is not as good as base classifiers in a few types, overall, better results would be obtained by using ensemble method. The reason is that meta classifier can integrate the outputs of the base classifiers and balance the advantages and disadvantages of each heterogeneous classifier.

Furthermore, as shown in two tables, we find that our method achieves the highest results on type I, mutipass and peripheral membrane proteins, which are respectively 93.61% (90.32%), 96.50% (95.65%) and 88.20% (88.06%); while those on type III only achieves 37.50% (33.33%). The reason may be that the used training dataset is highly unbalance which has too much number of type I, mutipass and peripheral membrane protein samples. In order to improve the overall accuracy during training, the random forest is biased towards large size types. Although we tried to set the sample weight to solve the problem of unbalance datasets, the accuracies of small-size types are still not as good as that of large-size types. Other effective strategies for solving imbalance problem may contribute to improving our method.

E. COMPARISON WITH THE EXISTING METHODS

Many researches have discussed on membrane protein types prediction, which are listed in Table 10. As seen from the table, we achieve slightly worse result compare with Chen's research [15] using jackknife test, but for independent test, however, we achieve better results. The reason may be the intrinsic property inside the dataset. Moreover, we find that

TABLE 8. Comparison of the Sn, Sp and MCC between the base classifiers and ensemble method using jackknife test.

Membrane protein types	Index	LAAC	LDC	TC	SPPI	ACF	EDP	RPSSM	PsePSSM	Stacking
Type I	Sn	0.8082	0.8443	0.7492	0.8082	0.7934	0.859	0.8918	0.8885	0.9361
	Sp	0.9098	0.9049	0.9204	0.9132	0.8485	0.9227	0.9409	0.9674	0.9746
	MCC	0.6715	0.6892	0.6474	0.6776	0.5595	0.7321	0.7909	0.8465	0.8952
Type 2	Sn	0.3878	0.3429	0.2628	0.4944	0.0125	0.4936	0.625	0.6891	0.7628
	Sp	0.9813	0.9833	0.9939	0.9724	0.9942	0.984	0.9854	0.9816	0.9864
	MCC	0.4804	0.4496	0.4379	0.5143	0.2669	0.5841	0.6902	0.7171	0.7891
Type 3	Sn	0	0.0417	0.2917	0.0833	0	0.1667	0.25	0.25	0.375
	Sp	1	1	1	0.9997	1	0.9972	0.9988	0.9997	0.9997
	MCC	0	0.2034	0.5386	0.2341	0	0.2223	0.3845	0.4611	0.5792
Type 4	Sn	0.2955	0.3636	0.1591	0.4091	0.3409	0.5909	0.75	0..6818	0.7227
	Sp	1	0.9994	0.9984	0.9972	0.9981	0.9978	0.9984	0.9991	0.9984
	MCC	0.5409	0.5653	0.3002	0.5173	0.498	0.6786	0.8046	0.7849	0.8185
Mutipass	Sn	0.8845	0.8769	0.9027	0.8716	0.8587	0.9012	0.9339	0.9354	0.965
	Sp	0.9322	0.9379	0.8184	0.9571	0.8712	0.8624	0.9519	0.9731	0.9752
	MCC	0.8189	0.8193	0.709	0.8387	0.7249	0.754	0.8851	0.9124	0.94
Lipid-chain-anchored	Sn	0.298	0.3709	0.5232	0.4503	0.1457	0.4371	0.4305	0.4636	0.6159
	Sp	0.9926	0.9916	0.9868	0.9832	0.9981	0.9929	0.9955	0.9939	0.9923
	MCC	0.4273	0.4865	0.5692	0.4839	0.3274	0.5576	0.5822	0.59	0.6871
GPI-anchored	Sn	0.4396	0.4066	0.3516	0.522	0.1154	0.489	0.5275	0.6978	0.8022
	Sp	0.9899	0.9905	0.9938	0.9902	0.9945	0.9938	0.9964	0.9932	0.9935
	MCC	0.5437	0.5212	0.5035	0.6124	0.2349	0.6193	0.675	0.762	0.831
Peripheral	Sn	0.7951	0.8082	0.7311	0.7705	0.7426	0.6623	0.8443	0.8836	0.882
	Sp	0.8598	0.8651	0.8958	0.8776	0.8549	0.9159	0.9091	0.9041	0.9451
	MCC	0.5808	0.5985	0.5892	0.5892	0.5337	0.5725	0.6965	0.7165	0.7926

TABLE 9. Comparison of the Sn, Sp and MCC between the base classifiers and ensemble method using independent test.

Membrane protein types	Index	LAAC	LDC	TC	SPPI	ACF	EDP	RPSSM	PsePSSM	Stacking
Type I	Sn	0.8176	0.8446	0.7275	0.8243	0.8153	0.8491	0.8761	0.8896	0.9032
	Sp	0.9499	0.938	0.9224	0.9347	0.8946	0.9578	0.9727	0.9838	0.9892
	MCC	0.6948	0.6796	0.6117	0.6574	0.5626	0.7403	0.8093	0.8615	0.8932
Type 2	Sn	0.3974	0.3718	0.2564	0.5256	0.2051	0.6026	0.7051	0.7692	0.8718
	Sp	0.9845	0.9894	0.9922	0.9819	0.9948	0.9915	0.9838	0.9911	0.9873
	MCC	0.3433	0.3707	0.3008	0.4146	0.2851	0.5763	0.5494	0.6799	0.6905
Type 3	Sn	0	0	0	0.3333	0	0	0	0.3333	0.3333
	Sp	1	1	0.9998	1	1	0.9991	0.9991	0.9991	0.9988
	MCC	0	0	-0.0006	0.5771	0	-0.0011	-0.0011	0.3324	0.3076
Type 4	Sn	0.3333	0.25	0.0833	0.4167	0.5	0.4167	0.5	0.6667	0.75
	Sp	0.9993	0.9998	0.9988	0.9984	0.9975	0.9993	0.9986	0.9993	0.9995
	MCC	0.4353	0.4322	0.1161	0.415	0.4182	0.5092	0.4986	0.6955	0.7828
Mutipass	Sn	0.8711	0.864	0.8689	0.86	0.8374	0.8949	0.9225	0.9305	0.9565
	Sp	0.9241	0.941	0.8277	0.9654	0.8998	0.867	0.9607	0.9757	0.9616
	MCC	0.7317	0.7354	0.6534	0.7491	0.6686	0.7216	0.8345	0.8581	0.8915
Lipid-chain-anchored	Sn	0.2105	0.1842	0.3158	0.2985	0.2105	0.3421	0.3684	0.4474	0.4737
	Sp	0.9953	0.9923	0.9893	0.9928	0.9974	0.9932	0.9965	0.9956	0.9963
	MCC	0.2396	0.1721	0.2475	0.2686	0.2935	0.3191	0.4173	0.455	0.4966
GPI-anchored	Sn	0.5	0.413	0.3478	0.4783	0.87	0.5435	0.587	0.7174	0.8478
	Sp	0.9951	0.9953	0.9932	0.9916	0.9984	0.9939	0.9972	0.9949	0.996
	MCC	0.5061	0.4432	0.3448	0.419	0.1738	0.5107	0.6339	0.6521	0.7657
Peripheral	Sn	0.7725	0.7905	0.7275	0.741	0.7477	0.6847	0.8333	0.8941	0.8806
	Sp	0.9149	0.9164	0.9283	0.9211	0.9221	0.9393	0.947	0.9465	0.973
	MCC	0.5752	0.5908	0.5749	0.5673	0.5744	0.5729	0.6971	0.7356	0.8131

TABLE 10. Comparison of the overall prediction accuracy with existing approaches.

Method	Test method	
	Jacknife test(%)	Independent test(%)
AAC based on Least Euclidean distance [64]	51.7	64.4
AAC based on Covariance [9]	52.0	37.2
PsePSSM based on ensemble method combining KNN [9]	85.0	91.6
Physicochemical properties based on ensemble method combining SVM [41]	—	91.0
Fusion representation based on SVM[15]	89.4	92.6
PsePSSM based on LLDA[32]	87.2	88.7
PsePSSM and DC based on GPP and KNN[12]	84.0	90.2
PsePSSM based on PCA and KNN[32]	77.65	80.66
Our method	89.23	93.49

MIC-GA feature selection performs better than PCA does when both adopt PsePSSM extraction method. However, comparing with the method which adopts LLDA to reduce the dimensions of feature, our results seem to be bad using jacknife test [32]. Exploring the reason, we find that LLDA is supervised dimensionality reduction method which might have learned the test sample during the process of dimensionality reduction so that it works well using jackknife test and performs poorly using independent test. Overall, the prediction results show that our method has better performance in predicting membrane protein types.

Additionally, our method can be incremental learning, thus it has better extensibility than existing method. It is easy to quickly embed a new feature extraction method into our system once new feature extraction method makes improvement for prediction accuracy of membrane protein types. The detailed approach is as follows:

Step 1: Extracting features by using the new feature extraction methods.

Step 2: Obtaining the best feature subset and optimal classifier parameters by MIC-GA feature selection.

Step 3: Training the base classifier with the best feature subset and optimal classifier parameters.

Step 4: The meta classifier combines the outputs of the base classifiers and learns again to obtain final result.

By this way, the new feature extraction method can be quickly embedded into our system. From above steps, we find that only meta classifier need to be retrained, and the best feature subsets and optimal classifier parameters for previous extraction methods are unnecessary to be adjusted. The old valid knowledge would not be washed away when learning new knowledge. Our method could be updated accordingly and thus pretty extensible.

IV. CONCLUSION

In the study, fusion representation is used to extract the information from original sequence for predicting membrane protein types. Among these feature extraction methods, SPPI is firstly used in our study. Experimental results indicate that SPPI works well in predicting the types of membrane proteins. For dealing with the curse of high-dimensional features and the problem of feature heterogeneity after fusion representation, we propose the MIC-GA feature selection and ensemble method based on stacked generalization respectively. Numerous positive results prove that our method could contribute to solve these problems. The final experimental results also indicate higher effectiveness of our method for prediction of membrane protein than any existing method do. Furthermore, our system can be incremental learning, any new feature extraction methods would be quickly embedded into our framework to make potential improvement. At last, as demonstrated in a series of researches [9], [15], [24], [27], [30] which provide friendly and convenient webservices for users, we shall make efforts in our future works to offer a webservice based on our method for prediction of membrane protein types.

REFERENCES

[1] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes," *J. Mol. Biol.*, vol. 305, no. 3, pp. 567–580, Jan. 2001.

[2] M. S. Almén, K. J. Nordström, R. Fredriksson, and H. B. Schiöth, "Mapping the human membrane proteome: A majority of the human membrane proteins can be classified according to function and evolutionary origin," *BMC Biol.*, vol. 7, no. 1, p. 50, Aug. 2009.

[3] P. R. Sanders *et al.*, "A set of glycosylphosphatidyl inositol-anchored membrane proteins of *plasmodium falciparum* is refractory to genetic deletion," *Infection Immunity*, vol. 74, no. 7, pp. 4330–4338, Mar. 2006.

[4] Z.-P. Feng, X. Zhang, P. Han, N. Arora, R. F. Anders, and R. S. Norton, "Abundance of intrinsically unstructured proteins in *P. falciparum* and other apicomplexan parasite proteomes," *Mol. Biochem. Parasitol.*, vol. 150, no. 2, pp. 256–267, Dec. 2006.

[5] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, "How many drug targets are there?" *Nature Rev. Drug Discovery*, vol. 5, no. 12, pp. 993–996, Dec. 2006.

[6] K.-C. Chou and D. W. Elrod, "Prediction of membrane protein types and subcellular locations," *Proteins, Struct. Function Bioinf.*, vol. 34, no. 1, pp. 137–153, Jan. 1999.

[7] J. Dev *et al.*, "Structural basis for membrane anchoring of HIV-1 envelope spike," *Science*, vol. 353, no. 6295, p. 172, Jul. 2016.

[8] K. Oxenoid *et al.*, "Architecture of the mitochondrial calcium uniporter," *Nature*, vol. 533, no. 7602, pp. 269–273, May 2016.

[9] K. C. Chou and H. B. Shen, "MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM," *Biochem. Biophys. Res. Commun.*, vol. 360, no. 2, pp. 339–345, Aug. 2007.

[10] L. Rundo *et al.*, "NeXt for neuro-radiosurgery: A fully automatic approach for necrosis extraction in brain tumor MRI using an unsupervised machine learning technique," *Int. J. Imag. Syst. Technol.*, vol. 28, no. 1, pp. 21–37, Nov. 2017.

[11] L. Wang, Z. Yuan, X. Chen, and Z. Zhou, "The prediction of membrane protein types with NPE," *IEICE Electron. Express*, vol. 7, no. 6, pp. 397–402, Mar. 2010.

[12] T. Wang, T. Xia, and X.-M. Hu, "Geometry preserving projections algorithm for predicting membrane protein types," *J. Theor. Biol.*, vol. 262, no. 2, pp. 208–213, Jan. 2010.

[13] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins, Struct. Function Bioinf.*, vol. 43, no. 3, pp. 246–255, May 2001.

[14] C. Huang and J.-Q. Yuan, "A multilabel model based on Chou's pseudo–amino acid composition for identifying membrane proteins with both single and multiple functional types," *J. Membrane Biol.*, vol. 246, no. 4, pp. 327–334, Apr. 2013.

[15] Y.-K. Chen and K.-B. Li, "Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition," *J. Theor. Biol.*, vol. 318, pp. 1–12, Feb. 2013.

[16] F. Ali and M. Hayat, "Classification of membrane protein types using voting feature interval in combination with Chou's pseudo amino acid composition," *J. Theor. Biol.*, vol. 384, pp. 78–83, Nov. 2015.

[17] J. Wang *et al.*, "ProClusEnsem: Predicting membrane protein types by fusing different modes of pseudo amino acid composition," *Comput. Biol. Med.*, vol. 42, no. 5, pp. 564–574, May 2012.

[18] S. Zhang and X. Duan, "Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC," *J. Theor. Biol.*, vol. 437, pp. 239–250, Jan. 2018.

[19] Y.-D. Cai, G.-P. Zhou, and K.-C. Chou, "Support vector machines for predicting membrane protein types by using functional domain composition," *Biophys. J.*, vol. 84, no. 5, pp. 3257–3263, May 2003.

[20] K.-C. Chou and Y.-D. Cai, "Using GO-PseAA predictor to identify membrane proteins and their types," *Biochem. Biophys. Res. Commun.*, vol. 327, no. 3, pp. 845–847, Feb. 2005.

[21] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices1," *J. Mol. Biol.*, vol. 292, no. 2, pp. 195–202, Sep. 1999.

[22] H. Saini, G. Raicar, and S. P. Lal, "Protein fold recognition using genetic algorithm optimized voting scheme and profile bigram," in *Proc. IEEE Int. Conf. Netw., Archit., Storage*, Mar. 2016, pp. 82–89.

[23] S. Ding, Y. Li, Z. Shi, and S. Yan, "A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile," *Biochimie*, vol. 97, no. 2, pp. 60–65, Feb. 2014.

[24] L. Zhang, X. Zhao, and L. Kong, "Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition," *J. Theor. Biol.*, vol. 355, pp. 105–110, Aug. 2014.

[25] L. Zou, C. Nan, and F. Hu, "Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles," *Bioinformatics*, vol. 29, no. 24, pp. 3135–3142, Dec. 2013.

[26] S. Zhang, F. Ye, and X. Yuan, "Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM," *J. Biomol. Struct. Dyn.*, vol. 29, no. 6, pp. 1138–1146, Apr. 2012.

[27] T. T. Wang and J. Yang, "Using the nonlinear dimensionality reduction method for the prediction of subcellular localization of Gram-negative bacterial proteins," *Mol. Diversity*, vol. 13, no. 4, pp. 475–481, 2009.

[28] M. Hayat and A. Khan, "MemHyb: Predicting membrane protein types by hybridizing SAAC and PSSM," *J. Theor. Biol.*, vol. 292, no. 1, pp. 93–102, Jan. 2012.

[29] S. Wang and S. Liu, "Protein sub-nuclear localization based on effective fusion representations and dimension reduction algorithm LDA," *Int. J. Mol. Sci.*, vol. 16, no. 12, pp. 30343–30361, Dec. 2015.

[30] D. Yu *et al.*, "Enhancing membrane protein subcellular localization prediction by parallel fusion of multi-view features," *IEEE Trans. Nanobiosci.*, vol. 11, no. 4, pp. 375–385, Dec. 2012.

[31] G.-S. Han, Z.-G. Yu, and V. Anh, "A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC," *J. Theor. Biol.*, vol. 344, pp. 31–39, Mar. 2014.

[32] T. Wang, J. Yang, H.-B. Shen, and K.-C. Chou, "Predicting membrane protein types by the LLDA algorithm," *Protein Peptide Lett.*, vol. 15, no. 9, pp. 915–921, Feb. 2008.

[33] S. Wang, B. Nie, K. Yue, Y. Fei, W. Li, and D. Xu, "Protein subcellular localization with Gaussian kernel discriminant analysis and its kernel parameter selection," *Int. J. Mol. Sci.*, vol. 18, no. 12, p. 2718, Dec. 2017.

[34] Q. Zou, S. Wan, Y. Ju, J. Tang, and X. Zeng, "Pretata: Predicting TATA binding proteins with novel features and dimensionality reduction strategy," *BMC Syst. Biol.*, vol. 10, no. 4, p. 114, Dec. 2016.

[35] S. Wan, M.-W. Mak, and S.-Y. Kung, "Mem-ADSVM: A two-layer multi-label predictor for identifying multi-functional types of membrane proteins," *J. Theor. Biol.*, vol. 398, pp. 32–42, Mar. 2016.

[36] S. K. Golmohammadi, L. Kurgan, B. Crowley, and M. Reformat, "Classification of cell membrane proteins," in *Proc. Frontiers Converg. Biosci. Inf. Technol.*, Oct. 2007, pp. 153–158.

[37] M. Hayat and A. Khan, "Mem-PHybrid: Hybrid features-based prediction system for classifying membrane protein types," *Anal. Biochem.*, vol. 424, no. 1, pp. 35–44, May 2012.

[38] M. Hayat and A. Khan, "Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition," *J. Theor. Biol.*, vol. 271, no. 1, pp. 10–17, Feb. 2011.

[39] S. Wan, M. W. Mak, and S. Y. Kung, "Mem-mEN: Predicting multi-functional types of membrane proteins by interpretable elastic nets," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 4, pp. 706–718, Aug. 2016.

[40] S.-Q. Wang, J. Yang, and K.-C. Chou, "Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition," *J. Theor. Biol.*, vol. 242, no. 4, pp. 941–946, Oct. 2006.

[41] L. Nanni and A. Lumini, "An ensemble of support vector machines for predicting the membrane protein type directly from the amino acid sequence," *Amino Acids*, vol. 35, no. 3, pp. 573–580, Oct. 2008.

[42] S. Wan, M.-W. Mak, and S.-Y. Kung, "Benchmark data for identifying multi-functional types of membrane proteins," *Data Brief*, vol. 8, pp. 105–107, May 2016.

[43] E. S. Sankari and D. Manimegalai, "Predicting membrane protein types using various decision tree classifiers based on various modes of general PseAAC for imbalanced datasets," *J. Theor. Biol.*, vol. 435, no. 25, pp. 208–217, Dec. 2017.

[44] A. Höglund, P. Dönnes, T. Blum, H.-W. Adolph, and O. Kohlbacher, "MultiLoc: Prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs, and amino acid composition," in *Proc. German Conf. Bioinf.*, 2005, pp. 45–59.

[45] S. Anishetty, G. Pennathur, and R. Anishetty, "Tripeptide analysis of protein structures," *BMC Struct. Biol.*, vol. 2, no. 1, p. 9, Dec. 2002.

[46] W.-S. Bu, Z.-P. Feng, Z. Zhang, and C.-T. Zhang, "Prediction of protein (domain) structural classes based on amino-acid index," *Eur. J. Biochem.*, vol. 266, no. 3, pp. 1043–1049, Dec. 2010.

[47] K. A. Dill, "Theory for the folding and stability of globular proteins," *Biochemistry*, vol. 24, no. 6, pp. 1501–1509, Mar. 1985.

[48] Z.-G. Yu, V. Anh, and K.-S. Lau, "Fractal analysis of measure representation of large proteins based on the detailed HP model," *Phys. A, Stat. Mech. Appl.*, vol. 337, no. 1, pp. 171–184, Jan. 2004.

[49] J. Shen *et al.*, "Predicting protein–protein interactions based only on sequences information," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 11, pp. 4337–4341, Mar. 2007.

[50] A. Sánchez-Flores, E. Pérez-Rueda, and L. Segovia, "Protein homology detection and fold inference through multiple alignment entropy profiles," *Proteins, Struct. Function Bioinf.*, vol. 70, no. 1, pp. 248–256, Jan. 2010.

[51] L. R. Murphy, A. Wallqvist, and R. M. Levy, "Simplified amino acid alphabets for protein fold recognition and implications for folding," *Protein Eng.*, vol. 13, no. 3, pp. 149–152, Mar. 2000.

[52] P. Y. Chou and G. D. Fasman, "Prediction of protein conformation," *Biochemistry*, vol. 13, no. 2, pp. 222–245, Jan. 1974.

[53] S. Basu, A. Pan, C. Dutta, and J. Das, "Chaos game representation of proteins," *J. Mol. Graph. Model.*, vol. 15, no. 5, pp. 279–289, Oct. 1997.

[54] H. B. Rao, F. Zhu, G. B. Yang, Z. R. Li, and Y. Z. Chen, "Update of PROFEAT: A Web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence," *Nucleic Acids Res.*, vol. 39, pp. W385–W390, Jul. 2011.

[55] S. Kawashima, H. Ogata, and M. Kanehisa, "AAindex: Amino acid index database," *Nucleic Acids Res.*, vol. 27, no. 1, pp. 368–369, Jan. 1999.

[56] S. F. Altschul *et al.*, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.

[57] D. N. Reshef *et al.*, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, Dec. 2011.

[58] L. Wei, P. Xing, J. Tang, and Q. Zou, "PhosPred-RF: A novel sequence-based predictor for phosphorylation sites using sequential information only," *IEEE Trans. Nanobiosci.*, vol. 16, no. 4, pp. 240–247, Jun. 2017.

[59] R. Gheshlaghi, M. A. Mahdavi, and A. Maali, "Suitability of sequence-based feature vector for classification algorithm improves accuracy of human protein-protein interaction prediction: A red blood cell case study," *Current Bioinf.*, vol. 11, no. 2, pp. 291–300, 2016.

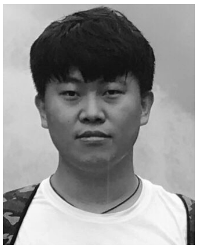
[60] M. S. Nobile *et al.*, "Computational intelligence for parameter estimation of biochemical systems," in *Proc. IEEE Congr. Evol. Comput.*, Jul. 2018, pp. 1–8.

[61] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–249, Dec. 1992.

[62] K. M. Ting and I. H. Witten, "Issues in stacked generalization," *J. Artif. Intell. Res.*, vol. 10, no. 1, pp. 271–289, May 1999.

[63] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[64] H. Nakashima, K. Nishikawa, and T. Ooi, "The folding type of a protein is relevant to the amino acid composition," *J. Biochem.*, vol. 99, no. 1, pp. 153–162, Jan. 1986.



LEI GUO received the B.E. degree from the Yunnan Police Officer Academy, China, in 2016. He is currently pursuing the master's degree with the School of Information Science and Engineering, Yunnan University, China. His research interests include machine learning, data mining, and bioinformatics.

SHUNFANG WANG received the Ph.D. degree in probability theory and mathematical statistics from Yunnan University, China, in 2005. Since 2016, she has been supervising Ph.D. students. She is currently a Professor with the School of Information Science and Engineering, Yunnan University. Her research interests include machine learning, bioinformatics, and applied statistics.

ZHENFENG LEI received the B.S. degree in computer science from Zhengzhou University in 2015 and the M.A.Eng. degree (Hons.) in computer science from Yunnan University in 2017. He is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Xiamen University, Xiamen, China. He had special insights in the field of protein sub-cellular localization. His current research interests include data mining and deep learning techniques, knowledge graph, recommended system, and bio-information for health care. He received the First Prize of the China Graduate Contest on Application, Design, and Innovation of Mobile Terminal in 2018.

XUEREN WANG is currently a Professor with the Department of Statistics, School of Mathematics and Statistics, Yunnan University, China. His research interests include applied mathematics and applied statistics.