Received July 21, 2020, accepted July 26, 2020, date of publication July 31, 2020, date of current version August 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3013409

A Cytokine Protein Identification Model Based on the Compressed PseKRAAC Features

XING GAO, (Member, IEEE), AND GUILIN LI
Department of Software Engineering, School of Informatics, Xiamen University, Xiamen 361005, China
Corresponding author: Guilin Li (glli@xmu.edu.cn)

ABSTRACT Cytokine proteins, which form a complex cytokine regulatory network, participate in a variety of important physiological functions of the human body. Identification of cytokine proteins is very important and has attracted the attention of many researchers. In this paper, we propose a MRMD-cosine model based on the PseKRAAC features to identify the cytokine proteins. First, the PseKRAAC feature extraction method is used to extract four kinds of feature sets from the cytokine proteins, named type1 g-gap, type1 lambda, type2 g-gap and type2 lambda feature sets. Then the MRMD algorithm is used to remove the redundant features from the feature sets. Three kinds of metrics are used by the MRMD algorithm to measure the redundancy of a feature set, which are the Euclidean distance, Cosine similarity and Tanimoto coefficient. Bagging and random forest algorithms are used to construct the classification models based on the compressed feature set. The experimental results show that the MRMD-cosine model based on the type1 lambda feature set constructed by the random forest algorithm can achieve the best performance among all models. Finally, we compare the performance of the MRMD-cosine model with another state-of-art model, named greedy based feature compression model based on the CNT features. It shows that the MRMD-cosine model uses only 15% features of the greedy based model to achieve a better accuracy.

INDEX TERMS MRMD, feature compression, cytokine identification.

I. INTRODUCTION

Cytokine is a kind of low molecular weight soluble protein induced by immunogen, mitogen or other stimulants. It can regulate innate immunity and adaptive immunity, hematopoiesis, cell growth and repair of damaged tissues. Cytokines can be divided into interleukin, interferon, tumor necrosis factor superfamily, colony stimulating factor, chemokines, growth factors and so on. Cytokines have multiple physiological characteristics, such as pleiotropy, overlap, antagonism, synergy and so on. They form a complex regulatory network and participate in a variety of important physiological functions of the human body.

Identification of cytokine proteins is very important and researchers have proposed several kinds of machine learning based models to identify the cytokine proteins [1]–[16]. As the number of features extracted from the cytokine data set is large, some kinds of feature selection methods are used to compress the feature set [23]–[35]. In paper [36], a greedy based feature compression model based on the CNT feature set is proposed to classify the cytokine proteins. The greedy based model, constructed by the SVM (Support Vector Machine) algorithm [37]–[53], is composed of 167 features and can achieves the accuracy of 87.3%.

In this paper, we utilize the PseKRAAC methods [54] to extract features from the cytokine proteins to construct the classification models. Four kinds of feature sets are extracted by the PseKRAAC methods, which are the type1 g-gap feature set, the type1 lambda feature set, the type2 g-gap feature set and the type2 lambda feature set. There are 155 features in the original feature set, which means some redundant features are contained in the feature set. Then the MRMD (Max Relevance Max Distance) based dimensionality reduction algorithm [55] is used to compress the four feature sets. MRMD wants to find the feature subset with maximum relevance with the classification, and the maximum distance between features in the subset at the same time. Such kind of feature set has strong correlation with classification and low redundancy within the feature set. MRMD utilizes three kinds of metrics to evaluate the redundancy of features in the feature set, which are the Euclidean distance, Cosine similarity and Tanimoto coefficient. Then two machine learning algorithms, bagging and random forest, are used to construct the classification models to identify the cytokines based on the compressed

The associate editor coordinating the review of this manuscript and approving it for publication was Leyi Wei.

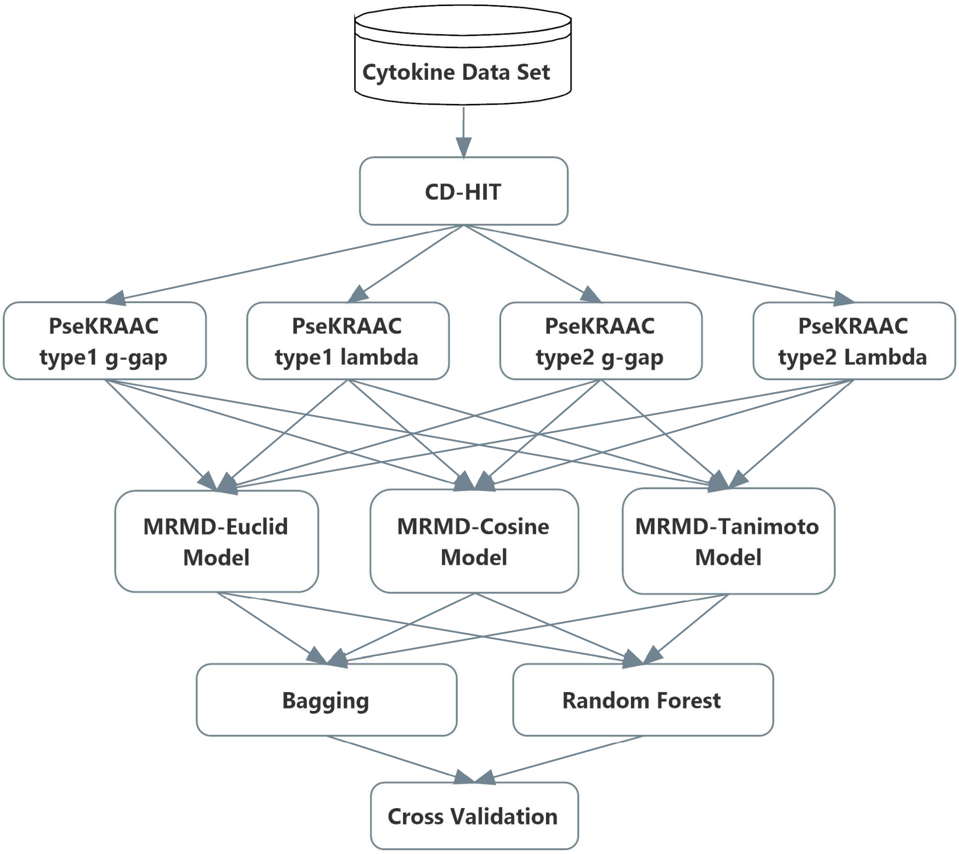


FIGURE 1. Framework for cytokine protein identification.

PseKRAAC feature set. Finally, a MRMD-Cosine model based on the type1 lambda feature set constructed by the random forest algorithm achieves the best performance among all models. Finally, we compare the performance of the MRMD-cosine model with the greedy based model based on the CNT feature set. It shows that the accuracy of the MRMD-cosine model is 87.7%, which is better than that of the greedy based model. Furthermore, the number of features used by the MRMD-cosine model is only 25, which is much smaller than that of the greedy based model.

The contributions of the paper are as follows. (1) A MRMD-cosine model based on the type1 lambda feature set constructed by the random forest algorithm is proposed to classify the cytokine proteins. (2) Compared with a state-of-art greedy based model, the MRMD-cosine model uses only 15% features of the greedy based model to achieve a better accuracy.

The organization of the paper is as follows: in section 2, we introduce the methods to construct the cytokine identification model. In section 3, five groups of experiments are done to evaluate the performance of models proposed in this paper. Finally, conclusions are drawn.

II. METHODS

Figure 1 shows the construction procedure of the cytokine identification model. First, the cytokines data set are processed by the CD-HIT algorithm [56], which remove the redundant instances to balance the number of positive and negative instances in the data set. Then the PseKRAAC method is used to extract four kinds of features from the cytokine data set, which are the type1 g-gap feature set, the type1 lambda feature set, the type2 g-gap feature set and the type2 lambda feature set. There are 155 features contained in each kind of feature set. Then the MRMD feature compression algorithm is used to compress the four feature sets. MRMD utilizes three kinds of metrics to evaluate the redundancy of features in the feature set, which are the Euclidean distance, Cosine similarity and Tanimoto coefficient. After the four feature sets are compressed by the MRMD algorithm, bagging and random forest algorithm are used to construct the machine learning model to classify the cytokines. Finally, the cross validation method is used to evaluate the performance of the classification models.

A. DATASET

The positive instances of cytokine data set are downloaded from the Uniprot [57]–[59] database. The negative instance data set is constructed according to the PFAM families of the positive instance. The longest proteins of all PFAM families, except the PFAM families of the positive instances, are extracted from the Uniprot database to form the negative instance data set. To balance the number of positive and negative instances in the data set, the CD-HIT algorithm is used to delete the redundant negative instances from the data set. There are 9299 negative instances and 9645 positive instances in the cytokine data set.

B. PseKRAAC FEATURE EXTRACTION METHOD

As we all know, there are 20 kinds of amino acids that make up proteins. It is a common feature extraction method to classify protein sequences based on their structural characteristics [60]–[65]. However, due to the large number of features caused by 20 kinds of amino acids and their combinations, people hope to classify 20 kinds of amino acids according to their physical and chemical characteristics. Since each class contains many amino acids with similar properties, the amino acids belonging to one class can be treated as a whole, which is called RAAC (Reduced AAC), so the number of features can be reduced [66]. Such kind of feature extraction method is called PseKRAAC (Pseudo K-tuple Reduced Amino Acids Composition) descriptor [67], [68]. 16 kinds of classification methods are listed in [69]. In this paper, two classification methods, type1 and type2, were used to extract features from cytokine protein. Two methods, g-gap and lambda correlation [70], have been proposed to describe the structural character of RAAC in proteins. Therefore, based on the above two RAAC types and two structural character description methods, we can extract the four feature sets: type1 g-gap, type1 lambda, type2 g-gap and type2 lambda.

C. MRMD FEATURE COMPRESSION ALGORITHM

Feature selection is widely used to select important features aiming to improve the predictive performance [71]–[75]. The main idea of MRMD (Max Relevance Max Distance based dimensionality reduction) is as follows. First, it calculates the correlation between each feature and classification by Pearson coefficient, which means Max Relevance. Second, the distance between features in the feature set is calculated to find the feature set with low redundancy. The larger the distance between features, the lower the correlation between them and the lower redundancy of the selected feature set. MRMD provides three methods to calculate the distance between features: Euclidean distance1, Cosine similarity2 and Tanimoto coefficient3. The feature subset selected based on the above idea has strong correlation with classification and low redundancy within the feature set.

$$ED(\vec{X}, \vec{Y}) = \sqrt{\sum_{k=1}^{N} (x_k - y_k)^2} \tag{1}$$

$$\mathrm{COS}(\vec{X}, \vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{\|\vec{X}\| \cdot \|\vec{Y}\|} \tag{2}$$

$$\mathrm{TC}(\vec{X}, \vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{\|\vec{X}\|^2 + \|\vec{Y}\|^2 - \vec{X} \cdot \vec{Y}} \tag{3}$$

D. BAGGING AND RANDOM FOREST ALGORITHMS

For a given training set S, m training samples are extracted from S by boosting sampling in each round. After N rounds are conducted and N sample sets are obtained. It should be noted that the N training sets are independent of each other. A sample set is used to construct a prediction model by some kinds of machine learning algorithm each time. And we get N prediction models. To solve the classification problem, the N models vote to get the classification results. All kinds of machine learning algorithms can be used to construct the prediction models for bagging. Random Forest is a special case of bagging, which use the decision tree as the machine learning algorithm to construct the prediction model [76]–[83].

III. EXPERIMENTS

In this section, five groups of experiments are done to test the performance of models constructed by combining different kinds of feature sets, feature redundancy metrics and machine learning algorithms. The four feature sets extracted by PsekRAAC methods are type1 g-gap, type1 lambda, type2 g-gap and type2 lambda. For each feature set, three kinds of feature redundancy metrics are used, which are the Euclidean distance, Cosine similarity and Tanimoto coefficient. Finally, the bagging and random forest algorithm are used to construct the classification model.

TABLE 1. Parameters set for the experiments.

Algorithm	Parameter Name	Value
Bagging	bagSizePercent	100
	calcOutOfBag	False
	numExecutionSlots	1
	numIterations	10
	classifier	REPTree
REPTree	maxDepth	-1
	minNum	2
	minVarianceProp	0.001
	numDecimalPlaces	2
Random Forest	maxDepth	0
	numTrees	100
	numFeatures	0

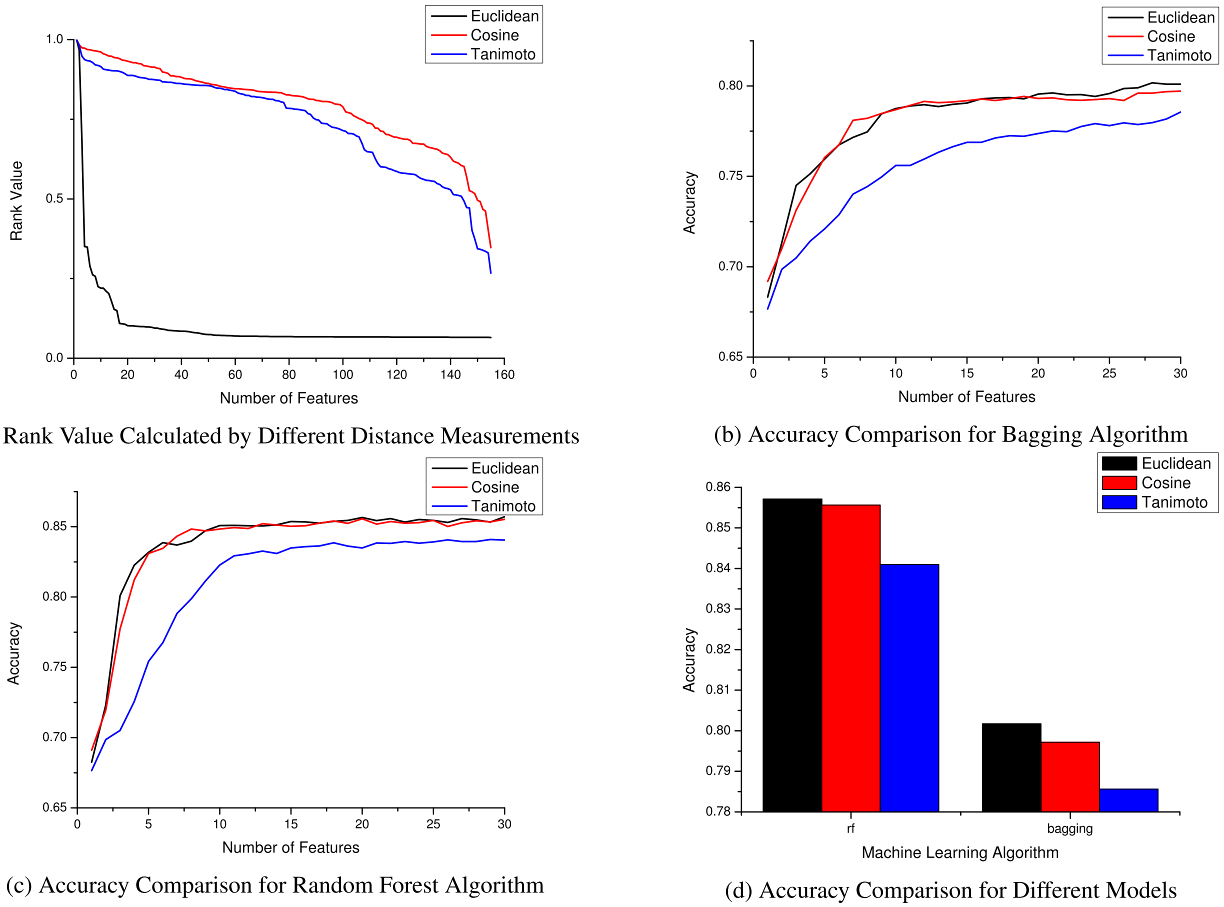
Accuracy (ACC), defined by Formula (4), is used to evaluate the performance of all classification models. The 10-fold cross-validation is used to calculate the accuracy for each model. Weka [84] is used to do all the experiments. Details of the parameters used in the experiments are shown in table 1.

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

where TP represents the True Positive, FP represents False Positive, TN represents true negative, and FN represents False Negative.

A. PERFORMANCE FOR THE PseKRAAC FEATURE SET OF TYPE1 G-GAP

Firstly, the MRMD feature compression method utilizes the Euclidean distance, the Cosine similarity and the Tanimoto



(a) Rank Value Calculated by Different Distance Measurements

(b) Accuracy Comparison for Bagging Algorithm

(c) Accuracy Comparison for Random Forest Algorithm

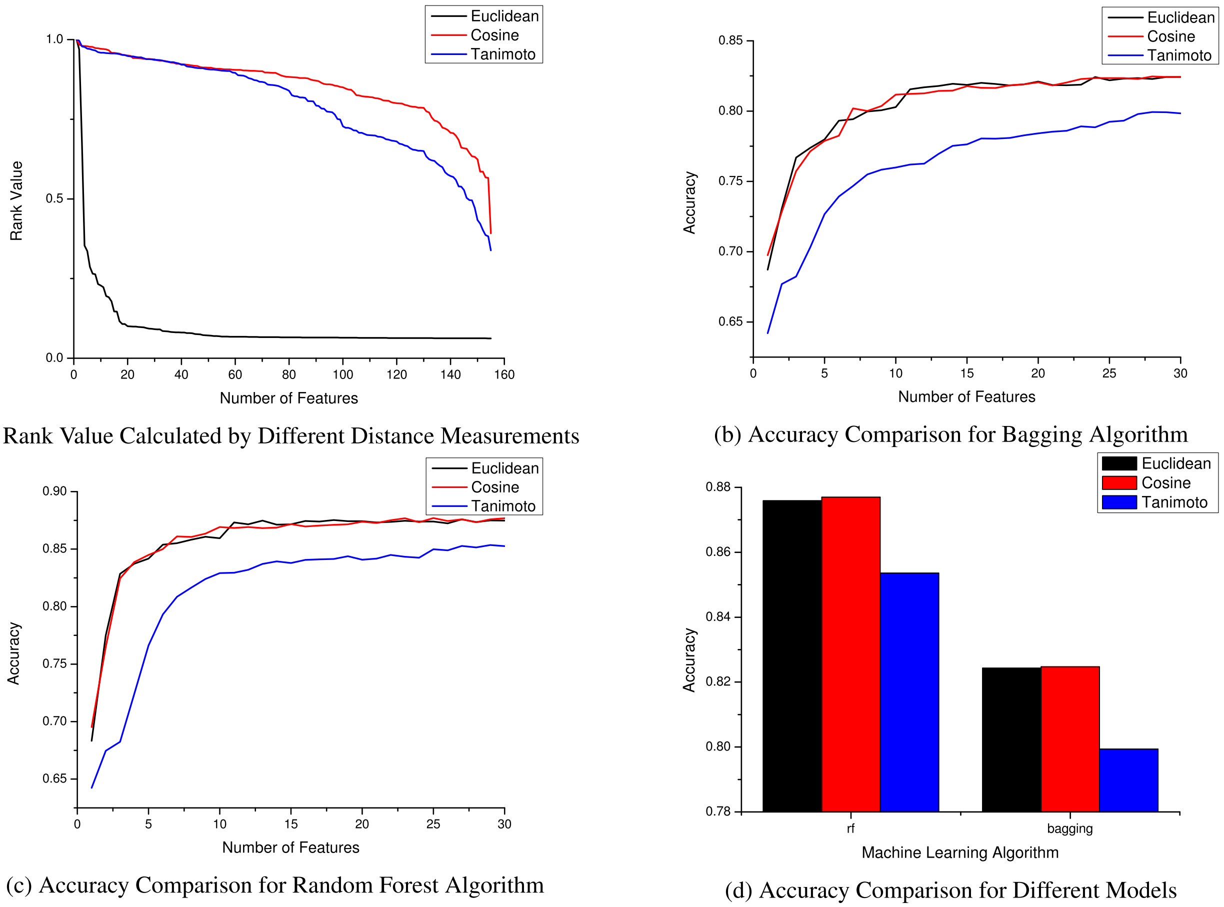
(d) Accuracy Comparison for Different Models

FIGURE 2. Performance comparison for PseKRAAC feature set of type1 G-gap.

coefficient to measure the redundancy among features. The results are shown in Figure 2. The rank values of the 155 features calculated by MRMD method are shown in Figure 2a. The experimental results show that the range of rank values calculated by the three methods for each feature is between 0 and 1, and the more important the feature, the larger the rank value. When Euclidean distance is used as the measurement method, the rank value of features decays rapidly. The rank value of the first 17 features decays rapidly from 1 to 0.109. When the feature number exceeds 24, the change rate of rank values is very small. However, the change rate of rank values calculated by cosine similarity and Tanimoto coefficient is much gentler than that of the Euclidean distance. With the increase of the feature number, the rank values of the features decrease approximately linearly. The rank values after 100 features show accelerated decay. At the same time, the rank values calculated by Cosine similarity are slightly larger than those calculated by Tanimoto coefficient.

Secondly, according to the rank values of each feature calculated by the three distance methods, we select the features from 1 to 30 in order, and form the feature set. Bagging and random forest algorithms are used to identify cytokine proteins, and accuracy is used as the metric to evaluate different classification models. The experimental results are shown in Figure 2b and Figure 2c. As shown in Figure 2b, in which bagging is used as the classification algorithm, when the feature number in the feature set is small, the classification accuracy is poor. With the increase of the feature number, the classification accuracy has been improved significantly. When the feature number in the feature set reaches a certain number, the improvement of classification accuracy is very limited by adding new features to the feature set. The performance comparison among the three distance methods is as follows. The classification models using Euclidean distance and Cosine similarity as the measurement achieves similar classification accuracy. The results also show that the model using the Tanimoto coefficient method is the worse than the other two models, and there is a big gap between it and the other two methods. The experimental results also show that increasing the feature number in the feature set has a great impact on the improvement of the classification performance of the Tanimoto based model. However, if there are too many features in the feature set, the running time of the classifier will be greatly affected in the face of the classification task with large amount of data.

As shown in Figure 2c, when random forest is used as the classification algorithm, the overall performance of models with different distance measurements is similar to that of



(a) Rank Value Calculated by Different Distance Measurements

(b) Accuracy Comparison for Bagging Algorithm

(c) Accuracy Comparison for Random Forest Algorithm

(d) Accuracy Comparison for Different Models

FIGURE 3. Performance comparison for PseKRAAC feature set of type1 lambda.

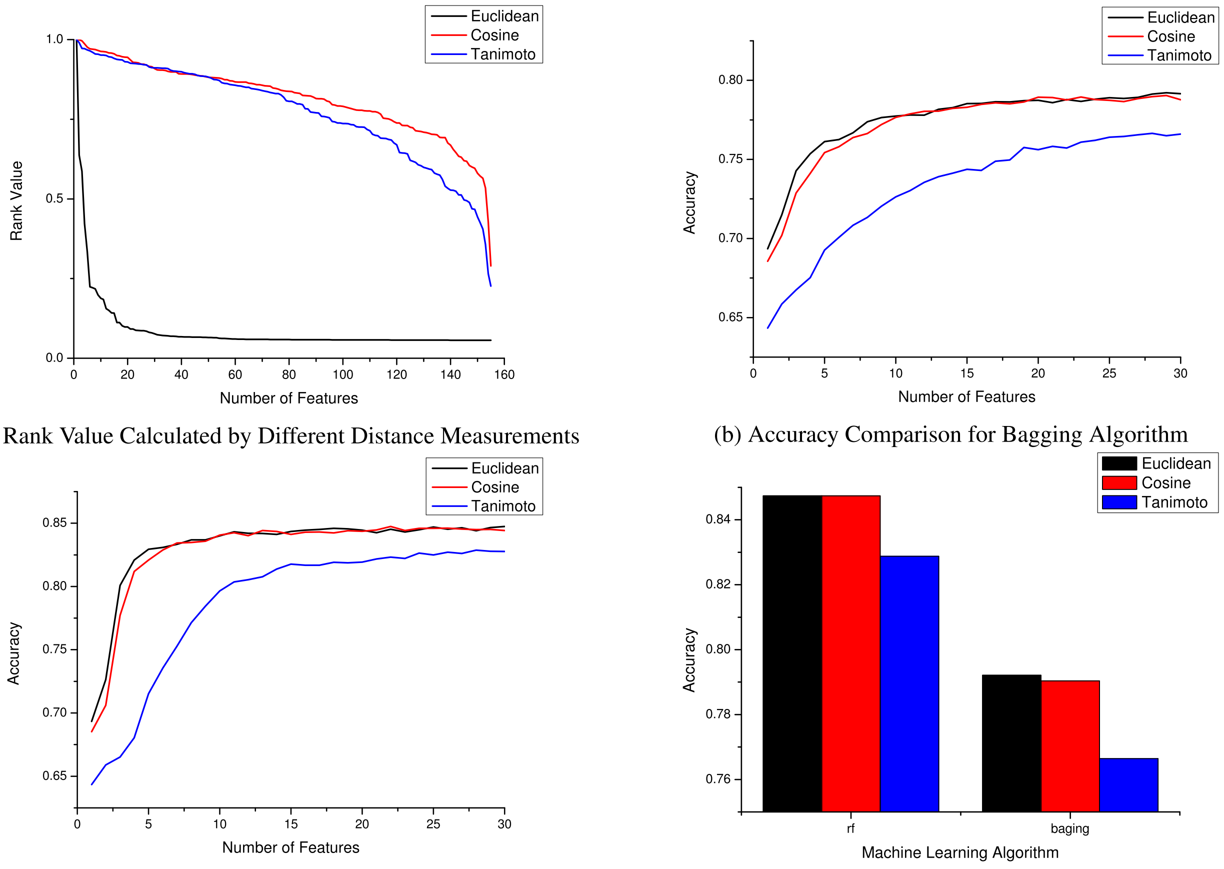
bagging algorithm. It should be noted that when using random forest algorithm, with the increase of the feature number in the feature set, the classification accuracy of the random forest algorithm is significantly better than that of the bagging algorithm. For example, when the feature number is 9, the classification accuracy of Euclidean model and Cosine model using random forest classification algorithm can reach 84.7%. While the classification accuracy of Euclidean model and Cosine model using bagging classification algorithm is only 78.5%, with a difference of 6.2%. In terms of classification accuracy, the accuracy of the classification models using random forest algorithm is significantly better than that of the classification models using bagging algorithm.

In order to compare the impact of the two machine learning algorithms on classification accuracy, we compare the highest accuracy achieved by all the Euclidean, cosine and Tanimoto based models constructed with random forest and bagging algorithm as classifier respectively. The comparison results are shown in Figure 2d. It shows that the classification accuracy of random forest based model is obviously better than that based on bagging algorithm. At the same time, the model using Euclidean distance as the redundancy measurement among features is the best, the model using cosine similarity takes the second place, and there is a small gap with Euclidean distance. The performance of Tanimoto coefficient is the worst, and there is a big gap between it and the other two models. Therefore, for the feature set extracted by type1 g-gap method, the random forest based model with the Euclidean distance achieves the best accuracy.

B. PERFORMANCE FOR THE PseKRAAC FEATURE SET OF TYPE1 LAMBDA

The experimental results for the PseKRAAC Feature Set of Type1 Lambda are shown in Figure 3. The rank values of the 155 features calculated by MRMD method are shown in Figure 3a. The experimental results show that the rank value of the first 18 features decays rapidly from 1 to 0.107. When the feature number exceeds 22, the change rate of rank values is very small. The change rate of rank values calculated by cosine similarity and Tanimoto coefficient decrease approximately linearly. After 100 features, the rank values drop accelerately. At the same time, the rank values of Cosine similarity are slightly larger than those of Tanimoto coefficient.

The experimental results by using Bagging and random forest algorithms are shown in Figure 3b and Figure 3c. As shown in Figure 3b, in which bagging is used as the classification algorithm. The classification models using Euclidean



(a) Rank Value Calculated by Different Distance Measurements

(b) Accuracy Comparison for Bagging Algorithm

(c) Accuracy Comparison for Random Forest Algorithm

(d) Accuracy Comparison for Different Models

FIGURE 4. Performance comparison for PseKRAAC feature set of type2 G-gap.

distance and Cosine similarity as the measurement achieves similar classification accuracy. The model using the Tanimoto coefficient method is worse than the other two models, and there is a big gap between it and the other two methods.

As shown in Figure 3c, when random forest is used as the classification algorithm, the classification accuracy of the random forest algorithm is significantly better than that of the bagging algorithm. When the feature number is 7, the classification accuracy of Cosine model using random forest classification algorithm can reach 86.1%. While the classification accuracy of Euclidean model using bagging classification algorithm is only 80.2%, with a difference of 5.9%. In terms of classification accuracy, the accuracy of the classification models using random forest algorithm is significantly better than that of the classification models using bagging algorithm.

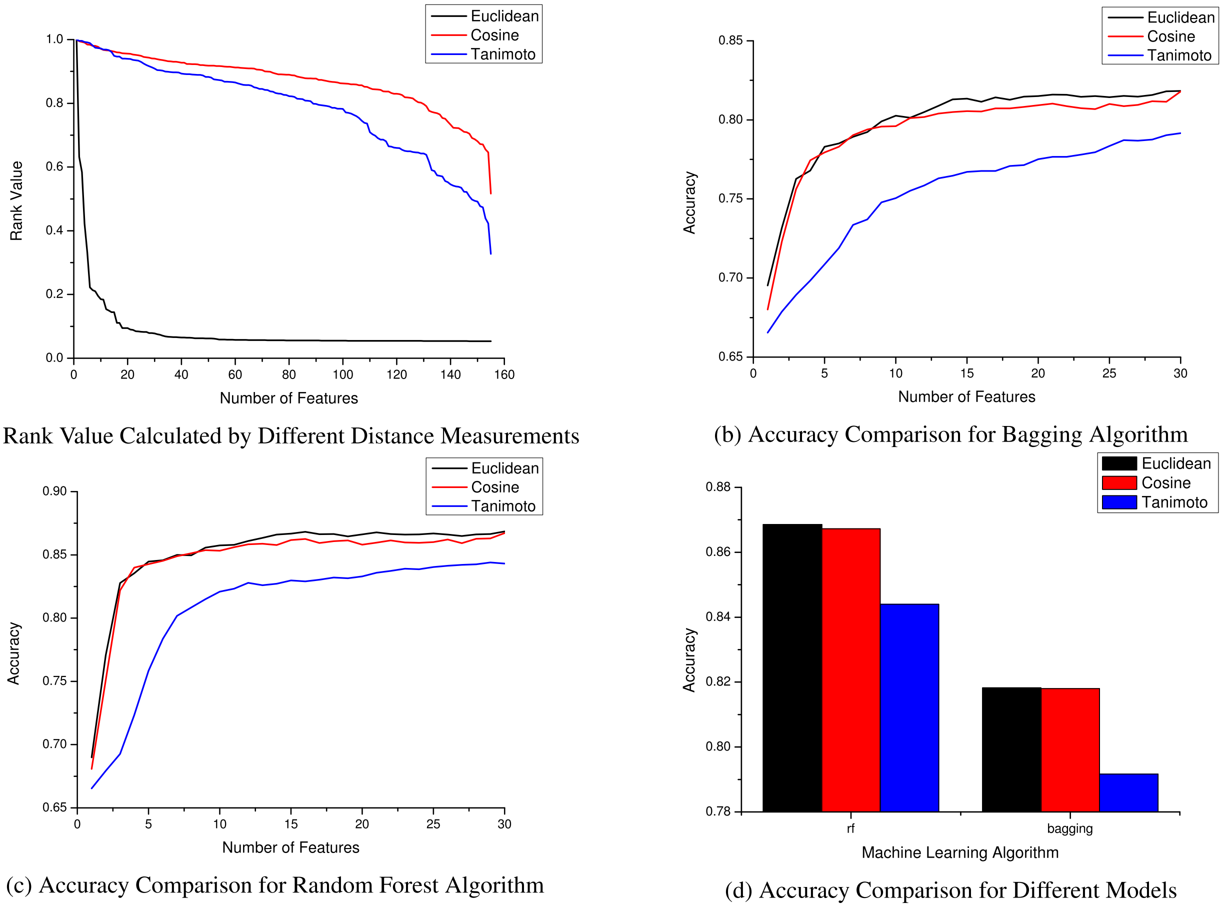
The highest accuracy achieved by the Euclidean, cosine and Tanimoto based models constructed with random forest and bagging algorithm as classifier are compared in Figure 3d. It shows that even the worst model of random forest, which is the Tanimoto based model, is better than the best model of bagging algorithm. Therefore, for the feature set extracted by type1 lambda method, the random forest based model with the Cosine similarity achieves the best accuracy.

C. PERFORMANCE FOR PseKRAAC FEATURE SET OF TYPE2 G-GAP

The experimental results for the PseKRAAC Feature Set of Type2 G-gap are shown in Figure 4. The rank values of the 155 features calculated by MRMD method are shown in Figure 4a. It shows that the rank value drops rapidly from 1 to 0.101, that takes only 18 features. With the increase of the feature number, the change rate of rank values becomes very small. The change rate of rank values of cosine similarity and Tanimoto coefficient decrease smoothly at first. After 140 features, the rank values drop significantly.

The experimental results by using Bagging and random forest algorithms are shown in Figure 4b and Figure 4c. As shown in Figure 4b, in which bagging is used as the classification algorithm. The classification accuracy of Euclidean based models are better than that of the Cosine based models when the feature number is small. When the feature number exceeds 10, the accuracy of the two kinds of models are very similar. The model based on the Tanimoto coefficient method is worse than the other two models.

As shown in Figure 4c, when random forest is used as the classification algorithm, the classification accuracy of the random forest algorithm is significantly better than that of the bagging algorithm. The classification accuracy of Euclidean



(a) Rank Value Calculated by Different Distance Measurements

(b) Accuracy Comparison for Bagging Algorithm

(c) Accuracy Comparison for Random Forest Algorithm

(d) Accuracy Comparison for Different Models

FIGURE 5. **Performance comparison for PseKRAAC feature set of type2 lambda.**

based models are very similar. The model based on the Tanimoto coefficient method is worse than the other two models.

The highest accuracy achieved by the Euclidean, Cosine and Tanimoto based models constructed with random forest and bagging algorithm as classifier are compared in Figure 4d. It shows that the random forest based model with the Euclidean distance achieves the best accuracy.

D. PERFORMANCE FOR THE PseKRAAC FEATURE SET OF TYPE2 LAMBDA

The experimental results for the PseKRAAC Feature Set of Type2 G-gap are shown in Figure 5. The rank values of the 155 features calculated by MRMD method are shown in Figure 5a. It shows that the rank value drops rapidly from 1 to 0.095, that takes only 18 features. With the increase of the feature number, the change rate of rank values becomes very small. The change rate of rank values of cosine similarity and Tanimoto coefficient decrease smoothly at first. After 130 features, the rank values drop accelerately.

The experimental results by using Bagging and random forest algorithms are shown in Figure 5b and Figure 5c. As shown in Figure 5b, in which bagging is used as the classification algorithm. When the feature number is below 9, the accuracy of the two kinds of models are very similar. The accuracy of Euclidean based models are better than that of the Cosine based models when the feature number exceeds 9. The model based on the Tanimoto coefficient method is worse than the other two models.

As shown in Figure 5c, when random forest is used as the classification algorithm, the accuracy of the random forest algorithm is significantly better than that of the bagging algorithm. The accuracy of the Euclidean based model is the best.

The highest accuracy achieved by the Euclidean, Cosine and Tanimoto based models constructed with random forest and bagging algorithm as classifier are compared in Figure 5d. It shows that the random forest based model with the Euclidean distance achieves the best accuracy.

E. COMPARISON OF DIFFERENT KINDS OF CLASSIFICATION MODELS

In this section, we compare the performance of the classification models for the four feature sets with the best accuracy. The four experiments above show that, for all feature sets, the models constructed by the random forest algorithm is better than that constructed by the bagging algorithm. Furthermore, for the type1 g-gap, type2 g-gap and type2 lambda feature sets, the Euclidean distance metric

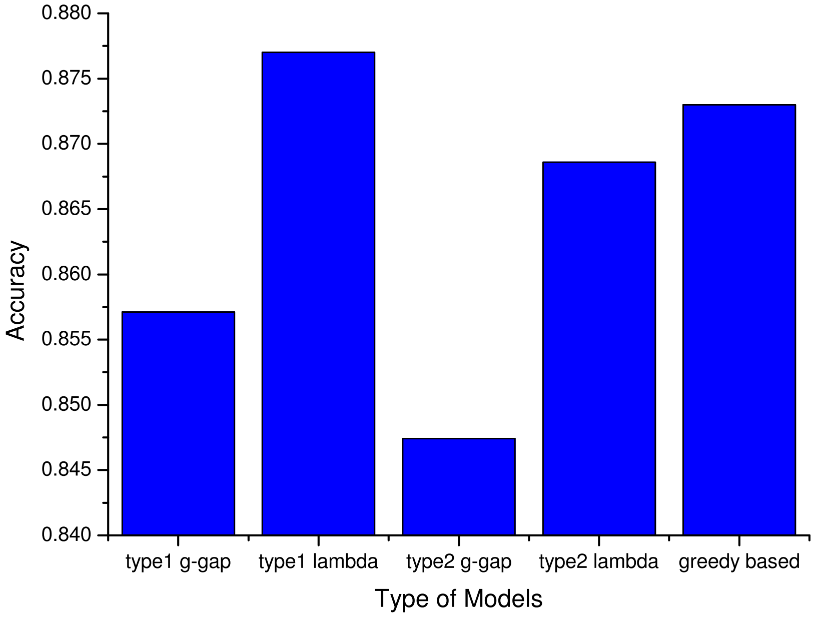


FIGURE 6. Comparison among different kinds of classification models.

achieves the best accuracy. For type1 lambda feature set, the Cosine similarity metric achieves the best accuracy. The four accuracy are compared in Figure 6, which shows that the feature set of type1 lambda is the best among all the models. We can conclude that the type1 lambda model processed by MRMD-cosine compression and random forest algorithm is the best model.

In paper, a greedy based feature compression model based on the CNT feature set is proposed to classify the cytokine proteins. The greedy based model, constructed by the SVM (Support Vector Machine) algorithm, is composed of 167 features and can achieves the accuracy of 87.3%. The accuracy of MRMD-cosine model proposed in this paper is 87.7%, which is composed of 25 features. It means that the MRMD-cosine model uses only 15% features of the greedy based model to achieve a better accuracy.

IV. CONCLUSION

In this paper, four kinds of PseKRAAC based feature sets are extracted from the cytokine data set. Three kinds of feature redundancy calculation metrics, which are the Euclidean distance, Cosine similarity and Tanimoto coefficient, are used to compress the features in the feature set by the MRMD feature compression algorithm. The bagging and random forest algorithm are used to construct the machine learning model to classify the cytokines. The experimental results show that the MRMD-cosine model based on type1 lambda feature set is the best model, which uses only 15% features of the greedy based model to achieve a better accuracy.

REFERENCES

[1] Q. Zou, W. Chen, Y. Huang, X. Liu, and Y. Jiang, "Identifying multi-functional enzyme by hierarchical multi-label classifier," *J. Comput. Theor. Nanosci.*, vol. 10, no. 4, pp. 1038–1043, Apr. 2013.

[2] P. K. Papasaikas, P. G. Bagos, Z. I. Litou, and S. J. Hamodrakas, "A novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden Markov models," *SAR QSAR Environ. Res.*, vol. 14, nos. 5–6, pp. 413–420, Oct. 2003.

[3] Y. Yabuki, T. Hirokawa, H. Mukai, and M. Suwa, "GRIFFIN: A system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model," *Nucleic Acids Res.*, vol. 33, no. 2, pp. 148–153, 2005,

[4] C.-H. Lu and J.-K. Hwang, "Prediction of protein subcellular localization. Proteins: Structure," *Function Genetics*, vol. 64, no. 3, pp. 643–651, 2006.

[5] H. Nielsen JE, S. Brunak, and G. von Heijne, "A neural network method for identifcation of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites," *Int. J. Neural Syst.*, vol. 8, nos. 5–6, pp. 581–599, 1997.

[6] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[7] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: Gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA," *RNA*, vol. 25, no. 2, pp. 205–218, Feb. 2019.

[8] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990.

[9] W. R. Pearson, "Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the smith-waterman and FASTA algorithms," *Genomics*, vol. 11, no. 3, pp. 635–650, Nov. 1991.

[10] G. S. Ladics, G. A. Bannon, A. Silvanovich, and R. F. Cressman, "Comparison of conventional FASTA identity searches with the 80 amino acid sliding window FASTA search for the elucidation of potential identities to known allergens," *Mol. Nutrition Food Res.*, vol. 51, no. 8, pp. 985–998, Aug. 2007.

[11] N. Huang, H. Chen, and Z. Sun, "CTKPred: An SVM-based method for the prediction and classification of the cytokine superfamily," *Protein Eng., Des. Selection*, vol. 18, no. 8, pp. 365–368, Aug. 2005.

[12] S. Lata and G. P. S. Raghava, "CytoPred: A server for prediction and classification of cytokines," *Protein Eng. Des. Selection*, vol. 21, no. 4, pp. 279–282, Jan. 2008.

[13] C. Z. Cai, "SVM-prot: Web-based support vector machine software for functional classification of a protein from its primary sequence," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3692–3697, Jul. 2003.

[14] L. Yu, F. Xu, and L. Gao, "Predict new therapeutic drugs for hepatocellular carcinoma based on gene mutation and expression," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 8, Jan. 2020.

[15] X. Zeng, S. Yuan, X. Huang, and Q. Zou, "Identification of cytokine via an improved genetic algorithm," *Frontiers Comput. Sci.*, vol. 9, no. 4, pp. 643–651, Aug. 2015.

[16] Q. Zou, "An approach for identifying cytokines based on a novel ensemble classifier," *BioMed Res. Int.*, vol. 2013, Oct. 2013, Art. no. 686090.

[17] Y. Liu, F. Tang, and Z. Zeng, "Feature selection based on dependency margin," *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1209–1221, Jun. 2015.

[18] Z. HLaZ, "Manipulating data and dimension reduction methods: Feature selection," in *Encyclopedia of Complexity and Systems Science*. Berlin, Germany. Springer 2009, 5348–5359.

[19] H. Liu HM, R. Setiono, and Z. Zhao, "Feature selection: An ever evolving frontier in data mining," in *Proc. JMLR*, Hyderabad, India, 2010, pp. 4–13.

[20] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

[21] A. Unler and A. Murat, "A discrete particle swarm optimization method for feature selection in binary classification problems," *Eur. J. Oper. Res.*, vol. 206, no. 3, pp. 528–539, Nov. 2010.

[22] al. YLe, "An improved particle swarm optimization for feature selection," *J. Bionic. Eng.*, vol. 8, no. 2, pp. 191–200, 2011.

[23] Y. Zhao, F. Wang, S. Chen, J. Wan, and G. Wang, "Methods of microRNA promoter prediction and transcription factor mediated regulatory network," *Biomed. Res. Int.*, vol. 2017, Apr. 2017, Art. no. 7049406.

[24] L. Cheng, P. Wang, R. Tian, S. Wang, Q. Guo, M. Luo, W. Zhou, G. Liu, H. Jiang, and Q. Jiang, "LncRNA2Target v2.0: A comprehensive database for target genes of lncRNAs in human and mouse," *Nucleic Acids Res.*, vol. 47, no. 1, pp. 140–s144, Jan. 2019.

[25] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, "Prediction and validation of disease genes using HeteSim scores," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 3, pp. 687–695, May 2017.

[26] L. Yu, J. Zhao, and L. Gao, "Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome," *Artif. Intell. Med.*, vol. 77, pp. 53–63, Mar. 2017.

[27] L. Yu, B. Wang, X. Ma, and L. Gao, "The extraction of drug-disease correlations based on module distance in incomplete human interactome," *BMC Syst. Biol.*, vol. 10, no. 4, p. 111, Dec. 2016.

[28] Y. Qiao, Y. Xiong, H. Gao, X. Zhu, and P. Chen, "Protein-protein interface hot spots prediction based on a hybrid feature selection strategy," *BMC Bioinf.*, vol. 19, no. 1, p. 14, Dec. 2018.

[29] J.-X. Tan, S.-H. Li, Z.-M. Zhang, C.-X. Chen, W. Chen, H. Tang, and H. Lin, "Identification of hormone binding proteins based on machine learning methods," *Math. Biosci. Eng.*, vol. 16, no. 4, pp. 2466–2480, 2019.

[30] X. Zeng, W. Wang, C. Chen, and G. G. Yen, "A consensus community-based particle swarm optimization for dynamic community detection," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2502–2513, Jun. 2020, doi: 10.1109/TCYB.2019.2938895.

[31] H. Xu, W. Zeng, D. Zhang, and X. Zeng, "MOEA/HD: A multiobjective evolutionary algorithm based on hierarchical decomposition," *IEEE Trans. Cybern.*, vol. 49, no. 2, pp. 517–526, Feb. 2019.

[32] H. Xu, W. Zeng, X. Zeng, and G. G. Yen, "An evolutionary algorithm based on minkowski distance for many-objective optimization," *IEEE Trans. Cybern.*, vol. 49, no. 11, pp. 3968–3979, Nov. 2019.

[33] T. Song, A. Rodriguez-Paton, P. Zheng, and X. Zeng, "Spiking neural p systems with colored spikes," *IEEE Trans. Cognit. Develop. Syst.*, vol. 10, no. 4, pp. 1106–1115, Dec. 2018.

[34] X. Chen, M. J. Pérez-Jiménez, L. Valencia-Cabrera, B. Wang, and X. Zeng, "Computing with viruses," *Theor. Comput. Sci.*, vol. 623, pp. 146–159, Apr. 2016.

[35] L. Yu, S. Yao, L. Gao, and Y. Zha, "Conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments," *Frontiers Genet.*, vol. 9, p. 745, Jan. 2019.

[36] G. Li and X. Gao, "The feature compression algorithms for identifying cytokines based on CNT features," *IEEE Access*, vol. 8, pp. 83645–83654, 2020.

[37] X.-J. Zhu, C.-Q. Feng, H.-Y. Lai, W. Chen, and L. Hao, "Predicting protein structural classes for low-similarity sequences by evaluating different features," *Knowl.-Based Syst.*, vol. 163, pp. 787–793, Jan. 2019.

[38] H. Wang, Y. Ding, J. Tang, and F. Guo, "Identification of membrane protein types via multivariate information fusion with Hilbert–Schmidt independence criterion," *Neurocomputing*, vol. 383, pp. 257–269, Mar. 2020.

[39] Y. Xiong, Q. Wang, J. Yang, X. Zhu, and D.-Q. Wei, "PredT4SE-stack: Prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method," *Frontiers Microbiol.*, vol. 9, p. 2571, Oct. 2018.

[40] Y. Xiong, J. Liu, W. Zhang, and T. Zeng, "Prediction of heme binding residues from protein sequences with integrative sequence profiles," *Proteome Sci.*, vol. 10, no. 1, p. S20, 2012.

[41] X. Zhu, J. He, S. Zhao, W. Tao, Y. Xiong, and S. Bi, "A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of saccharomyces cerevisiae," *Briefings Funct. Genomics*, vol. 15, pp. 367–376, Oct. 2019.

[42] Z. Liao, D. Li, X. Wang, L. Li, and Q. Zou, "Cancer diagnosis through IsomiR expression with machine learning method," *Current Bioinf.*, vol. 13, no. 1, pp. 57–63, Feb. 2018.

[43] L. Chao, L. Wei, and Q. Zou, "SecProMTB: A SVM-based classifier for secretory proteins of mycobacterium tuberculosis with imbalanced data set," *Proteomics*, vol. 2019, Apr. 2019, Art. no. e1900007.

[44] H. Bu, J. Hao, J. Guan, and S. Zhou, "Predicting enhancers from multiple cell lines and tissues across different developmental stages based on SVM method," *Current Bioinf.*, vol. 13, no. 6, pp. 655–660, Nov. 2018.

[45] C. Meng, S. Jin, L. Wang, F. Guo, and Q. Zou, "AOPs-SVM: A sequence-based classifier of antioxidant proteins using a support vector machine," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 224, Sep. 2019.

[46] L. Wei, Q. Zou, M. Liao, H. Lu, and Y. Zhao, "A novel machine learning method for cytokine-receptor interaction prediction," *Combinat. Chem. High Throughput Screening*, vol. 19, no. 2, pp. 144–152, Jan. 2016.

[47] B. Liu, C.-C. Li, and K. Yan, "DeepSVM-fold: Protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks," *Briefings Bioinform.*, Oct. 2019, doi: 10.1093/bib/bbz098.

[48] B. Liu, S. Chen, K. Yan, and F. Weng, "IRO-PsekGCC: Identify DNA replication origins based on pseudo k-Tuple GC composition," *Frontiers Genet.*, vol. 10, p. 842, Sep. 2019.

[49] Y. Cao, S. Wang, Z. Guo, T. Huang, and S. Wen, "Synchronization of memristive neural networks with leakage delay and parameters mismatch via event-triggered control," *Neural Netw.*, vol. 119, pp. 178–189, Nov. 2019.

[50] X. Zeng, N. Ding, A. Rodríguez-Patón, and Q. Zou, "Probability-based collaborative filtering model for predicting gene–disease associations," *BMC Med. Genomics*, vol. 10, no. S5, p. 76, Dec. 2017.

[51] X. Zhang, Q. Zou, A. Rodriguez-Paton, and X. Zeng, "Meta-path methods for prioritizing candidate disease miRNAs," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 1, pp. 283–291, Jan. 2019.

[52] Q. Zou, J. Li, L. Song, X. Zeng, and G. Wang, "Similarity computation strategies in the microRNA-disease network: A survey," *Briefings Functional Genomics*, vol. 15, no. 1, pp. 55–64, 2015.

[53] R. Cao, Z. Wang, Y. Wang, and J. Cheng, "SMOQ: A tool for predicting the absolute residue-specific quality of a single protein model with support vector machines," *BMC Bioinf.*, vol. 15, no. 1, p. 120, 2014.

[54] Y. Zuo, Y. Li, Y. Chen, G. Li, Z. Yan, and L. Yang, "PseKRAAC: A flexible Web server for generating pseudo K-tuple reduced amino acids composition," *Bioinformatics*, vol. 33, no. 1, pp. 122–124, Jan. 2017.

[55] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346–354, Jan. 2016.

[56] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT suite: A Web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, Mar. 2010.

[57] T. UniProt Consortium, "The universal protein resource (UniProt)," *Nucleic Acids Res.*, vol. 35, no. 1, pp. 193–197, Jan. 2007.

[58] The UniProt Consortium, "UniProt: The universal protein knowledgebase," *Nucleic Acids Res.*, vol. 45, no. 1, pp. 158–169, Jan. 2017.

[59] C. H. Wu, "The universal protein resource (UniProt): An expanding universe of protein information," *Nucleic Acids Res.*, vol. 34, no. 90001, pp. 187–191, Jan. 2006.

[60] W. Chen, P. Feng, and F. Nie, "IATP: A sequence based method for identifying anti-tubercular peptides," *Medicinal Chem.*, vol. 15, p. 1–7, Oct. 2019.

[61] Y. Shen, J. Tang, and F. Guo, "Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC," *J. Theor. Biol.*, vol. 462, pp. 230–239, Feb. 2019.

[62] Y. Shen, Y. Ding, J. Tang, Q. Zou, and F. Guo, "Critical evaluation of Web-based prediction tools for human protein subcellular localization," *Briefings Bioinf.*, Nov. 2019, doi: 10.1093/bib/bbz106.

[63] J. Zhang and B. Liu, ''A review on the recent developments of sequence-based protein feature extraction methods,'' *Current Bioinf.*, vol. 14, no. 3, pp. 190–199, Mar. 2019.

[64] K. Patil and U. Chouhan, ''Relevance of machine learning techniques and various protein features in protein fold classification: A review,'' *Current Bioinf.*, vol. 14, no. 8, pp. 688–697, Dec. 2019.

[65] C. Meng, ''Identification of proteins of tobacco mosaic virus by using a combined method of feature extraction,'' *Proteins Proteomics*, vol. 1868, no. 6, 2020, Art. no. 140406.

[66] W. Chen, P. Feng, T. Liu, and D. Jin, ''Recent advances in machine learning methods for predicting heat shock proteins,'' *Current Drug Metabolism*, vol. 20, no. 3, pp. 224–228, May 2019.

[67] H. Yang, ''Identification of Secretory Proteins in Mycobacterium tuberculosis Using Pseudo Amino Acid Composition,'' *Biomed. Res. Int.*, vol. 2016, Apr. 2016, Art. no. 5413903.

[68] H. Tang and W. H. Chen Lin, ''Identification of Immunoglobulins using Chou's pseudo amino acid composition with feature selection technique,'' *Mol. Biosyst.*, vol. 12, no. 4, pp. 1269–1275, 2016.

[69] Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, G. I. Webb, A. I. Smith, R. J. Daly, K.-C. Chou, and J. Song, ''IFeature: A Python package and Web server for features extraction and selection from protein and peptide sequences,'' *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, Jul. 2018.

[70] K. C. Chou, ''Prediction of protein cellular attributes using pseudo-amino acid composition,'' *Proteins*, vol. 43 pp. 246–255, Apr. 2001.

[71] R. Su, X. Liu, and L. Wei, ''MinE-RFE: Determine the optimal subset from RFE by minimizing the subset-accuracy–defined energy,'' *Briefings Bioinf.*, vol. 21, no. 2, pp. 687–698, Mar. 2020.

[72] R. Su, X. Liu, L. Wei, and Q. Zou, ''Deep-resp-forest: A deep forest model to predict anti-cancer drug response,'' *Methods*, vol. 166, pp. 91–102, Aug. 2019.

[73] R. Su, X. Liu, G. Xiao, and L. Wei, ''Meta-GDBP: A high-level stacked regression model to improve anticancer drug response prediction,'' *Briefings Bioinf.*, vol. 21, no. 3, pp. 996–1005, May 2020.

[74] Z. Wang, W. He, J. Tang, and F. Guo, ''Identification of highest-affinity binding sites of yeast transcription factor families,'' *J. Chem. Inf. Model.*, vol. 60, no. 3, pp. 1876–1883, Mar. 2020.

[75] Y. Ding, J. Tang, and F. Guo, ''Identification of drug-side effect association via multiple information integration with centered kernel alignment,'' *Neurocomputing*, vol. 325, pp. 211–224, Jan. 2019.

[76] X. Gao and G. Li, ''A KNN model based on manhattan distance to identify the SNARE proteins,'' *IEEE Access*, vol. 8, pp. 112922–112931, 2020.

[77] G. Li, ''Identification of SNARE proteins through a novel hybrid model,'' *IEEE Access*, vol. 8, pp. 117877–117887, 2020.

[78] L. Wei, S. Wan, J. Guo, and K. K. Wong, ''A novel hierarchical selective ensemble classifier with bioinformatics application,'' *Artif. Intell. Med.*, vol. 83, pp. 82–90, Nov. 2017.

[79] L. Wei, P. Xing, G. Shi, Z. Ji, and Q. Zou, ''Fast prediction of protein methylation sites using a sequence-based feature selection technique,'' *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 4, pp. 1264–1273, Jul. 2019.

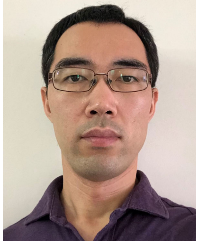
[80] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, ''Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier,'' *Artif. Intell. Med.*, vol. 83, pp. 67–74, Nov. 2017.

[81] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, ''ACPred-FL: A sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides,'' *Bioinformatics*, vol. 34, no. 23, pp. 4007–4016, 2018.

[82] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, and L. Wei, ''DUNet: A deformable network for retinal vessel segmentation,'' *Knowl.-Based Syst.*, vol. 178, pp. 149–162, Apr. 2019.

[83] Y. Ding, J. Tang, and F. Guo, ''Identification of drug-target interactions via multiple information integration,'' *Inf. Sci.*, vols. 418–419, pp. 546–560, Dec. 2017.

[84] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and H. I. Witten, ''The WEKA data mining software: An update,'' *SIGKDD Explor.*, vol. 1, pp. 10–18, May 2009.



XING GAO (Member, IEEE) was born in Yangzhou, China, in 1980. He received the B.S degree in computer science and technology from the China University of Mining and Technology, in 2002, and the M.S. and Ph.D. degrees in computer software and theory from the Harbin Institute of Technology, Harbin, China, in 2009.

From 2009 to 2013, he was an Assistant Professor with the School of Software, Xiamen University, Xiamen, China, where he has been an Associate Professor with the School of Informatics, since 2013. He has authored more than 30 articles. His research interests include bioinformatics, feature engineering, machine learning, and deep learning.

GUILIN LI was born in Harbin, China, in 1979. He received the B.S. and M.S. degrees in computer science and technology, and the Ph.D. degree in computer software and theory from the Harbin Institute of Technology, Harbin, in 2003 and 2009, respectively.

From 2009 to 2013, he was an Assistant Professor with the School of Software, Xiamen University, Xiamen, China. Since 2013, he has been an Associate Professor with the School of Informatics, Xiamen University. He has authored more than 30 articles. His research interests include bioinformatics, feature engineering, machine learning, and deep learning.