Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 8629776, 14 pages
https://doi.org/10.1155/2021/8629776

Research Article

Protein Subcellular Localization Based on Evolutionary Information and Segmented Distribution

Danyu Jin and Ping Zhu

School of Science, Jiangnan University, Wuxi 214122, China

Correspondence should be addressed to Ping Zhu; zhuping@jiangnan.edu.cn

Received 31 October 2021; Revised 6 December 2021; Accepted 7 December 2021; Published 31 December 2021

Academic Editor: Nianyin Zeng

The prediction of protein subcellular localization not only is important for the study of protein structure and function but also can facilitate the design and development of new drugs. In recent years, feature extraction methods based on protein evolution information have attracted much attention and made good progress. Based on the protein position-specific score matrix (PSSM) obtained by PSI-BLAST, PSSM-GSD method is proposed according to the data distribution characteristics. In order to reflect the protein sequence information as much as possible, AAO method, PSSM-AAO method, and PSSM-GSD method are fused together. Then, conditional entropy-based classifier chain algorithm and support vector machine are used to locate multilabel proteins. Finally, we test Gpos-mPLoc and Gneg-mPLoc datasets, considering the severe imbalance of data, and select SMOTE algorithm to expand a few sample; the experiment shows that the AAO + PSSM* method in the paper achieved 83.1% and 86.8% overall accuracy, respectively. After experimental comparison of different methods, AAO + PSSM* has good performance and can effectively predict protein subcellular location.

1. Introduction

Cells are the basic unit of life, and various organelles in organisms are also called subcells, which are further subdivided into cells, including mitochondria, cell membrane, and nucleus. Many of the life activities of living organisms are performed by proteins, and thousands of proteins can only function at specific locations in living organisms. With the advent of the postgene era, a variety of biological information has exploded, and many new protein sequences have been excavated. However, the traditional experimental localization prediction methods overconsume the experimental cost and time [1], so it is urgent to build an efficient and accurate computational model to predict the subcellular location of proteins. For the newly discovered unknown protein, selecting suitable models with good performance to predict its subcellular location can help us further understand the life activities of the protein in the organism. Therefore, protein subcellular localization is of certain significance to the study of protein function and structure and also helps us to recognize new proteins and better understand complex biological functions.

In recent years, more and more models have been proposed to predict protein subcellular localization, and the accuracy and calculation speed have been improved continuously. Therefore, protein subcellular localization prediction has become a major focus in biological information research. The prediction model of protein subcellular localization mainly consists of two parts: one is to select a reasonable method to extract protein information features to a great extent; the other is to build a classification prediction model to obtain better results.

At present, feature extraction methods mainly include the following: methods based on amino acid sequence information, methods based on protein evolution information, methods based on gene ontology, methods based on amino acid physical and chemical properties. The traditional method of amino acid composition [2–4] includes the amino acid frequency in the protein sequence. Although this method is simple and easy to understand, the arrangement

of amino acids is not considered. Subsequently, considering the sequence of amino acids, Chou et al. propose the pseudoamino Acid Composition (PseAAC) method [5] by adding the physical and chemical properties of amino acids. Since then, this method has been widely used in the prediction of protein subcellular localization [6, 7]. Gene ontology involves a vocabulary package of genes and gene products that integrates cell components, molecular functions, and biological processes. Many scholars make use of GO information [8, 9] and make continuous improvement on this basis, achieving good results. In recent studies, the field has focused on feature extraction methods that use protein evolution information to extract feature information and greatly improve classification accuracy. With continuous exploration, methods such as PsePSSM [10], PSSM-S [11], DipCPSSM [12], and PSSM-SAA [13] are proposed and apply to the prediction of protein subcellular localization, achieving good results.

For the classification prediction model of protein subcellular localization, the traditional algorithms mainly include support vector machine (SVM) [14, 15], K-nearest neighbor algorithm (KNN) [16], random forest [9], deep learning [17], and integrated learning [18]. K-nearest neighbors is a simple and mature classification method, whose principle is to select the category with the most frequent occurrence among the K-nearest neighbors as the judgment category, but its classification results depend very much on dataset balance and k-value selection. Random forest is an algorithm that synthesizes multiple decision trees based on the idea of integration and has high flexibility and robustness. Although deep learning, integrated learning, and other classification algorithms have made some progress in protein subcellular localization prediction, they still have disadvantages such as high computational complexity. However, the traditional SVM has better performance and generalization in solving the nonlinear classification problems, so it has been widely used in the prediction of protein subcellular localization. Many researchers have improved the basis of SVM to propose classifiers with higher accuracy and better performance. In fact, most proteins are located at one subcellular site, but some proteins exist at two or more subcellular sites simultaneously. Multisite protein research is also a major focus of protein subcellular localization. Good multilabel classification algorithm can fully explore the relationship between tags and improve classification accuracy. Common multilabel classification algorithms include binary relevance [19], classifier chains [20], and MLKNN [21].

Because the classification effect of some classifiers depends very much on the balance of datasets, the classification performance will be greatly affected if there is a serious imbalance of datasets. The most common method to solve this problem is to directly reduce some majority samples or add minority samples on the basis of the original data, but these methods are easy to lose data information and make minority samples collinear. SMOTE (synthetic minority oversampling technique) algorithm [22] is the oversampling algorithm proposed by Chawla et al., which improves the above problems to some extent and has a good performance in solving the problem of unbalanced data classification. But it only applies to single label classification problems. SMOTE is an improved method based on random sampling, using the K-nearest neighbor method to artificially create minority samples, resulting in a rough balance of samples.

2. Materials and Methods

2.1. AAO (Consensus Sequence-Based Occurrence). AAO method [23] starts with the PSSM matrix of protein and extracts the information of amino acid evolution in the process of protein evolution.

If two proteins share a common evolutionary ancestor, they are homologous. Homologous proteins have similar amino acid arrangements and, in general, similar functions. Considering the evolutionary information of proteins, this study selects the site-specific score matrix PSSM to further extract protein-related information. The site-specific score matrix is to extract as much protein sequence information as possible by taking the evolutionary information of protein sequence into consideration on the basis of protein sequence. For each protein sequence selected in this study, PSI-BLAST [24] is used to search and compare related homologous sequences in SWISS-PROT database, and then homologous information was numerized by PSSM matrix. The PSSM matrix obtained by a protein sequence $P = p_1 p_2 \cdots p_L$ of length L is shown as

$$\text{PSSM} = \begin{bmatrix} U^0_{1\longrightarrow 1} & U^0_{1\longrightarrow 2} & \cdots & U^0_{1\longrightarrow j} & \cdots & U^0_{1\longrightarrow 20} \\ U^0_{2\longrightarrow 1} & U^0_{2\longrightarrow 2} & \cdots & U^0_{2\longrightarrow j} & \cdots & U^0_{2\longrightarrow 20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ U^0_{i\longrightarrow 1} & U^0_{i\longrightarrow 2} & \cdots & U^0_{i\longrightarrow j} & \cdots & U^0_{i\longrightarrow 20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ U^0_{L\longrightarrow 1} & U^0_{L\longrightarrow 2} & \cdots & U^0_{L\longrightarrow j} & \cdots & U^0_{L\longrightarrow 20} \end{bmatrix}. \quad (1)$$

PSSM is a matrix of $L \times 20$ dimensions, where L is the length of the protein sequence, and the numbers 1~20 represent the 20 amino acids that make up the protein. $U^0_{i\longrightarrow j}$ represents the probability score of the $i\,(i = 1, 2, \cdots, L)$ amino acid in the protein sequence being encoded as the $j\,(j = 1, 2, \ldots, 20)$ amino acid in the process of evolution. In other words, each row of the PSSM matrix reflects the probability that the current amino acid residue will be replaced by twenty amino acids.

Consensus sequence is an amino acid sequence consisting of the most commonly occurring residues at each position in a set of protein homologous sequences. AAO method is to replace the amino acid at each position in the protein sequence with the amino acid with the highest probability score, that is, the maximum value of each row in the PSSM matrix, to obtain a new common sequence $P_{\text{new}} = o_1 o_2 \cdots o_L$ and then calculate the occurrence frequency $f_i\,(i = 1, 2, \cdots, 20)$ of 20 amino acids. So the amino acid composition of a protein can be represented by a 20-dimensional vector as

$$W_{AAO} = [f_1, f_2, \ldots, f_{20}]^T. \quad (2)$$

2.2. PSSM-AAO (Simi-Occurrence). PSSM-AAO method [23] is a feature of protein evolution extracted from PSSM matrix. Each column of the PSSM matrix is summed up to represent the evolution of the j-th $(j = 1, 2, \ldots, 20)$ amino acid in the whole protein sequence.

First, the elements in the PSSM matrix are normalized by the following method as

$$U_{i\longrightarrow j} = \frac{U^0_{i\longrightarrow j} - (1/20)\sum_{k=1}^{20} U^0_{i\longrightarrow k}}{\sqrt{(1/20)\sum_{l=1}^{20}\left(U^0_{i\longrightarrow l} - (1/20)\sum_{k=1}^{20} U^0_{i\longrightarrow k}\right)^2}}. \quad (3)$$

The standardized score matrix is obtained, denoted as $\mathrm{PSSM}_{\mathrm{std}}$, as follows:

$$\mathrm{PSSM}_{\mathrm{std}} = \begin{bmatrix} U_{1\longrightarrow 1} & U_{1\longrightarrow 2} & \cdots & U_{1\longrightarrow j} & \cdots & U_{1\longrightarrow 20} \\ U_{2\longrightarrow 1} & U_{2\longrightarrow 2} & \cdots & U_{2\longrightarrow j} & \cdots & U_{2\longrightarrow 20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ U_{i\longrightarrow 1} & U_{i\longrightarrow 2} & \cdots & U_{i\longrightarrow j} & \cdots & U_{i\longrightarrow 20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ U_{L\longrightarrow 1} & U_{L\longrightarrow 2} & \cdots & U_{L\longrightarrow j} & \cdots & U_{L\longrightarrow 20} \end{bmatrix}. \quad (4)$$

Based on the above matrix, the protein sequence is further expressed as equations (5) and (6), namely, the PSSM-AAO feature sequence:

$$W_{\mathrm{PSSM-AAO}} = \left(\overline{U_1}, \overline{U_2}, \overline{U_3}, \ldots, \overline{U_{20}}\right)^T, \quad (5)$$

$$\overline{U_j} = \frac{1}{L}\sum_{i=1}^{L} U_{i\longrightarrow j}, \quad j = 1, 2, \ldots, 20. \quad (6)$$

2.3. PSSM-GSD (Grouped Segmented Distribution)

2.3.1. Theoretical Basis. The central tendency and dispersion degree of data are two basic characteristics of data distribution. Most data show the law of fluctuation around a certain center within a certain range, which is the central tendency of data. The main statistics include mean, median, and mode. However, the distribution of data outside the range cannot be fully demonstrated by these statistics, so statistics such as range, quartile difference, and variance reflect the degree of dispersion of data.

When a set of data is arranged in ascending order, the value corresponding to a certain percentile is called the percentile of this percentile. For the p-th $(p = 1, 2, \ldots, 20)$ percentile, strictly define the p-th percentile as a number such that at least P% of the data items are less than or equal to this value, and at least $(100 - p)$% of the data items are greater than or equal to this value. In particular, the median is the 50th percentile, the lower quartile is the 25th percentile, and the upper quartile is the 75th percentile. These three numbers are the quartiles, which comprehensively reflect the central tendency and dispersion degree of a group of data.

2.3.2. PSSM-GSD. The length of each protein sequence is different, but the input classifier has the same requirement for the number of features, and the PSSM matrix contains a lot of information. The above 20-dimensional vector extracted from the protein PSSM matrix does not fully extract the protein sequence information and its amino acid distribution. Therefore, based on the PSSM matrix, a new feature algorithm is proposed to reflect the piecewise distribution of amino acid evolution information along the protein sequence to add more local information. Improving the PSSM-SD method [11] proposed by Dehzangi et al., a new feature extraction algorithm PSSM-GSD method is proposed considering the high and low replacement scores of protein sequences.

Since the PSSM matrix reflects the evolutionary information of proteins, the larger the element value in the matrix is, the more likely the homologous proteins at this position are to be replaced by this amino acid during the evolution. The elements in the PSSM matrix were divided into 3 equal fractions according to the score of amino acid replacement, and the low-probability replacement group 1, medium-probability replacement group 2, and high-probability replacement group 3 are obtained (see Table 1).

$m = \max(U_{i\longrightarrow j})$ is the maximum value of all elements of the PSSM matrix, $n = \min(U_{i\longrightarrow j})$ is the minimum value of all elements of the PSSM matrix, $l = ((m - n)/3)$ indicates the numeric length of each group, and $i = 1, 2, \ldots, L$; $j = 1, 2, \ldots, 20$.

Therefore, the PSSM matrix of each protein can be transformed into a PSSM grouping matrix denoted by $\mathrm{PSSM}_{\mathrm{group}}$ as

$$\mathrm{PSSM}_{\mathrm{group}} = \begin{bmatrix} G_{1\longrightarrow 1} & G_{1\longrightarrow 2} & \cdots & G_{1\longrightarrow j} & \cdots & G_{1\longrightarrow 20} \\ G_{2\longrightarrow 1} & G_{2\longrightarrow 2} & \cdots & G_{2\longrightarrow j} & \cdots & G_{2\longrightarrow 20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ G_{i\longrightarrow 1} & G_{i\longrightarrow 2} & \cdots & G_{i\longrightarrow j} & \cdots & G_{i\longrightarrow 20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ G_{L\longrightarrow 1} & G_{L\longrightarrow 2} & \cdots & G_{L\longrightarrow j} & \cdots & G_{L\longrightarrow 20} \end{bmatrix}. \quad (7)$$

The $G_{i\longrightarrow j}$ $(i = 1, 2, \ldots, L; j = 1, 2, \ldots, 20)$ values are 1, 2, and 3, indicating the category to which the element belongs.

In order to reflect the centralization trend and dispersion degree of data as much as possible, the distribution of evolutionary information of 20 amino acids in protein sequence is considered, and the lower quartile, median quartile, upper quartile, 1st percentile, and 100th percentile of statistics are introduced. The distribution along the protein chain is described by five chain lengths (percentages), which contain the position coordinates of the first (1%), 25%, 50%, 75%, and 100% of a group [25]. To facilitate understanding, the extraction method of split distribution will be described in detail below (see Figure 1).

First calculate the position I_j^{k1} where the first element of group k appears. For column j $(j = 1, 2, \ldots, 20)$ of $\mathrm{PSSM}_{\mathrm{group}}$, the total numbers T_{1j}, T_{2j} and T_{3j} of groups 1, 2, and 3 are calculated respectively, and the calculation formula is as

$$T_{kj} = \sum_{i=1}^{L} \left|G_{i\longrightarrow j} = k\right|, \quad (8)$$

Table 1: PSSM matrix grouping.

Group	Grouping condition	Mark
Group 1	$n \le U_{i\longrightarrow j} \le n + l \times 1$	1
Group 2	$n + l \times 1 < U_{i\longrightarrow j} \le n + l \times 2$	2
Group 3	$n + l \times 2 < U_{i\longrightarrow j} \le m$	3

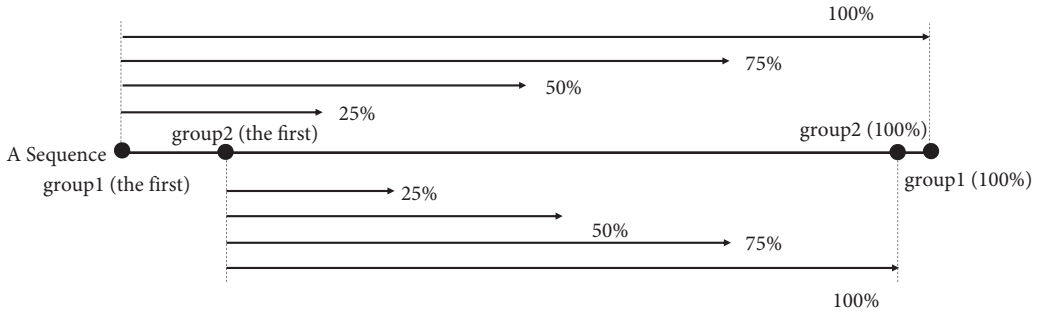

Figure 1: Schematic diagram of segmentation distribution feature extraction of sequence: the first 1%, 25%, 50%, 75%, and 100% distributions of two groups in the same sequence.

$|G_{i\longrightarrow j} = k|\,(k = 1, 2, 3)$ refers to if $G_{i\longrightarrow j}$ belongs to the group k, it is 1or otherwise 0.

Then, for the j-th ($j = 1, 2, \ldots, 20$) column of $\mathrm{PSSM}_{\mathrm{group}}$, from the first row of $\mathrm{PSSM}_{\mathrm{group}}$ matrix to the I_j^k row, the numbers S_{1j}, S_{2j}, and S_{3j} of groups 1, 2, and 3 are calculated respectively. The calculation formula is

$$S_{kj} = \sum_{i=1}^{I_j^k} \left| G_{i\longrightarrow j} = k \right|. \tag{9}$$

When (S_{kj}/T_{kj}) reaches 25%, 50%, 75%, and 100%, the corresponding I_j^k is obtained; then, the j-th column of $\mathrm{PSSM}_{\mathrm{group}}$ can be represented as the 15-dimensional feature vector $\mathrm{GSD}_j = (I_j^{11}, \ldots, I_j^{15}, I_j^{21}, \ldots, I_j^{25}, I_j^{31}, \ldots, I_j^{35})$. Further, the protein sequence can be represented as the 300-dimensional feature vector:

$$W_{\mathrm{PSSM-GSD}} = \left(\mathrm{GSD}_1, \mathrm{GSD}_2, \ldots, \mathrm{GSD}_{20}\right)^T. \tag{10}$$

Due to the different lengths of protein sequences, the values of feature vector $W_{\mathrm{PSSM-GSD}}$ vary greatly, so Z_{score} standardization is performed on it to facilitate subsequent studies.

2.4. Feature Fusion. In order to consider as much protein sequence information as possible, based on the idea of feature fusion, this study proposes a new feature extraction algorithm for protein subcellular localization prediction. In this algorithm, the selected feature extraction algorithms AAO and PSSM-AAO are combined with the new method PSSM-GSD, which not only consider the evolutionary information of protein sequence, but also indirectly extract the evolutionary arrangement information of amino acids, greatly enriching the protein information. The protein sequence information after fusion can be expressed as

$$W_P = W_{\mathrm{AAO}} + W_{\mathrm{PSSM-AAO}} + W_{\mathrm{PSSM-GSD}}. \tag{11}$$

Since the feature AAO, PSSM-AAO, and PSSM-GSD of protein have all been standardized, the protein sequence can be transformed into $340(20 + 20 + 300)$ dimensional feature vector by directly combining the three feature values.

3. Classifier Chains

Many proteins function at only one subcellular site, but in practice some proteins exist at two or more subcellular sites simultaneously. It is also very important to predict the proteins of these multisubcellular sites, and there is a certain relationship between the location tags of these proteins. How to construct a fast and accurate multilabel classification algorithm is the key to solve the localization of multilocation proteins.

For multilabel classification problems, the most simple and common method is the binary relevance method, which transforms multilabel problems into multiple binary classification problems, and each binary classifier corresponds to a label to be predicted. BR method is easy to understand and operate, but if there is a correlation between labels, it will have a great impact on its prediction ability. Based on BR algorithm, classifier chains (CC) is an algorithm which further considers the relationship between tags. Its core idea is to arrange tags in a chain in a certain order, and the input of the last binary classifier in the chain depends on the input and output of the previous classifier. The principle of CC algorithm determines that the algorithm is very dependent on the ordering of tags on the chain. If the prediction result of initial tags on the chain is not accurate, the error information will be transmitted along the chain. Therefore, label sorting is a very important part of the algorithm.

Information entropy is a concept proposed by the famous scientist Shannon to measure the uncertainty or disorder of random events. For the information source X, its information entropy is

$$H(X) = E\left(\log\frac{1}{P(a_i)}\right) = -\sum_{i=1}^{q} P(a_i)\log\ P(a_i). \quad (12)$$

When the information entropy is smaller, the uncertainty or disorder of the random event is smaller. For deterministic events, the information entropy is 0. If a binary classification problem is a deterministic event, that is, all samples belong to the same category, the prediction results of other samples will hardly be wrong. Therefore, for labels with lower information entropy, there is less uncertainty in prediction, so the labels can be sorted according to the information entropy of them from low to high.

However, the information entropy does not take into account the influence between the front and rear labels on the chain, so the conditional entropy is introduced to define the average uncertainty of the output of the latter symbol X_2 when the former symbol X_1 is defined, and the calculation formula is

$$\begin{aligned} H(X_2 \mid X_1) &= \sum_{i=1}^{q} P(a_i) H(X_2 \mid X_1 = a_i) \\ &= -\sum_{i=1}^{q}\sum_{j=1}^{q} P(a_i) P(a_j \mid a_i)\log\ P(a_j \mid a_i) \\ &= -\sum_{i=1}^{q}\sum_{j=1}^{q} P(a_i a_j)\log\ P(a_j \mid a_i). \end{aligned} \quad (13)$$

For label sorting, this paper firstly calculates the information entropy of all labels and takes the label with the lowest information entropy as the first label on the CC chain. Then, the conditional entropy of the remaining labels on the basis of the on-chain labels is calculated, and the label with the lowest conditional entropy is added to the chain until all labels are sorted. Finally, CC algorithm based on conditional entropy sorting is combined with support vector machine classifier to perform protein subcellular localization.

4. Experimental Results and Analysis

4.1. Experimental Data and Pretreatment. In this study, protein data are derived from the Cell-PLoc 2.0 database [26] (http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/), which is a set of web servers for predicting the subcellular localization of proteins from different organisms, and is an improvement of Chou and Shen on Cell-Ploc 1.0.

In order to better test the comprehensive performance and generalization ability of the proposed method, the Gram-positive protein Gpos-mPLoc dataset and Gram-negative protein Gneg-mPLoc dataset are selected for experiments. The Gpos-mPLoc dataset (see Table 2) contains 519 different proteins with 4 subcellular site labels, 515 proteins in one subcellular site, and 4 proteins in two subcellular sites. Since a protein located in cytoplasm contains amino acid U, the protein sequence is deleted. The Gneg-mPLoc dataset (see Table 3) contains 1392 different proteins with 8 subcellular site labels, 1328 proteins in one subcellular site, and 64 proteins in two subcellular sites.

Table 2: Subcellular location distribution of Gpos-mPLoc protein.

Subcellular site marker	Protein subcellular class	Number of proteins
1	Cell membrane	174
2	Cytoplasm	207
3	Cell wall	18
4	Extracell	123

Table 3: Subcellular location distribution of Gneg-mPLoc protein.

Subcellular site marker	Protein subcellular class	Number of proteins
1	Cell inner membrane	557
2	Cell outer membrane	124
3	Cytoplasm	410
4	Extracell	133
5	Fimbrium	32
6	Flagellum	12
7	Nucleoid	8
8	Periplasm	180

As can be seen from Table 2, the data of Gram-positive proteins selected in this paper have serious data imbalance. The protein samples located in the cell wall only contain 18 proteins, which is about 1/10 of other types of subcellular proteins. As can be seen from Table 3, the data distribution of Gram-negative proteins is also seriously unbalanced, with only 0.5%, 0.8%, and 2.2% of the proteins located in the fimbrium, flagellum, and nucleoid. The pie chart of Gpos-mPLoc data distribution in Figure 2(a) also shows that the number of proteins located in the cell wall only accounts for 3% of the total amount, while the distribution of other three types of data is relatively balanced. As a result, the classifier tends to predict a few samples as a majority of samples, resulting in extremely low classification accuracy and ultimately affecting the prediction performance of the classifier.

In order to minimize the classification errors caused by unbalanced data distribution, Gpos-mPLoc is taken as an example to show some data preprocessing results. Select SMOTE algorithm to generate new sample of Gram-positive protein in cell membrane, cell wall, and extracell to make the data relatively balanced and finally use the new and original data as experimental dataset for experimental analysis. In this paper, we select the data after SMOTE oversampling and compare with the original data distribution, finally drawing Figures 2 and 3. In the figure we can see that the protein data after SMOTE is evenly distributed in the 4 subcellular sites with very similar percentage, which greatly alleviates the imbalance of the original data.

4.2. Evaluation Index. In the prediction of protein subcellular localization, there are many indicators that can be used to evaluate the performance of the model. In this study, four commonly used test indicators are selected: sensitivity (also known as recall rate), specificity, Matthews correlation coefficient (MCC), and overall accuracy, calculated as

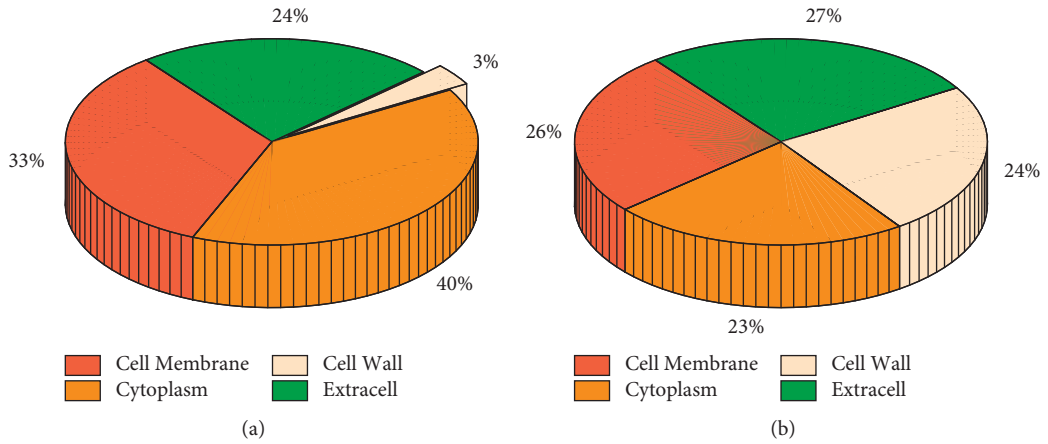

Figure 2: Pie chart of protein subcellular location distribution. (a) Pie chart of subcellular position distribution in Gpos-mPLoc. (b) Pie chart of subcellular position distribution in the dataset after SMOTE.

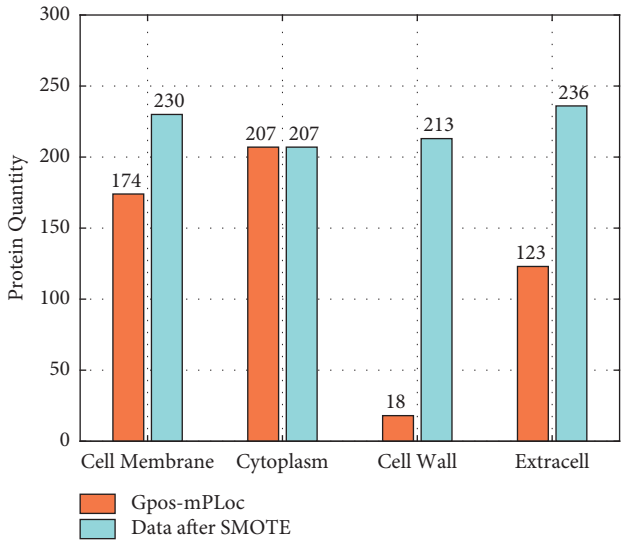

Figure 3: Histogram of protein subcellular location distribution. The comparison of distribution of Gram-positive protein dataset Gpos-mPLoc and data after SMOTE.

$$\mathrm{SN}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}, \tag{14}$$

$$\mathrm{SP}_i = \frac{\mathrm{TN}_i}{\mathrm{TN}_i + \mathrm{FP}_i}, \tag{15}$$

$$\mathrm{MCC}_i = \frac{(\mathrm{TP}_i \times \mathrm{TN}_i) - (\mathrm{FN}_i \times \mathrm{FP}_i)}{\sqrt{(\mathrm{TP}_i + \mathrm{FN}_i) \times (\mathrm{TN}_i + \mathrm{FP}_i) \times (\mathrm{TP}_i + \mathrm{FP}_i) \times (\mathrm{TN}_i + \mathrm{FN}_i)}}, \tag{16}$$

$$OA = \frac{\sum_{i=1}^{4} \mathrm{TP}_i}{N}, \tag{17}$$

where for the class i sample, the positive sample is the class i sample, and the negative sample is the other class samples except the class i sample. TP_i refers to the number of samples judged as positive; in fact, it is the positive; TN_i refers to the number of samples judged as negative; in fact, it is the negative; FN_i refers to the number of samples judged as negative; in fact, it is the positive; FP_i refers to the number of samples judged as positive; in fact, it is the negative. N is the

number of all protein samples. The protein sample data selected in this study have four subcellular classes, so $i = 1, 2, 3, 4$.

4.3. Results and Analysis. In order to evaluate the performance of the proposed algorithm and verify the necessity and advantages of multifeature fusion for protein subcellular localization prediction, the corresponding protein feature information is extracted by AAO method, PSSM-AAO method, PSSM-GSD method, and three feature fusion methods on Gpos-mPLoc dataset. Support vector machine (SVM) is used for subcellular localization of protein datasets. LIBSVM [11, 27] is used to construct the prediction model classifier. The linear kernel function is used for the classifier, and the default settings are used for other related parameters. In this paper, the retention method (n-fold cross-validation) is used to verify the prediction performance of model classification. In each round of validation, one protein sample in the dataset is selected as the test set, and the rest protein samples are used as the training set to train the classifier, and the verification results are unique. As the retention method selects as much training data as possible in each iteration, the results obtained by this method are the closest to the expected results of the entire training dataset, and the retention method is also recognized as one of the most objective and reliable test methods.

Due to the randomness of SMOTE algorithm, the mean of 10 experimental results is selected as the final result in this paper, the method marked with * has the test after oversampling on Gpos-mPLoc dataset, the method without * uses the original Gpos-mPLoc dataset, the classification algorithm is SVM, and the feature fusion algorithm proposed in this paper is AAO + PSSM* (see Table 4).

As can be seen from Table 4, the overall accuracy of AAO* method, PSSM-AAO* method, PSSM-GSD* method, and multifeature fusion method on Gram-positive protein dataset is 44.5%, 71%, 73.5%, and 83.1%, respectively. In this paper, the feature extraction method AAO + PSSM* has the best overall performance, not only the best accuracy, but also high sensitivity and specificity.

In order to compare the performance of various feature extraction methods more directly, the results of the above indicators are displayed by the bar chart. As can be seen from Figures 4(a) and 5, the classification effect of AAO* method is obviously weaker than other methods, and the overall accuracy is the lowest, and some evaluation indexes are even lower than 0.6, which is still far behind existing protein characterization models. PSSM-AAO* and PSSM-GSD* perform well. The SN, SP, and MCC indexes of PSSM-AAO* are higher in cytoplasm, and the SN, SP, and MCC indexes of PSSM-GSD* in cell wall are as high as 0.9.

It can be seen from Figure 5 that after combining the three feature extraction algorithms, the new method AAO + PSSM* occupies an absolute advantage in almost all aspects and is the best in sensitivity except for a slight disadvantage in cell membrane and cytoplasm. In specificity, almost all feature extraction methods perform well, and AAO + PSSM* is as high as 0.9 at all four locations, indicating that the specificity of the proposed method is good. In the Matthews correlation coefficient, the performance on the cell membrane is poor, but the other positions are better than other methods. It can be seen that the single feature extraction method has a better result for predicting proteins in some subcellular locations, and after the combination of multiple features, the multiple features of the protein sequence are fully extracted, and the advantages of several feature extraction algorithms are greatly brought into play. The SN, SP, MCC, and OA indexes of the new multifeature fusion method are significantly improved compared with the AAO* method and are significantly improved compared with PSSM-AAO* and PSSM-GSD* methods.

Due to the defect of severe imbalance in the selected dataset, the overall accuracy of the fusion feature extraction method is only 66.2% without combined SMOTE algorithm, but increased by 16.9% after adding SMOTE algorithm to balance the dataset, getting a good classification result. In Figure 6, it can be seen intuitively that the protein dataset has a great improvement in the experimental classification after SMOTE oversampling. Although all indicators decrease after SMOTE data in the cytoplasm, the sensitivity, specificity, and Matthews correlation coefficient in cell membrane, cell wall, and extracellular are increased to varying degrees. The sensitivity in cell wall and extracellular cell is increased by 0.642 and 0.21, and the Matthews correlation coefficient in cell membrane, cell wall, and extracellular cell is increased by 0.172, 0.593, and 0.307, respectively. Therefore, after using SMOTE algorithm to deal with Gpos-mPLoc dataset, the data imbalance is greatly relieved and the classification performance is improved.

In order to better illustrate the advantages of the support vector machine classification method selected in this paper, the nearest neighbor classification method and random forest are selected in the experiment to compare the results (see Table 5). The fusion algorithm AAO + PSSM proposed in this paper is used for feature extraction methods, and SMOTE oversampling is used for all data.

As can be seen from Table 5 and Figure 7, the result of random forest is not satisfactory, the overall accuracy is only 68.1%, while the nearest neighbor method is better than random forest, and the classification accuracy is 73.7%. Random forest has good generalization ability, but it may be due to the small amount of data in this paper or the similar decision tree generated, so the classification result is not ideal. The support vector machine has the highest accuracy of 83.1% among the three classification methods, which has been greatly improved. The three extracellular indicators of SVM are optimal, and the sensitivity of SVM in cell membrane, cytoplasm, and extracellular cells is the best, and the sensitivity of SVM in cell wall is as high as 0.975. The specificity indexes of the four subcellular locations were all above 0.95, and the SVM method shows good specificity. In terms of Matthews correlation coefficient, SVM is the highest among the four subcellular locations, especially in the cytoplasm and extracell. Therefore, the support vector machine classification algorithm is very good, and its overall classification accuracy is high and has a low rate of missed classification and misclassification rate.

Table 4: Comparison of single feature and fusion feature on Gpos-mPLoc dataset.

Feature extraction method	Subcellular site				OA
	Cell membrane	Cytoplasm	Cell wall	Extracell	
(SN)					
AAO*	0.700	0.432	0.101	0.499	0.445
PSSM-AAO*	0.622	0.827	0.743	0.716	0.710
PSSM-GSD*	0.806	0.692	0.962	0.840	0.735
AAO + PSSM	0.718	0.976	0.333	0.707	0.662
AAO + PSSM*	**0.744**	**0.872**	**0.975**	**0.917**	**0.831**
(SP)					
AAO*	0.999	0.940	0.966	1.000	—
PSSM-AAO*	0.989	0.979	0.957	0.998	—
PSSM-GSD*	0.956	0.925	0.976	0.965	—
AAO + PSSM	0.843	0.994	0.984	0.896	—
AAO + PSSM*	**0.958**	**0.968**	**0.986**	**0.982**	—
(MCC)					
AAO*	0.769	0.361	0.070	0.523	—
PSSM-AAO*	0.725	0.833	0.735	0.781	—
PSSM-GSD*	0.774	0.628	0.927	0.810	—
AAO + PSSM	0.558	0.972	0.358	0.595	—
AAO + PSSM*	**0.730**	**0.848**	**0.951**	**0.902**	—

The bold values are the results of the proposed method in the paper on the Gpos-mPLoc dataset for clearer comparison.

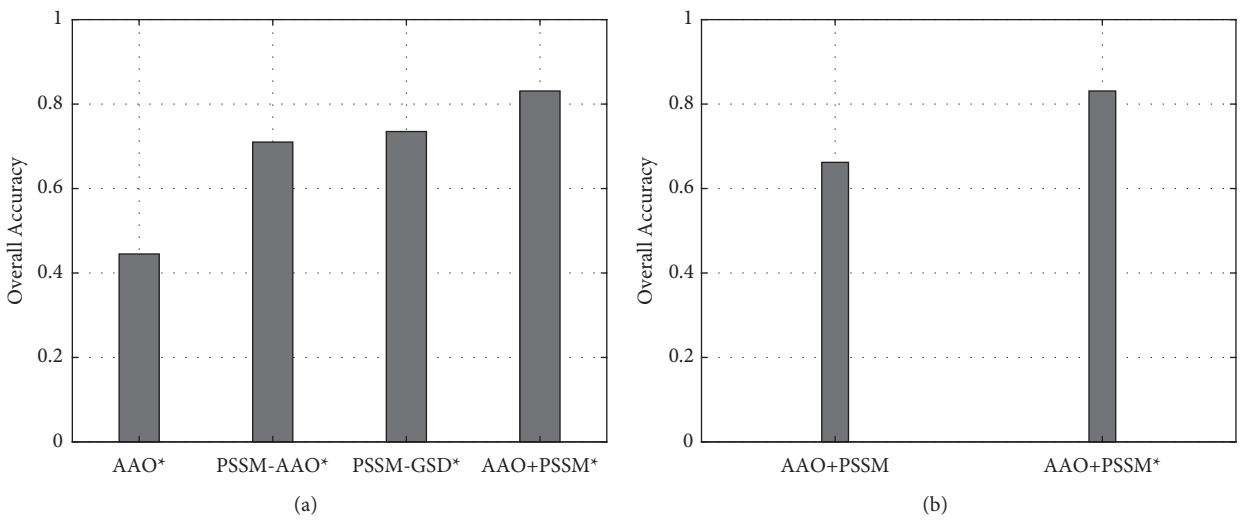

Figure 4: Comparison of overall accuracy of different methods. (a) Classification accuracy of all kinds of feature extraction algorithms. (b) Classification accuracy of AAO + PSSM on the original data and data after SMOTE.

In order to evaluate the proposed method more objectively, two existing methods are selected and tested on Gram-positive protein dataset respectively. One of them is that Yu and Zhang in 2021 combining amino acid composition and PSSM matrix to get AAO and PSSM-AAO feature fusion algorithm, using MLSMOTE balance dataset and finally using MLKNN classification prediction, putting forward AAO + PAAO* method [23]. The other is the PSORTb3.0 predictor [28] proposed by Yu et al. in 2010 (see Table 6).

As can be seen from Table 6 and Figure 8, compared with the other two prediction methods, the prediction results in this paper are better in terms of overall accuracy and test indexes of each subcellular location, and both of them have been greatly improved. In terms of overall accuracy, the proposed method is 3% higher than PSORTb3.0 classifier, which performs better among the existing algorithms. In addition, compared with AAO + PAAO* and PSORTb3.0, the sensitivity, specificity, and Matthews correlation coefficient of the proposed method in cell wall location are higher and more stable. Compared with the other two methods, the sensitivity and Matthews correlation coefficient of the proposed method at the four subcellular sites are the highest, and the specificity of the proposed method is the best except for the cell wall site, which is slightly inferior to PSORTb3.0. Therefore, the proposed method has excellent performance and good classification effect.

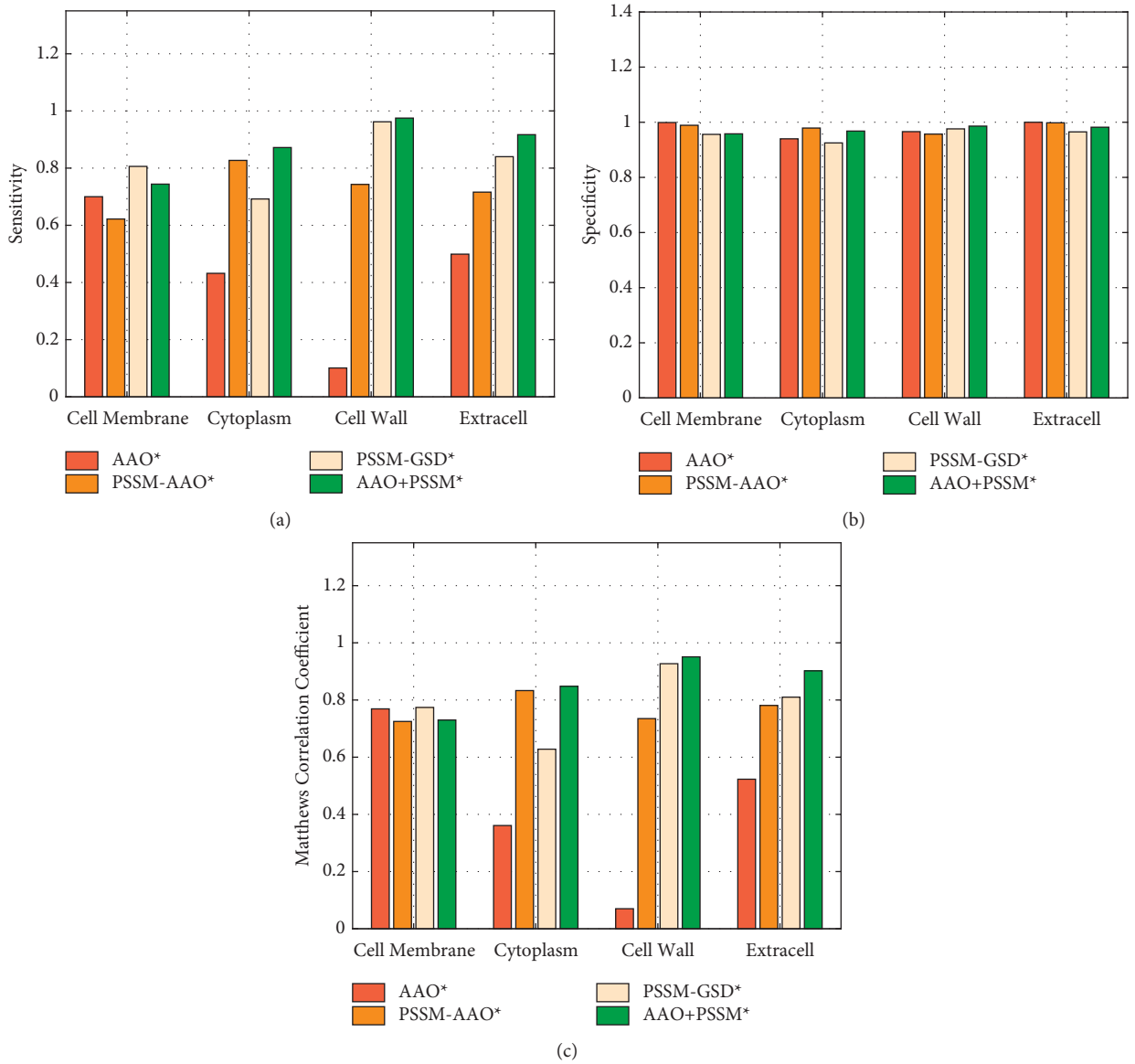

Figure 5: Comparison of evaluation index in various feature extraction algorithms. (a) Sensitivity, (b) specificity, and (c) Matthews correlation coefficient of AAO*, PSSM-AAO*, PSSM-GSD*, and AAO + PSSM*.

To measure the generalization ability of the proposed method, an experiment is performed on Gram-negative dataset (see Table 7). Gneg-mPLoc has larger data volume and more multilabel protein samples. Also we perform oversampling on the Gneg-mPLoc dataset and select the mean of ten SMOTE results as the final result in order to eliminate randomness. Cim refers to the inner membrane and com refers to the outer membrane.

The proposed method achieves an overall accuracy of 86.8% on the Gneg-mPLoc dataset, showing good performance. It is particularly good in cell outer membrane, fimbrium, flagellum, and nucleoid and each index is as high as 0.9. The sensitivity and the Matthews correlation coefficient are all above 0.8 at 8 subcellular sites. Therefore, the subcellular localization method proposed in this paper has good generalization ability and can achieve accurate localization of proteins on different datasets.

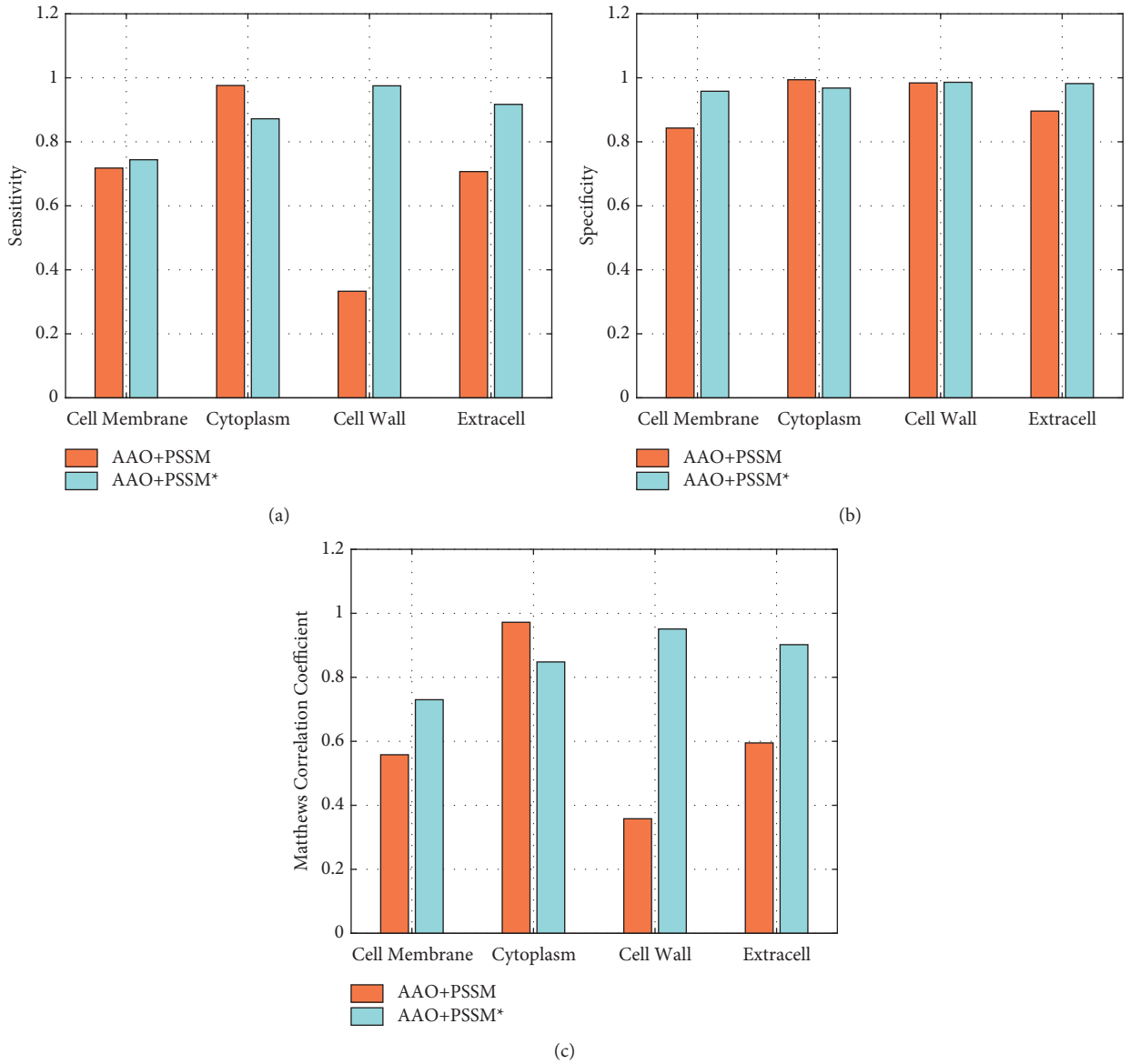

Figure 6: Comparison of evaluation index on various datasets. (a) Sensitivity, (b) specificity, and (c) Matthews correlation coefficient on raw data and data after SMOTE algorithm.

Table 5: Comparison of different classification methods on Gpos-mPLoc dataset.

Classification method	Subcellular site				OA
	Cell membrane	Cytoplasm	Cell wall	Extracell	
(SN)					
RF	0.631	0.554	0.910	0.623	0.681
KNN	0.357	0.658	0.992	0.825	0.737
SVM	**0.744**	**0.872**	**0.975**	**0.917**	**0.831**
(SP)					
RF	0.984	0.970	0.998	0.973	—
KNN	0.969	0.904	0.947	0.908	—
SVM	**0.958**	**0.968**	**0.986**	**0.982**	—
(MCC)					
RF	0.709	0.617	0.938	0.672	—
KNN	0.430	0.590	0.893	0.729	—
SVM	**0.730**	**0.848**	**0.951**	**0.902**	—

The bold values are the results of the proposed method in paper on the Gpos-mPLoc dataset for clearer comparison.

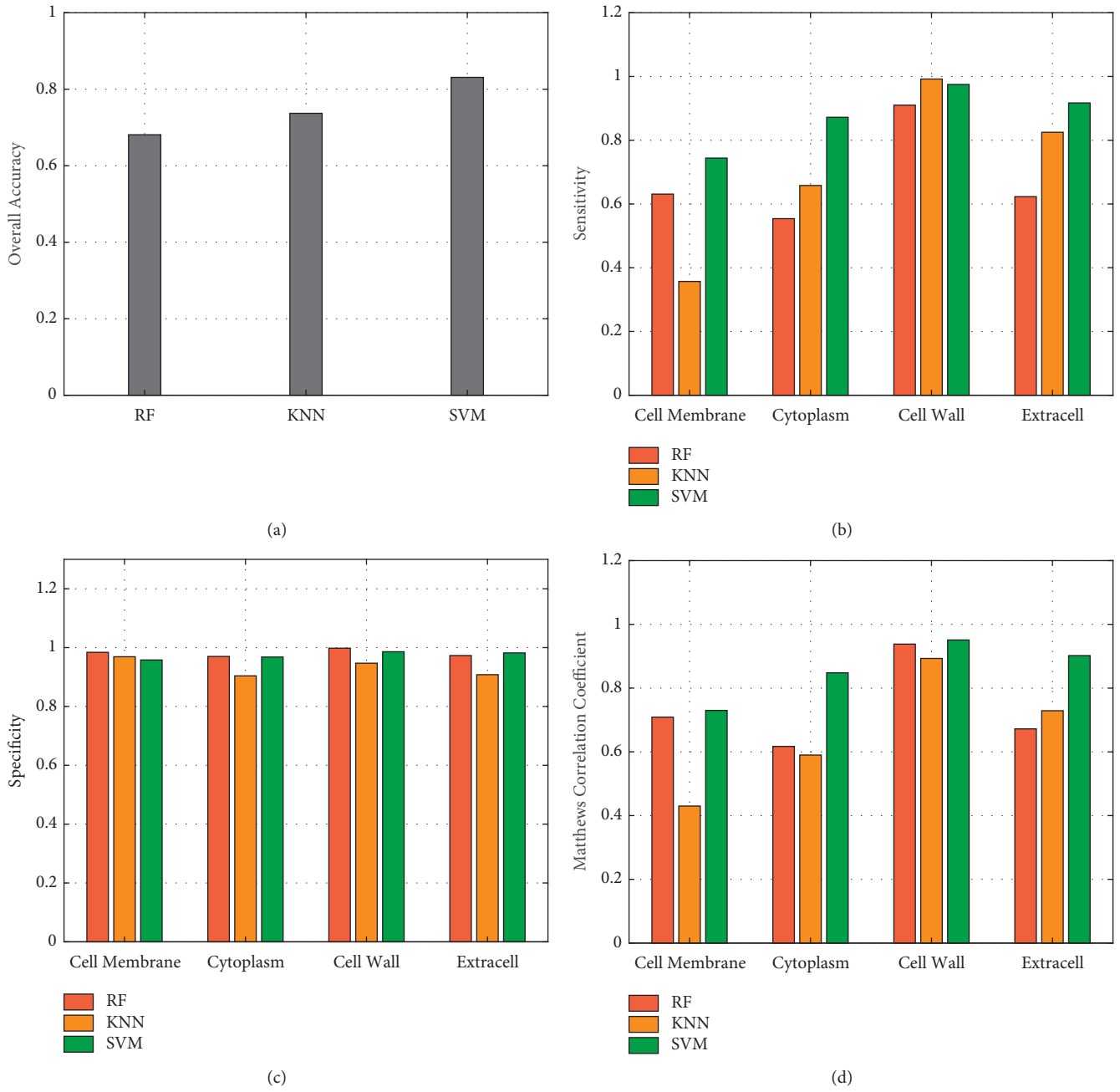

Figure 7: Comparison of evaluation index in various classification methods. (a) Overall accuracy, (b) sensitivity, (c) specificity, and (d) Matthews correlation coefficient of KNN, NBC, and SVM.

Table 6: Comparison of the proposed method with other methods on the Gpos-mPLoc dataset.

Method	Subcellular site				OA
	Cell membrane	Cytoplasm	Cell wall	Extracell	
(SN)					
AAO + PAAO*	0.625	0.832	0.799	0.669	0.751
PSORTb3.0	0.731	0.763	0	0.074	0.801
AAO + PSSM*	**0.744**	**0.872**	**0.975**	**0.917**	**0.831**
(SP)					
AAO + PAAO*	0.959	0.920	0.904	0.874	—
PSORTb3.0	0.910	0.760	1.000	0.954	—

Table 6: Continued.

Method	Subcellular site				OA
	Cell membrane	Cytoplasm	Cell wall	Extracell	
AAO + PSSM*	**0.958**	**0.968**	**0.986**	**0.982**	—
(MCC)					
AAO + PAAO*	0.659	0.740	0.693	0.545	—
PSORTb3.0	0.622	0.520	0	0.034	—
AAO + PSSM*	**0.730**	**0.848**	**0.951**	**0.902**	—

The bold values are the results of the proposed method in paper on the Gpos-mPLoc dataset for clearer comparison.

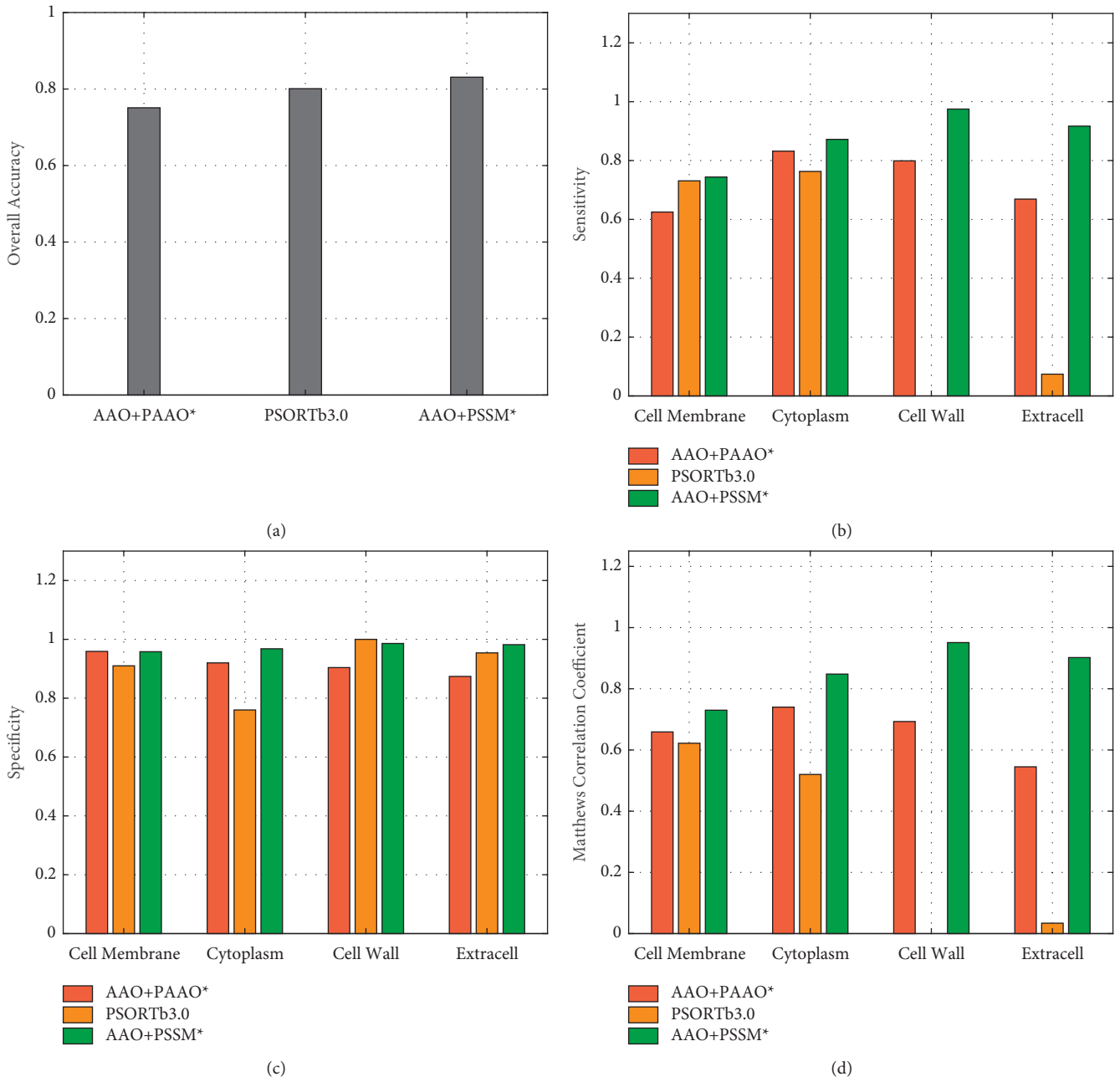

Figure 8: Comparison of evaluation index in the proposed method and other methods. (a) Overall accuracy, (b) sensitivity, (c) specificity, and (d) Matthews correlation coefficient of AAO + PAAO*, PSORTb3.0, and the method of this study.

Table 7: Experimental results of the proposed method on the Gneg-mPLoc dataset.

	Cim	Com	Cytoplasm	Extracell	Fimbrium	Flagellum	Nucleoid	Periplasm
SN	0.796	0.907	0.883	0.806	0.994	0.980	0.984	0.825
SP	0.992	0.990	0.995	0.986	0.999	0.995	0.996	0.984
MCC	0.845	0.907	0.905	0.820	0.992	0.968	0.974	0.827
OA					0.868			

5. Conclusions

Although the method AAO has some advantages, such that itis simple, the dimension of the feature vector is low, and the computing speed is fast, the feature only considers the overall evolution information of 20 amino acids, protein sequence arrangement of status after protein evolution, but ignores that the protein is arranged and other information, at the expense of the classification accuracy. Compared with AAO method, PSSM-AAO method and PSSM-GSD method have improved each index in Gram-positive protein dataset to a certain extent. These algorithms not only consider the evolution information of proteins, but also extract the sequence information of proteins to a certain extent, so the combination of AAO method, PSSM-AAO method, and PSSM-GSD method has achieved excellent classification effect. A single feature extraction method cannot completely extract protein information. Three feature extraction methods are combined to reflect the sequence information and evolutionary information of proteins in a relatively comprehensive way without increasing redundant information and time complexity as much as possible, laying a solid foundation for subsequent classification. In addition, the support vector machine also plays a great advantage in protein subcellular localization, which has better classification ability than k-nearest neighbor and random forest method. Comparing the proposed method with the existing protein subcellular localization methods, it can be found that the proposed method has improved to varying degrees in most indicators. But in fact, there is data imbalance in most datasets, which will seriously affect the accuracy of the model classification. Considering this, SMOTE oversampling algorithm is added in this study and the model performance is greatly improved.

Based on the existing research, this paper continuously deepens and improves the existing knowledge and proposes a new protein subcellular localization method. The main innovations and advantages are as follows:

(1) Based on the centralization trend and dispersion degree of the two most basic features of data distribution, a new feature extraction method PSSM-GSD is proposed. The feature method comprehensively reflects the data distribution of PSSM matrix and extracts protein evolution information as much as possible.

(2) SMOTE algorithm is used to greatly relieve the imbalance of protein dataset and improve the accuracy of protein subcellular localization to a certain extent.

(3) Considering that some proteins are located in two or more subcellular locations, the classifier chain algorithm is used for multilabel protein classification. The tags on the chain are sorted according to the conditional entropy from low to high, so as to avoid the influence of introducing wrong information on the final classification.

The overall accuracy of the proposed method is 83.1% and 86.8% on Gpos-mPLoc and Gneg-mPLoc datasets, and the overall performance is quite good and stable on different datasets, but there are still shortcomings. In the feature extraction method, we can further propose a method with more distinguishing ability and generalization ability. For example, improvements on classical algorithms such as PseAAC can be considered. In this paper, SVM with good classification ability is selected, which can be improved to enhance the comprehensive performance of the classification algorithm. Therefore, how to establish efficient and accurate feature extraction algorithm and classification algorithm is still the most important research of protein subcellular localization prediction.

Data Availability

The FASTA data used to support the findings of this study have been deposited in the Cell-PLoc 2.0 repository (http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

Acknowledgments

Publication costs are funded by the National Natural Science Foundation of China (grant no. 11271163).

References

[1] S. P. Qiao and B. Q. Yan, "Review of protein subcellular localization prediction," *Application Research of Computers*, vol. 9, 2014.

[2] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *Journal of Molecular Biology*, vol. 238, no. 1, p. 54, 1994.

[3] H. Yang, H. M. Xu, S. J. Yan, J. Chen, L. Geng, and Y. Yao, "Protein subcellular localization prediction based on reduced representation of amino acid and statistical characteristic," *Chinese Journal of Bioinformatics*, no. 2, pp. 29–36, 2015.

[4] X. J. Chen, X. Hu, and W. Xue, "Prediction of protein subcellular localization based on multilayer sparse coding," *Sheng wu gong cheng xue bao = Chinese journal of biotechnology*, vol. 35, no. 4, pp. 687–696, 2019.

[5] K. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins*, vol. 44, no. 1, p. 60, 2001.

[6] X. Xiao, X. Cheng, and S. Su, "pLoc-mGpos: incorporate key gene ontology information into general PseAAC for predicting subcellular localization of gram-positive bacterial proteins," *Natural Science*, vol. 09, no. 9, pp. 330–349, 2017.

[7] S. Zhang and D. Xin, "Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 437, p. 239, 2018.

[8] S. Wan, M. W. Mak, and S. Y. Kung, "HybridGO-loc: mining hybrid features on gene ontology for predicting subcellular localization of multi-location proteins," *PLoS One*, vol. 9, no. 3, Article ID e89545, 2014.

[9] L. L. Zhao, J. Wang, M. M. Nabil, and J. Zhang, "Deep forest-based prediction of protein subcellular localization," *Current Gene Therapy*, vol. 18, 2018.

[10] B. Yu, L. Shan, W. Qiu et al., "Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction," *BMC Genomics*, vol. 19, no. 1, p. 478, 2018.

[11] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, and A. Sattar, "Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 364, 2015.

[12] S. Wang and S. Liu, "Protein sub-nuclear localization based on effective fusion representations and dimension reduction algorithm LDA," *International Journal of Molecular Sciences*, vol. 16, no. 12, pp. 30343–30361, 2015.

[13] S. Wang, W. Li, F. Yu et al., "An improved process for generating uniform PSSMs and its application in protein subcellular localization via various global dimension reduction techniques," *IEEE Access*, vol. 7, no. 99, pp. 42384–42395, 2019.

[14] Z. Lei and D. Yang, "An SVM-based system for predicting protein subnuclear localizations," *BMC Bioinformatics*, vol. 6, 2005.

[15] S. Wan and M. W. Mak, "Predicting subcellular localization of multi-location proteins by improving support vector machines with an adaptive-decision scheme," *International journal of machine learning and cybernetics*, vol. 9, no. 3, pp. 399–411, 2018.

[16] S. P. Qiao, B. Yan, and J. Yan, "Ensemble learning for protein multiplex subcellular localization prediction based on weighted KNN with different features," *Applied Intelligence the International Journal of Artificial Intelligence Neural Networks & Complex Problem Solving Technologies*, vol. 48, 2018.

[17] K. Manaz, Y. Zheng, J. Chen et al., "SCLpred-EMS: subcellular localization prediction of endomembrane system and secretory pathway proteins by Deep N-to-1 Convolutional Neural Networks," *Bioinformatics*, vol. 36, no. 11, p. 11, 2020.

[18] S. Qiao and B. Yan, "Protein subcellular multi-localization prediction based on three-layer ensemble multi-label learning," *Journal of Computer Applications*, 2016.

[19] M. L. Zhang, L. I. Yukun, X. Y. Liu, and X. Gang, "Binary relevance for multi-label learning: an overview," *Frontiers of Computer Science*, vol. 12, no. 2, 2018.

[20] R. Wang, S. Ye, K. Li, and S. Kwong, "Bayesian network based label correlation analysis for multi-label classifier chain," *Information Sciences*, vol. 554, no. 8, 2020.

[21] S. K. Srivastava and S. K. Singh, "Multi-label classification of Twitter data using modified ML-KNN," *Advances in Data and Information Sciences*, pp. 31–41, 2019.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial IntelLigence Research*, vol. 16, pp. 321–357, 2002.

[23] J. Yu and J. Zhang, "Prediction of protein subcellular localization based on feature fusion and balanced dataset," *Information Technology and Informatization*, vol. 24, no. 3, pp. 137–139+142, 2021.

[24] S. F. Altschul, T. L. Madden, A. A. Schffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, p. 3389, 1997.

[25] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim, "Prediction of protein folding class using global description of amino acid sequence," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 19, pp. 8700–8704, 1995.

[26] K. C. Chou and H. B. Shen, "Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms," *Natural Science*, vol. 2, pp. 1090–1103, 2010.

[27] D. L. Zhang, S. Wang, and X. Y. Zhang, "Application of LIBSVM regression algorithm in coke strength prediction," *Iron & Steel*, vol. 53, no. 11, pp. 14–21, 2018.

[28] N. Y. Yu, J. R. Wagner, M. R. Laird et al., "PSORTb 3.0," *Bioinformatics*, vol. 26, no. 13, pp. 1608–1615, 2010.