Neural Computing and Applications
https://doi.org/10.1007/s00521-020-04792-z

ORIGINAL ARTICLE

SulSite-GTB: identification of protein S-sulfenylation sites by fusing multiple feature information and gradient tree boosting

Minghui Wang[1,2] · Xiaowen Cui[1,2] · Bin Yu[1,2,3] · Cheng Chen[1,2] · Qin Ma[4] · Hongyan Zhou[1,2]

Received: 5 February 2019 / Accepted: 17 February 2020

Abstract
Protein cysteine S-sulfenylation is an essential and reversible post-translational modification that plays a crucial role in transcriptional regulation, stress response, cell signaling and protein function. Studies have shown that S-sulfenylation is involved in many human diseases such as cancer, diabetes and arteriosclerosis. However, experimental identification of protein S-sulfenylation sites is generally expensive and time-consuming. In this study, we proposed a new protein S-sulfenylation sites prediction method SulSite-GTB. First, fusion of amino acid composition, dipeptide composition, encoding based on grouped weight, *K* nearest neighbors, position-specific amino acid propensity, position-weighted amino acid composition and pseudo-position specific score matrix feature extraction to obtain the initial feature space. Secondly, we use the synthetic minority oversampling technique (SMOTE) algorithm to process the class imbalance data, and the least absolute shrinkage and selection operator (LASSO) are employed to remove the redundant and irrelevant features. Finally, the optimal feature subset is input into the gradient tree boosting classifier to predict the S-sulfenylation sites, and the five-fold cross-validation and independent test set method are used to evaluate the prediction performance of the model. Experimental results showed the overall prediction accuracy is 92.86% and 88.53%, respectively, and the AUC values are 0.9706 and 0.9425, respectively, on the training set and the independent test set. Compared with other prediction methods, the results show that the proposed method SulSite-GTB is significantly superior to other state-of-the-art methods and provides a new idea for the prediction of post-translational modification sites of other proteins. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/SulSite-GTB/.

Keywords S-sulfenylation sites · Multi-information fusion · SMOTE · LASSO · Gradient tree boosting

1 Introduction

Protein post-translational modifications (PTMs) play a crucial regulatory role in cell life, affecting cell differentiation, signaling and regulation, stress response, gene expression regulation and protein–protein interactions [1]. Abnormal post-translational modifications may cause serious illnesses such as heart disease, diabetes [2]. Or the occurrence of certain diseases may be manifested in specific post-translational modification abnormalities [3]. Therefore, the identification of post-translational modification sites of proteins not only clarifies the molecular mechanism in cells but also reveals the intrinsic relationship between post-translational modifications and certain diseases. Currently, various regulatory enzymes involved in PTMs have become targets of many drugs. S-sulfenylation is a critical reversible post-translational modification that occurs when a cysteine residue thiol is covalently bonded to hydroxide, and the thiol function is oxidized to sulfenic, sulfinic and sulfurous acids. Previous studies have shown that S-sulfenylation modifications are involved in a variety of biological processes, including transcription

✉ Bin Yu
yubin@qust.edu.cn

1 College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China

2 Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China

3 School of Life Sciences, University of Science and Technology of China, Hefei 230027, China

4 Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA

Published online: 13 March 2020

factors, signaling proteins, metabolic enzymes, protein stabilizing regulators, cytoskeletal components, protein functions and protein interactions [4–10]. Therefore, understanding the regulatory mechanisms of S-sulfenylation is critical, and the first step is to determine the S-sulfenylation sites.

Nowadays, with the advancement of science and technology, there are several proteomic strategies for identifying S-sulfenylation sites of proteins. For example, Leonard et al. [11] identified S-sulfenylated protein in cells by liquid chromatography–tandem mass spectrometry (LC–MS/MS). Paulsen et al. [5] used the probe DYn-2 to detect S-sulfenylated proteins in human cells. However, these strategies are costly and time-consuming and lack of information on sites and redox diversity in experimental identification, both of which lead to false positive identification. Therefore, there is an urgent need to use the machine learning method to identify protein S-sulfenylation sites.

With the development of the post-genome era, bioinformatics calculation methods can quickly and efficiently identify protein post-translational modification sites [12], and machine learning algorithms are widely used in sites prediction. Only when protein sequence character data are converted to numeric vectors can it be accepted by machine learning algorithms. Therefore, feature extraction of protein sequences is a key step in the construction of protein post-translational modification sites prediction models. At present, the feature extraction methods of protein sequences are mainly based on sequence information, physical and chemical properties, structural information and evolutionary information. Weng et al. [13] developed the S-palmitoylation sites predictor MDD-Palm, which uses amino acid composition and position-specific scoring matrices to extract protein sequence information. Cui et al. [14] proposed a protein ubiquitination sites predictor UbiSitePred, which uses binary encoding, pseudo-amino acid composition, the composition of k-spaced amino acid pairs and position-specific propensity matrix to extract protein sequence feature information. Chen et al. [15] constructed the S-glutathionylation sites prediction model GSHSite, which converts the character information of the protein sequence into a numerical vector by BLOSUM62 and amino acid pair composition. Xie et al. [16] constructed the protein nitration and nitrosylation sites predictor DeepNitro, using positional amino acid distributions, sequence contextual dependencies, physicochemical properties and position-specific scoring matrix to extract protein sequence features.

Feature fusion will bring redundant information and generate "dimension disaster," which will cause trouble in the calculation and even affect the prediction result. Therefore, it is necessary to select the optimal feature subset of the fusion information and reduce the noise and eliminate the redundant information. Meanwhile, the valuable features are retained to the maximum extent, and the efficiency, performance and robustness of the prediction model are improved. Wuyun et al. [17] constructed a lysine acetylation sites prediction model using the Pearson correlation coefficient and stepwise feature selection method to select the optimal feature subset. Cai et al. [18] predicted protein ubiquitination sites and selected the optimal feature subsets using maximum correlation minimum redundancy and IFS. Wen et al. [19] constructed a methylation sites model and optimized the features by information gain to obtain the optimal feature subset. Zhao et al. [20] constructed the protein glycation sites predictor Glypre and used the maximum relevance minimum redundancy method and greedy feature selection algorithm to denoise the fusion features. Yu et al. [21] proposed a variety of features to propose lysine glycation sites predictor PredGly, using XGboost to optimize feature vectors and selecting the optimal feature subset to improve the model prediction performance.

However, the predictive performance of constructing a protein post-translational modification sites model is affected by the class imbalance of the training dataset. The imbalance of the dataset can lead to the prediction result to be biased toward multiple categories. Ning et al. [22] predicted protein succinylation sites, using bootstrap sampling and ensemble learning to solve the data imbalance problem. The accuracy and MCC on the training set were 89.14% and 0.79 via ten-fold cross-validation, and the accuracy on independent dataset was 84.5%. Zuo and Jia [23] developed the human protein carbonylation sites predictor CarSite, which used a one-sided selection resampling method, on ten-fold cross-validation of the Jia dataset; CarSite obtained rates of sensitivity corresponding to K/P/R/T-type peptides of 21%, 22%, 19% and 18% higher than those obtained by iCar-PseCp, respectively. Hu et al. [24] proposed the protein-nucleotide-binding residue predictor TargetSOS. To overcome the imbalance of sample dataset, a new supervised oversampling algorithm was proposed. Cross-validation test and independent validation test proved the effectiveness of TargetSOS. Jia and Zuo [25] established S-sulfenylation sites predictor S-SulfPred, using one-side selection undersampling and synthesis minority oversampling techniques to establish balance datasets. On independent datasets, the sensitivity was 74.62%, and specificity was 71.62%.

Nowadays, machine learning algorithms have become the core content and research hotspot in the field of artificial intelligence. At present, the most commonly used machine learning algorithms in bioinformatics are: artificial neural networks [26, 27], logistic regression [28, 29], random forest [30], support vector machines [14, 31],

ensemble learning [32, 33] and deep learning [34, 35] and so on. Johansen et al. [26] proposed a protein glycosylation sites predictor GlyNN based on an artificial neural network with an MCC value of 0.58. Li et al. [29] designed the phosphorylation sites predictor Quokka, which used logistic regression classifier. The independent test set shows that Quokka significantly improves prediction performance compared to the most advanced prediction tools used for phosphorylation prediction. Li et al. [30] constructed a species-specific acetylation sites predictor SSPKA based on random forest classifier. Qiu et al. [31] predicted the crotonylation sites using the support vector machine classifier with radial basis kernel function. The sensitivity, specificity, accuracy and Matthew's correlation coefficient were 71.69%, 98.7%, 94.43% and 0.778, respectively. Qiu et al. [32] constructed a lysine crotonylation sites predictor iKcr-PseEns using an ensemble random forest classifier. Sensitivity, specificity, accuracy and MCC were 90.53%, 95.27%, 94.49% and 0.826, respectively, through rigorous cross-validation. He et al. [35] proposed a deep architecture within multiple modalities to identify the ubiquitination sites, whose specificity, sensitivity, accuracy and MCC reached 66.4%, 66.7%, 66.43% and 0.221, respectively. The results showed that this method is superior to other prediction methods.

In the post-genome era, faced with such a large amount of post-translational modification data, it is urgent to develop a calculation method for timely identification of S-sulfenylation sites in proteins. Several calculation methods have been developed to predict S-sulfenylation sites. Based on the S-sulfenylation sites verified by the experiment, Bui et al. [36] developed the predictor MDD-SOH using the position-specific scoring matrix and BLOSUM62, and the support vector machine was used as the classifier. The specificity, sensitivity and MCC reached 0.70, 0.68 and 0.27, respectively. Bui et al. [37] developed a predictor SOHSite based on evolutionary information PSSM and 12AAindex for feature extraction and obtained Sn, Sp, ACC and MCC of 0.72, 0.69, 0.693 and 0.278, respectively. Xu et al. [38] proposed the predictor iSulf-Cys with an AUC value of 0.7155 via ten-fold cross-validation. The AUC and MCC were 0.7343 and 0.3315 on the independent test set, respectively. Sakka et al. [39] developed a predictive tool PRESS based on four structural features of proteins and SVM classifiers, which can effectively predict cysteine S-sulfenylation sites modification. Wang et al. [40] constructed an ensemble learning prediction tool SOHPRED based on Naïve Bayes, random forest and support vector machine classifier to identify human S-sulfenylation sites and obtained Sp, Sn and MCC of 0.712, 0.731 and 0.315, respectively. Jia and Zuo [25] designed the prediction model S-SulfPred, which used the one-side selection method to process the unbalanced dataset, and the support vector machine is employed as the classifier. On independent test set, the Sn, Sp, ACC and MCC reached 74.62%, 71.62%, 72.68% and 0.4441, respectively. Hasan et al. [41] constructed the predictor SulCysSite based on RF classifier. In the independent test set, Sp, Sn and MCC were 0.711, 0.766 and 0.343, respectively. Deng et al. [42] proposed a prediction model PredCSO, which adopted the maximum correlation feature selection algorithm and sequence backward elimination algorithm for feature selection. Combined with the bootstrap sampling, gradient tree boosting and majority voting, ACC, Sn, Sp and MCC were, respectively, 87.2%, 87.8%, 87.0% and 0.684. Ju and Wang [43] constructed the predictor Sulf-FSVM based on fuzzy support vector machine classifier. The predicted results of Sn, Sp, ACC, MCC and AUC on Xu's independent test set were 80.69%, 68.66%, 72.88%, 0.4712 and 0.8053, respectively. On Hasan's independent test set, Sn, Sp and MCC were 0.7981, 0.7969 and 0.4461, respectively. Wang et al. [44] used the C4.5 classification decision tree, CART and Best-first decision tree to design the S-sulfenylation sites predictor Fu-SulfPred. The results showed that on three independent test datasets, the MCC values of Fu-SulfPred model were 0.5437, 0.3736 and 0.6809, which are better than the MCC values of the S-SulfPred model.

It is worth noting that SOHSite, MDD-SOH, iSulf-Cys, SOHPRED and S-SulfPred are sequence-based prediction methods, and PRESS is a method based on protein structure. Despite these efforts, there is still room for improvement in the prediction model. First, these methods cannot fully elucidate the feature information of S-sulfenylation sites, and only one feature extraction method will lose part of the information of protein sequence. Second, MDD-SOH, SOHPRED, SulCysSite, SOHSite, iSulf-Cys, S-SulfPred, PredCSO and Fu-SulfPred do not carry out feature selection on the feature vector space, which may contain redundant information and noise, thus affecting the performance of the prediction model. Thirdly, MDD-SOH, SOHSite, SulCysSite and iSulf-Cys models are built based on unbalanced sample datasets, and the class imbalance of datasets is not processed.

Inspired by this, this paper proposes a new method to predict protein S-sulfenylation sites called SulSite-GTB, where SulSite represents S-sulfenylation Sites, and GTB represents Gradient Tree Boosting. First, we use amino acid composition, dipeptide composition, encoding based on grouped weight, K nearest neighbors, position-specific amino acid propensity, position-weighted amino acid composition and pseudo-position specific score matrix to extract the protein sequence features information. Through the five-fold cross-validation, we determine the optimal parameters of feature extraction. Secondly, the SMOTE algorithm is used to deal with the imbalance of positive and

negative data. Again, LASSO is selected to determine the optimal feature subset, removing redundant and uncorrelated features, and providing important feature information to the classifier. Finally, the S-sulfenylation sites prediction model based on a gradient boosting tree is constructed. The overall prediction accuracy, AUC and MCC values are 88.53%, 0.9425 and 0.7740, respectively, on independent test sets. The results show that SulSite-GTB method significantly improved the predictive performance of S-sulfenylation sites and non-S-sulfenylation sites compared to the most advanced tools for S-sulfenylation sites prediction.

2 Materials and methods

2.1 Datasets

To objectively and fairly evaluate the predictive model constructed in this paper, a reliable and rigorous benchmark dataset is significant. This article uses "Training data" as the training dataset to train the optimal parameters of the model. In addition, "Independent testing data" is used as an independent test set to evaluate the performance of the predictive model. Both the training set and the independent test set were constructed by Bui et al. [36]. The datasets were derived from the Carroll Lab, RedoxDB [45] and UniProtKB databases. CD-HIT [46] is an effective tool for clustering protein sequences based on specified sequence similarity values. To avoid redundancy and sequence homology, the CD-HIT program was used to remove homologous sequence fragments from the positive and negative datasets, ensuring that the pairwise sequence identity between any two peptides was less than 40%. The dataset contained a total of 1247 S-sulfenylation sites, which are considered positive samples, and 9446 non-S-sulfenylation sites are considered negative samples. The dataset is further divided into a training set and test set. The training set includes 1031 positive samples and 8028 negative samples. The test set contains 216 positive samples and 1418 negative samples. The sample sequence window size is 21.

2.2 Feature encoding

2.2.1 Amino acid composition

Amino acid composition (AAC) [47] is a feature extraction algorithm that reflects the frequency of 20 amino acids in each protein sequence. For a protein sequence sample fragment of a given form such as $S = R_1R_2\ldots R_N$, where N represents the total number of amino acids contained in the residue sequence, we calculate the frequency of each amino acid. The frequency of all 20 amino acids ("*ACDEFGHIKLMNPQRSTVWY*") uses an amino acid composition algorithm to convert the protein sequence into a 20-D vector, as shown in Eq. (1):

$$V_{AAC}(S) = (v_1, v_2, \ldots, v_{20}) \tag{1}$$

where $v_i = f_i / \sum_{i=1}^{20} f_i, (i = 1, 2, \ldots, 20)$, f_i represents the number of occurrences of the ith amino acid in the protein sequence.

2.2.2 Dipeptide composition

As an effective feature coding algorithm, it has been successfully applied to many biological problems [48, 49]. The dipeptide composition (DC) [49] is the frequency of occurrence of the amino acid pair for a given protein sequence, i.e., the frequency of the dipeptide. Totally, 20 common amino acids *ACDEFGHIKLMNPQRSTVWY* form the amino acid pair $AA, AC, AD, \ldots, YW, YY$, a total of 400. The feature vector elements extracted by the dipeptide composition algorithm are as shown in Eq. (2):

$$f_i = \sum_{j=1}^{n} R_j/(N-1), \quad i = 1, 2, \ldots, 400 \tag{2}$$

among them, R_j indicates the frequency of occurrence of the amino acid pair, and N is the length of the protein sequence.

2.2.3 Encoding based on grouped weight

The basic unit of protein is an amino acid, which forms various protein sequences through a series of biochemical reactions. Studies have found that different amino acids have different physical and chemical properties. Zhang et al. [50] proposed the encoding based on grouped weight (EBGW) method for feature extraction of protein sequences. According to the physicochemical properties (hydrophobicity and chargeability) of these amino acids, they are divided into four categories,

The hydrophobic group $C1 = \{G, A, V, L, I, M, P, F, W\}$
The polar group $C2 = \{Q, N, S, T, Y, C\}$
The negatively charged group $C3 = \{D, E\}$
The positively charged group $C4 = \{H, K, R\}$

The amino acid residue is partitioned on the basis of the following disjoint groups: $\{C1, C2\}$ versus $\{C3, C4\}$, $\{C1, C3\}$ versus $\{C2, C4\}$, $\{C1, C4\}$ versus $\{C2, C3\}$. Suppose a protein sequence $P = p_1p_2p_3\ldots p_N$ is divided according to the above division method. After mapping by Eqs. (3), (4) and (5), the sequence P becomes three binary sequences, which are respectively recorded as $H_1(P)$, $H_2(P)$, $H_3(P)$:

$$H_1(p) = \begin{cases} 1 & p_i \in \{C1, C2\} \\ 0 & p_i \in \{C3, C4\} \end{cases} \quad (3)$$

$$H_2(p) = \begin{cases} 1 & p_i \in \{C1, C3\} \\ 0 & p_i \in \{C2, C4\} \end{cases} \quad (4)$$

$$H_3(p) = \begin{cases} 1 & p_i \in \{C1, C4\} \\ 0 & p_i \in \{C2, C3\} \end{cases} \quad (5)$$

where $H_1(P)$, $H_2(P)$ and $H_3(P)$ are three second-order sequences of length N. Set a fixed parameter L, and the sub-sequences can be expressed as $[kN/L], \quad k = 1, 2, \ldots, L$. Where $[\bullet]$ is the integer operator. Then, we calculate the frequency of occurrence of 1 in each sub-sequence, and a $H_i(P)$ can be converted into L dimension vector. In conclusion, for a protein sequence of length N, $3L - D$ vector can be obtained.

2.2.4 *K* nearest neighbors

To take advantage of cluster information of local sequence fragments for predicting S-sulfenylated sites, we took the local sequence around S-sulfenylated sites in a query protein and extracted features from its similar sequences in both positive and negative datasets by a K nearest neighbors algorithm (KNN) [51]. The details are as follows:

1. According to the local sequence similarity, the KNN features are found in the positive dataset and the negative dataset, respectively. For two local sequences s_1 and s_2, the distance $\mathrm{Dist}(s_1, s_2)$ is defined as:

$$\mathrm{Dist}(s_1, s_2) = 1 - \frac{\sum_{i=-p}^{p} \mathrm{Sim}(s_1(i), s_2(i))}{2p + 1} \quad (6)$$

   where P represents the number of flanking residues in the central sites of the protein sequence fragment and i represents the position of the amino acid in the sequence fragment. Sim is an amino acid similarity scoring matrix.

$$\mathrm{Sim}(a, b) = \frac{M(a, b) - \min\{M\}}{\max\{M\} - \min\{M\}} \quad (7)$$

   where a and b are two amino acids, M is the BLOSUM62 substitution matrix, and $\max\{M\}/\min\{M\}$ is the maximum/minimum number in the BLOSUM62 substitution matrix.
2. After that, the corresponding KNN scores were then extracted as follows: (1) form a comparison dataset by combining positive and negative samples, (2) calculate the average distance from the sequence fragment to the comparison dataset, (3) sort the neighbors by the distances and choose the K nearest neighbors, (4) calculate the percentage of positive samples (S-sulfenylated sites) in its K nearest neighbors as the KNN features.

2.2.5 Position-specific amino acid propensity

A position-specific amino acid propensity was introduced in the prediction of protein phosphorylation sites [52], which used 20 natural amino acids and achieved excellent results. The amino acid at each position of the positive dataset is calculated by PSAAP and can be expressed as:

$$\left(a_{P,1,j}, a_{P,2,j}, \ldots, a_{P,21,j}\right)^T$$

where $a_{P,i,j}$ represents the frequency of the occurrence of the i amino acid at the position of j, and the character P represents the positive example sample, $i \in \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V, X\}$, $j = 1, 2, \ldots, 20$.

The number of negative training samples in this paper is approximately eight times the number of positive samples (1031:8028). We randomly built eight subsets, each containing 1031 negative samples. For the feature extraction of negative samples, the average amino acid composition and standard deviation of each position were calculated. It can be expressed as:

$$\left(\overline{a_{N,1,j}}, \overline{a_{N,2,j}}, \ldots, \overline{a_{N,21,j}}\right), \quad j = 1, 2, \ldots, 20$$
$$\left(\sigma_{N,1,j}, \sigma_{N,2,j}, \ldots, \sigma_{N,21,j}\right), \quad j = 1, 2, \ldots, 20$$

The PSAAP matrix $Z = \left(Z_{i,j}\right)_{21 \times 20}$ obtained from,

$$Z_{i,j} = \frac{a_{P,i,j} - \overline{a_{N,i,j}}}{\sigma_{N,i,j}}, \quad i = 1, 2, \ldots, 21, \quad j = 1, 2, \ldots, 20 \quad (8)$$

Each row of matrix Z represents an amino acid, and each column of matrix Z represents a position in each peptide fragment. When this feature is encoded, each protein sequence is converted into a 20-dimensional feature vector. The flow of the PSAAP feature extraction algorithm is shown in Fig. 1.

2.2.6 Pseudo-position specific scoring matrix

Jones et al. [53] proposed position specific scoring matrix (PSSM) and used the iterative PSI-BLAST search method to find evolutionary information of protein sequences, which has been widely used to predict the subcellular location of apoptotic proteins [54, 55] and the protein submitochondrial locations [56]. In this paper, the PSI-BLAST program [57] was mainly used to compare all protein sequences in the dataset with the non-redundant SwissProt database. The parameters of the PSI-BLAST in this process had an E-value threshold of 0.001 and a maximum number of iterations of 3. The remaining

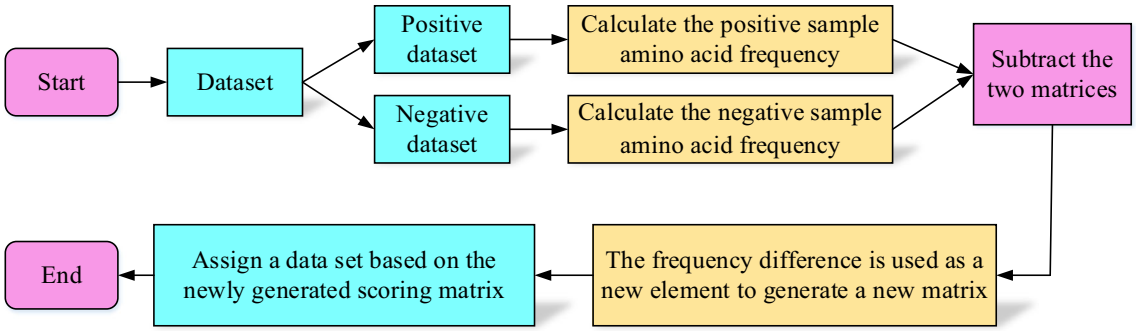

Fig. 1 The feature representation of PSAAP

parameters are set as the default values. For each protein sequence with a window size of N, feature extraction is performed to obtain its position specific scoring matrix, which is the matrix of $N \times 20$, as shown in Eq. (9).

$$P_{\mathrm{PSSM}} = \begin{pmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,20} \\ \vdots & \vdots & \vdots & \vdots \\ P_{i,1} & P_{i,2} & \cdots & P_{I,20} \\ \vdots & \vdots & \vdots & \vdots \\ P_{N,1} & P_{N,2} & \cdots & P_{N,20} \end{pmatrix} \tag{9}$$

This paper first normalizes the PSSM matrix and uses Eq. (10) to convert the elements between 0 and 1.

$$f(x) = \frac{1}{1+e^{x}} \tag{10}$$

that is,

$$\overline{P_{\mathrm{PSSM}}} = (\overline{P_1}, \overline{P_2}, \ldots, \overline{P_{20}}) \tag{11}$$

among them, $\overline{P_j}$ represents the composition of amino acid j in PSSM, and the model of $\overline{P_{\mathrm{PSSM}}}$ is called: PSSM-AAC [58].

However, this method only considers the average score of the amino acid residue mutated into the amino acid j in the protein sequence. Without considering the order information of the amino acid residues of the protein sequence, to overcome this disadvantage, PsePSSM [59] was proposed. The feature extraction process is as follows:

$$P^{\xi}_{\mathrm{PsePSSM}} = (\theta^{\xi}_1, \theta^{\xi}_2, \ldots, \theta^{\xi}_j, \ldots, \theta^{\xi}_{20})^T \tag{12}$$

$$\theta^{\xi}_j = \frac{1}{L-\xi}\sum_{i=1}^{L-\xi}(P_{i,j} - P_{(i+\xi),j})^2 \quad (j = 1, 2, \ldots, 20; \ \xi < N, \quad \xi \neq 0) \tag{13}$$

where θ^{ξ}_j represents the ξ-order correlation factor of amino acid j. To sum up, a protein sequence can be expressed by PsePSSM as Eq. (14).

$$P_{\mathrm{PsePSSM}} = (\overline{P_1}, \overline{P_2}, \ldots \overline{P_{20}}, \theta^1_1, \theta^1_2, \ldots, \theta^1_{20}, \ldots, \theta^{\xi}_1, \theta^{\xi}_2, \ldots, \theta^{\xi}_{20})^T \tag{14}$$

As we can see from Eq. (14), the PsePSSM algorithm generates $20 + 20 \times \xi$ dimensional vector for a protein sequence.

2.2.7 Position weight amino acid composition

To reveal sequence information around the cysteine S-sulfenylation sites, we used position weight amino acid composition to extract sequence position information for amino acid residues. Position weight amino acid composition feature extraction methods have been used to identify protein phosphorylation sites [60], predict methylation sites [61] and distinguish between lysine acetylation and lysine methylation [62].

Given a protein sequence P of length $2m+1$, the positional information of the amino acid residue a_i $(i = 1, 2, \ldots, 20)$ is given by Eq. (15),

$$c_i = \frac{1}{m(m+1)}\sum_{j=1}^{m} x_{i,j}\left(j + \frac{|j|}{m}\right) \tag{15}$$

where m represents the number of upstream or downstream residues of the central site in protein sequence fragment P. If a_i is the j residues in protein sequence fragment P, then $x_{i,j} = 1$, otherwise $x_{i,j} = 0$. Finally, protein fragment P is encoded as a 20-dimensional feature vector.

2.2.8 Feature space

To extract more characteristic information from the protein sequence, AAC, DC, EBGW KNN, PSAAP, PWAAC and PsePSSM coding methods are fused to obtain the feature space of the protein sequence, which is recorded as "ALL." The dimension of each feature vector obtained is $20 + 441 + 63 + 30 + 20 + 20 + 420 = 1014$.

2.3 SMOTE

SMOTE [63] is an oversampling algorithm proposed by Chawla et al., which aims to synthesize some new positive

samples to reduce the class imbalance. It has been applied to protein–protein interaction sites prediction [64], drug-target interactions prediction [65], predicting protein sub-mitochondrial localization [66] and tumor classification [67]. The algorithm is briefly introduced as follows:

Given a positive sample X, search for its K nearest neighbor samples. If the oversampling ratio is N, then N nearest neighbor samples are selected from the K nearest neighbor samples, denoted as $c_1, c_2, \ldots, c_N$. A random linear interpolation can be performed between the positive samples X and $c_1, c_2, \ldots, c_N$, and a new positive sample P_j is generated by Eq. (16):

$$P_j = X + \text{rand}(0,1) * (c_j - X), \quad j = 1, 2, \ldots, N \tag{16}$$

where $\text{rand}(0,1)$ represents the random number within $(0,1)$.

2.4 LASSO

High-dimensional feature sets typically contain noise and redundancy features that are detrimental to the model's predictive performance. For example, the amino acid index database contains 544 physicochemical properties of each amino acid. For a 21 residue peptide, the dimension of the amino acid index feature will be 11,424, and these 11,424 features are not equally important for predicting PTM sites. Therefore, feature selection is a very important step before model construction, which can measure the importance of all features and eliminate the features with less information. It can also reduce feature dimensions and eliminate misleading features so that the model can achieve better predictive performance. In this paper, LASSO is used for feature selection, and the optimal feature subset is selected to improve the prediction performance of the model.

Given a dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, where $x \in R^d, y \in R$ subject to the squared error as the loss function, the optimization goal is

$$\min_w \sum_{i=1}^{m} (y_i - w^T x_i)^2 \tag{17}$$

To reduce the risk of over-fitting, Tibshirani [68] proposed a compression estimation method LASSO (least absolute shrinkage and selection operator) in 1996, which has been successfully applied to predict protein ubiquitination sites [14], drug-target interactions prediction [65] and tumor classification [67]. The basic idea is to introduce $L1$ norm regularization from minimizing residuals sum of squares. The LASSO sparse representation coefficient w can be described as follows:

$$J(w) = \min_w \sum_{i=1}^{m} (y_i - w^T x_i)^2 + \lambda \|w\|_1 \tag{18}$$

The regularization parameter λ controls the degree of punishment of the sparse coefficient estimation, and $\|w\|_1$ is a $L1$ norm. Feature vector corresponding to a non-zero solution will appear in the final model. $\lambda \geq 0$ is the regularization parameter in the model, and it represents the balance between the complexity of the model and the degree of data fitting. Equation (18) is optimized using the coordinate gradient descent method.

2.5 Gradient tree boosting

Gradient tree boosting [69] is an ensemble learning classification algorithm for multiple decision trees with excellent learning ability, which has been successfully applied in many fields, such as symptom severity classification [70], single amino acid variations [71] and resultant protein solvent accessibility [72]. The basic idea is that when training each regression tree, GTB fits residual of a tree at each iteration. The residual is the predicted value of the tree minus the residual by the previous tree which multiplied by a learning rate η. In general, the smaller the η, the higher the accuracy of the resulting GTB model.

Training set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, given the x value and the corresponding value y, according to the principle of risk experience minimization, find the approximate value $\widehat{F}(x)$ of the objective function $F(x)$ and minimize the expected value of loss function $L(y, F(x))$.

$$\widehat{F} = \arg\min_F E_{x,y}[L(y, F(x))] \tag{19}$$

Suppose $h_m(x)$ is the base decision tree of step m and J_m is the number of leaf nodes in the base decision tree. The input space was divided into J_m disjoint regions $R_{1m}, R_{2m}, \ldots, R_{J_m m}$, and the output value in each region was predicted. The output $h_m(x)$ is expressed as a linear combination of input x:

$$h_m(x) = \sum_{j=1}^{J_m} b_{jm} I_{R_{jm}}(x) \tag{20}$$

where b_{jm} is the predicted value in the region R_{jm}.

Then, the coefficient b_{jm} is multiplied by the learning rate r_m, the value of the leaf node region is estimated by linear search, the approximate value of the residual is fitted, and the loss function is minimized. The model is iteratively updated as follows:

$$\begin{aligned} F_m(x) &= F_{m-1}(x) + r_m h_m(x), \\ r_m &= \arg\min_r \sum_{i=1}^{n} L(y_1, F_{m-1}(x_i) + r h_m(x_i)) \end{aligned} \tag{21}$$

2.6 Performance evaluation and model building

To build the best prediction model, the prediction performance of the model is evaluated by five-fold cross-validation and independent test. The dataset was randomly divided into five subsets of roughly equal size by using the five-fold cross-validation evaluation test. One of the five sets was selected as the test set, and the other four were selected as the training set, and the cross-validation process was repeated five times. The final results are generated by averaging the five validation results. For the independent test, the whole training dataset is used for the training model, and the independent test set is used to evaluate the performance of the training model.

To intuitively measure the predictive performance of the model, the following evaluation indexes were used in this study [56, 64, 73, 74]: sensitivity (Sn), specificity (Sp), accuracy (ACC) and Matthews correlation coefficient (MCC):

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{22}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{23}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{24}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}} \tag{25}$$

where TP represents the number of true positives, FP represents the number of false positives, TN represents the number of true negatives, and FN represents the number of false negatives. Sn is the percentage of correct prediction of positive data (cysteine S-sulfenylation), and Sp is the correct prediction of negative data (cysteine non-S-sulfenylation). Accuracy reflects the overall proportion of correct predictions, and the higher positive MCC value indicates the better prediction of correctly classified positive and negative data. The MCC value ranges from -1 to 1, and the values of Sn, Sp and ACC range from 0 to 1.

For convenience, the S-sulfenylation sites prediction method proposed in this paper is called SulSite-GTB, and the calculation process is shown in Fig. 2. The experimental environment is Windows Server 2012R2 Intel (R) Xeon (TM) CPU E5-2650 @ 2.30 GHz 2.30 GHz with 32.0 GB of RAM, MATLAB2014a and Python 3.6 programming implementation.

The steps of the SulSite-GTB method are described as:

1. Data collecting and preparation. Obtain training datasets and independent test sets, S-sulfenylation and non-S-sulfenylation sites and their corresponding label.
2. Feature extraction. We use AAC, DC, EBGW, PsePSSM, KNN, PSAAP and PWAAC to extract the sequence feature vectors. And the sequence information, physicochemical information and evolutionary information via multi-information fusion.
3. Balance data and feature selection. SMOTE algorithm is used to deal with data imbalance. For the feature vector of the protein sequence, LASSO is adapted to remove redundant and noise information, and the optimal feature subset is selected through the five-fold cross-validation, to provide good feature information for the classifier.
4. Model construction and evaluation. According to steps 2 and 3, the selected optimal feature subset and the corresponding category label are input into the gradient tree boosting classifier to predict the post-translational modification of the S-sulfenylation sites. The model can be evaluated via five-fold cross-validation on training data. The model is also tested using an independent test set. And we use AUC, ACC, Sn, Sp, MCC and AUPR values to verify the model performance.

3 Results

3.1 Sequence analysis of protein cysteine S-sulfenylation sites

To investigate the difference of the frequency and distribution of 20 amino acids around the S-sulfenylation sites and the non-S-sulfenylation sites, we used the Two Sample Logos [75] web server (http://www.twosamplelogo.org/cgi-bin/tsl/tsl.cgi) to analyze training dataset. Cysteine is placed in the middle of the fragment sequence, and the position of the flanking amino acid is described in the range of -10 to $+10$. The analysis results are shown in Fig. 3.

As can be seen from Fig. 3, there is a significant difference between the composition of the residues in the training dataset S-sulfenylated and non-S-sulfenylated cysteine. Positively charged residue lysine (*K*) at positions -10, -8, -7, -6, -4, -2, $+4$, $+7$, $+8$ and $+9$, negatively charged residues glutamic (E) at positions -10, -7, -5, -4, -3, $+1$, $+3$, $+4$ and $+7$, serine (S) at positions -3, -1, $+2$ and $+3$, threonine (T) was at positions -1 and $+1$, and proline (P) at positions -1, $+3$ and $+4$ was abundant, but not in negative samples. Residues such as tyrosine (Y), cysteine (C), histidine (H) and non-polar phenylalanine (F) appear significantly larger in the negative sample. Based on these

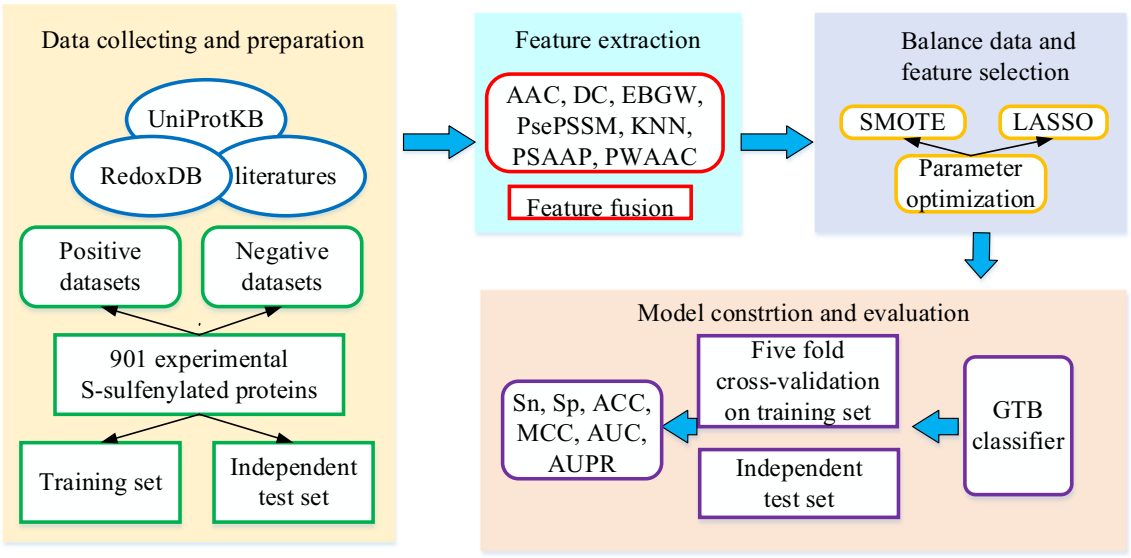

Fig. 2 Flowchart of the SulSite-GTB prediction method

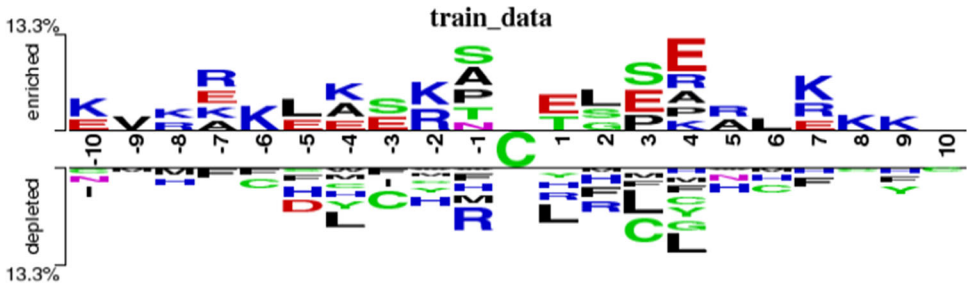

Fig. 3 Comparison of sequence identification of S-sulfenylation and non-S-sulfenylation in training datasets

characteristics, it can be inferred that the frequency difference in the appearance of various amino acids in the training dataset at different positions significantly affects the S-sulfenylation process of the cysteine. The distance plays a crucial role in distinguishing between S-sulfenylation sites and non-sulfenylation sites.

3.2 Selection of model parameters

The selection of parameters is significant for constructing predictive models. How to extract effective feature information from protein sequences is the key to the success of a protein post-translational modification sites prediction model. To better find the advantages and disadvantages of the characteristic parameters, it is usually necessary to compare, the S-sulfenylation sites and the non-S-sulfenylation sites protein sequence dataset as the research object and select the optimal parameters of the prediction model. In this paper, the EBGW and PsePSSM algorithm are used to extract the feature of the protein sequence. In the feature extraction process of these two algorithms, the selection of L value and ξ value plays a crucial role in the model construction. The ξ value represents the order information of the amino acid of the protein sequence. To find the optimal ξ value in the model, we set the values from 1 to 21. For different ξ values, the gradient tree boosting algorithm is used to classify the training set, and the results are tested by the five-fold cross-validation method, and the overall prediction accuracy is obtained, as shown in Table 1. EBGW converts protein sequence character information into $3L - D$ a numeric vector. For different parameter L values, the gradient tree boosting algorithm is used to classify the training dataset, and the results are tested by the five-fold cross-validation, and the overall prediction accuracy is acquired, as shown in Table 2.

From Table 1, it is concluded that changing the ξ value will result in different prediction results. As the value of ξ changes, the prediction accuracy and AUC value of the prediction model for the training dataset also change. When $\xi = 20$, the overall prediction accuracy and AUC value reached the highest 72.67% and 0.8085, respectively. Among them, comparing $\xi = 20$ and $\xi = 1$, ACC and AUC are 6.01% and 8.34% higher, respectively. At $\xi = 4$ and $\xi = 7$, ACC and AUC decreased, but the overall trend is growing. Parameter $\xi = 20$ is better than $\xi = 19$, and ACC and AUC are increased by 1.85% and 2.46%,

Table 1 Prediction results of S-sulfenylation sites for different ξ values on training set

ξ	$\xi = 1$	$\xi = 2$	$\xi = 3$	$\xi = 4$	$\xi = 5$	$\xi = 6$	$\xi = 7$	$\xi = 8$	$\xi = 9$	$\xi = 10$
ACC	66.66	67.68	68.06	67.88	68.04	68.62	68.39	69.09	68.95	68.79
AUC	0.7251	0.7403	0.7474	0.7445	0.7518	0.7618	0.7562	0.7642	0.7599	0.7625
ξ	$\xi = 11$	$\xi = 12$	$\xi = 13$	$\xi = 14$	$\xi = 15$	$\xi = 16$	$\xi = 17$	$\xi = 18$	$\xi = 19$	$\xi = 20$
ACC	68.89	69.80	69.76	69.71	69.98	69.80	70.25	70.75	70.82	72.67
AUC	0.7604	0.7701	0.7708	0.7672	0.7735	0.7698	0.7758	0.7835	0.7839	0.8085

Table 2 Prediction results of S-sulfenylation sites for different L values on training set

L	$L = 1$	$L = 2$	$L = 3$	$L = 4$	$L = 5$	$L = 6$	$L = 7$
ACC (%)	57.47	63.78	67.98	68.45	70.27	68.65	71.03
AUC	0.6303	0.7279	0.7709	0.7777	0.7972	0.7856	0.8080
L	$L = 8$	$L = 9$	$L = 10$	$L = 11$	$L = 12$	$L = 13$	$L = 14$
ACC (%)	70.61	72.51	72.03	70.32	70.30	72.87	74.29
AUC	0.7989	0.8250	0.8143	0.8102	0.8177	0.8302	0.8381
L	$L = 15$	$L = 16$	$L = 17$	$L = 18$	$L = 19$	$L = 20$	$L = 21$
ACC (%)	73.65	74.40	74.00	72.98	73.94	74.36	74.41
AUC	0.8349	0.8381	0.8424	0.8315	0.8320	0.8424	0.8391

respectively. Considering the overall consideration, the optimal $\xi = 20$ value is selected in the model. The PsePSSM algorithm is used to extract the features of the protein sequence, and each protein sequence generates $20 + 20 \times \xi = 20 + 20 \times 20 = 420 - D$ feature vector.

As can be seen from Table 2, by changing the L value, the prediction accuracy and AUC value in the training dataset also change. When the parameter $L = 21$, the prediction accuracy reaches a maximum of 74.41%. When $L = 17$ and $L = 20$, the AUC value reaches a maximum of 0.8424. When $L = 21$ compared with $L = 1$, the prediction accuracy and AUC were improved by 16.94% and 20.88%, respectively. At $L = 10$, $L = 11$ and $L = 12$, the prediction accuracy and AUC value decreased, but at $L = 13$, the accuracy and AUC began to increase. Therefore, $L = 21$ is selected as the optimal value of feature extraction parameter. By EBGW algorithm, each protein sequence can get $3L = 3 \times 21 = 63 - D$ vector.

3.3 The effect of feature extraction algorithm on prediction results

Choosing the appropriate feature extraction method is a crucial step for constructing the best predictive model for distinguishing S-sulfenylation and non-S-sulfenylation sites. In this paper, the predictive model is trained by eight feature extraction methods, such as AAC, DC, EBGW, KNN, PSAAP, PsePSSM, PWAAC and ALL (where ALL represents the feature vector after the fusion of the seven features). The PsePSSM algorithm is an evolutionary information feature based on sequence homology that converts character information into a numerical vector. The EBGW utilizes the physicochemical properties of amino acids to obtain numerical features of protein sequences. The AAC and DC calculate the frequency of the amino acid and the frequency of the amino acid pair, respectively. For each protein sequence, AAC, DC, EBGW, KNN, PSAAP, PsePSSM, PWAAC and ALL feature extraction methods are used to obtain the feature vectors of 20, 441, 63, 30, 20, 420, 20 and 1014 dimensions, respectively. To verify the effectiveness of the feature extraction algorithm, the gradient tree boosting is selected as the classifier to predict the S-sulfenylation sites, and ACC, Sn, Sp, MCC, AUC and AUPR are used as the measurement indexes to evaluate the prediction performance of the model.

The eight feature vectors are, respectively, inputted into the classifier of the gradient tree boosting for the prediction of S-sulfenylation sites, and the prediction results of different feature extraction algorithms in the training dataset are obtained by using the five-fold cross-validation method

Table 3 The prediction results of different features on the training dataset

Algorithm	ACC (%)	Sn (%)	Sp (%)	MCC	AUC	AUPR
AAC	74.88	77.75	71.97	0.5026	0.8247	0.8017
EBGW	74.41	81.28	67.42	0.4994	0.8391	0.8030
KNN	65.31	68.37	62.20	0.3065	0.7152	0.7143
PWAAC	75.12	80.34	69.82	0.5047	0.8375	0.8298
PSAAP	72.44	80.70	64.04	0.4541	0.8080	0.7923
PsePSSM	72.67	77.64	67.62	0.4570	0.8085	0.7961
DC	71.97	85.13	58.59	0.4637	0.8397	0.7788
All	84.25	83.29	85.23	0.6853	0.9164	0.8889

for evaluation, as shown in Table 3. In addition, this paper uses ROC(receiver operating characteristic, ROC) curve [76] and PR (precision recall, PR) curve [74] to compare the robustness of the prediction model under different feature extraction algorithms. Figure 4 is the ROC curve and PR curve obtained by the training set under the eight feature extraction methods.

It can be seen from Table 3 that with different feature extraction algorithms, the prediction accuracy of the training dataset is also different. We used the DC, PSAAP, AAC, EBGW, PWAAC, PsePSSM and KNN to obtain overall prediction accuracy of 71.97%, 72.44%, 74.88%, 74.41%, 75.12%, 72.67% and 65.31%, respectively. The AUC values of DC, PSAAP, AAC, EBGW, PWAAC, PsePSSM and KNN are 0.8397, 0.8080, 0.8247, 0.8391, 0.8375, 0.8085 and 0.7152, respectively. The fusion feature ALL obtained better prediction results, and each evaluation index exceeded the single feature extraction. The values of ACC, Sn, Sp, MCC, AUC and AUPR were 84.25%, 83.29%, 85.23%, 0.6853, 0.9164 and 0.8889. The ACC value increased by 9.37%–18.94%, the MCC value increased by 18.06–37.88%, and the AUPR value increased by 5.91%–17.46%. In summary, feature fusion can obtain more comprehensive feature information of protein sequences, which is helpful for predicting S-sulfenylation sites, and the prediction performance is better than a single feature extraction algorithm.

As can be seen from Fig. 4a, ROC curves with different feature extraction methods have different coverage areas. The ROC curve using ALL feature coding completely contains the ROC curve of other feature extraction algorithms, with the largest coverage area, i.e., the AUC value is 0.9164, which is significantly higher than the other seven methods. The area AUC value under the ROC curve of the DC is 0.8397. The AUC value of the ROC curve of the KNN feature algorithm is the minimum, which is 0.7152. The AUC of DC is 12.45% higher than that of the KNN algorithm. EBGW feature extraction algorithm is used to extract the physical and chemical properties of protein sequences, and the AUC value is 0.8391, which is inferior to the DC extraction algorithm. The ROC curve of the PWAAC feature extraction algorithm completely contains the ROC curve of the AAC, PsePSSM and PSAAP feature extraction algorithms, indicating that the prediction model of protein S-sulfenylate sites of the PWAAC feature extraction algorithm plays a certain positive role. The AUC values of ALL feature codes are 9.17%, 7.73%, 20.12%, 7.89%, 10.84%, 10.79% and 7.67% higher than those of AAC, EBGW, KNN, PWAAC, PsePSSM, PSAAP and DC, respectively. As can be seen from Fig. 4b, AUPR values of AAC, EBGW, KNN, PWAAC, PsePSSM, PSAAP, DC and ALL are, respectively, 0.8017, 0.8030, 0.7143, 0.8298,

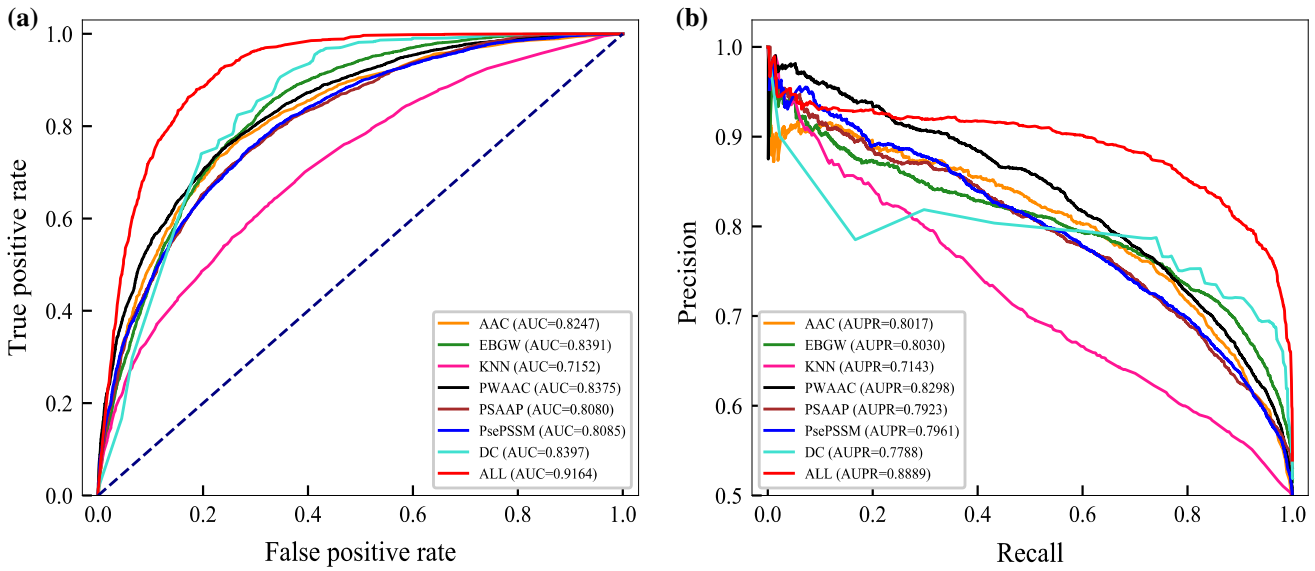

Fig. 4 Prediction results of different features algorithms on the training dataset. **a** The graph is the ROC curve, and **b** the graph is the PR curve

0.7923, 0.7961, 0.7768 and 0.8889. AUPR values of ALL are improved by 5.91%–17.46%. The AUPR values of ALL feature extraction are 8.72%, 8.59%, 17.46%, 5.91%, 9.66%, 9.28% and 11.01% higher than that of AAC, EBGW, KNN, PWAAC, PSAAP, PsePSSM, DC and ALL, respectively. Therefore, the ROC and PR curve of ALL cover the largest area, indicating that the ALL feature extraction algorithm provides the optimal feature subset for the prediction of S-sulfenylation sites.

By comparing the influences of different feature coding methods on the prediction results and comparing the robustness of eight different feature coding methods from ROC curve and PR curve, we determined that ALL features fused with seven features are the best feature extraction algorithm in this study. The applications of feature fusion algorithms are further compared, and the differences between the existing methods and this paper are explained in detail. In Table 4, we summarize 11 fusion feature extraction algorithms for cysteine S-sulfenylation sites prediction.

It can be clearly seen from Table 4 that the model SulSite-GTB fuses seven feature extraction methods to extract protein sequence information, physicochemical properties and evolutionary information. While other existing prediction methods only incorporate 2–4 feature extraction methods, PRESS only extracts structural information of protein sequences. Therefore, the method SulSite-GTB in this paper extracts protein characteristic information more comprehensively.

3.4 The effectiveness of the SMOTE algorithm

The imbalance of the dataset will lead to the biased prediction results. There is a severe imbalance between the datasets in this paper. The number of truly known S-sulfenylation sites is significantly smaller than that of non-S-sulfenylation sites, which will lead to the reduced prediction performance of the model. Therefore, to improve the generalization ability of classifier, we attempted to use the SMOTE algorithm to randomly undersample most samples of non-S-sulfenylation sites while randomly oversampling a few samples of known S-sulfenylation sites. This paper obtains the protein sequence feature vector through various feature extraction algorithms. To evaluate the effectiveness of the SMOTE algorithm, we compare balanced and unbalanced dataset. At the same time, on the basis of five-fold cross-validation, GTB classification algorithm is adopted to obtain the prediction results of the model, as shown in Table 5.

As can be seen from Table 5, for the training dataset, the prediction results are very different on the balanced dataset and the unbalanced dataset. The evaluation index Sn, MCC and AUC obtained on the balanced dataset after SMOTE processing has a greater advantage than the result on the unbalanced dataset. The indicators Sn and Sp measure the proportion of the positive and negative examples in the sample being correctly classified. On the unbalanced dataset, the numbers of positive sample and negative sample are 1031 and 8028, respectively. We can find the number of positive samples is far less than the number of negative samples, the index Sn value is small and Sp gets higher precision, which is mainly determined by the number of positive and negative samples.

The numbers of positive sample and negative sample are 7217 and 7097 after SMOTE algorithm. The AUC is the area under the ROC curve, which is the indicator to evaluate the robustness of model. Therefore, among all the above indicators, the index that can reasonably measure the performance of the prediction model is AUC, and the dataset is balanced by the SMOTE algorithm, and the AUC

Table 4 Comprehensive summary of the prediction model of cysteine S-sulfenylation sites in the fusion of multiple feature information

Model	Multiple features
MDD-SOH [36]	PSSM, BLOSUM62
S-SulfPred [25]	PSAAP, AAindex
SOHSite [37]	PSSM, 12AAindex
iSulf-Cys [38]	BE, PSAAP, AAindex
PRESS [39]	Neighboring amino acids in the protein sequence, neighboring amino acids in the 3D space, ASA, SS
SOHPRED [40]	AAC, physicochemical properties, CKSAAP, sequence profiles
SulCysSite [41]	AAindex, BE, PSSM, pCKSAAP
PredCSO [42]	PSSM, PSAAP, ASA, four-body statistical pseudo-potential
Sulf_FSVM [43]	AAF, BE, CKSAAP
Fu-SulfPred [44]	PSAAP, AAindex
SulSite-GTB	AAC, DC, EBGW, KNN, PSAAP, PsePSSM, PWAAC

Table 5 Comparison of predict result before and after SMOTE method on training dataset

Resampling method	ACC (%)	Sn (%)	Sp (%)	MCC	AUC
Without any resampling	88.31	1.75	99.41	0.0449	0.6639
SMOTE	89.29	86.24	92.39	0.7875	0.9506

value is increased by 28.67%. The SMOTE algorithm balances the dataset by generating a "synthesized" sample of S-sulfenylation sites to increase the proportion of S-sulfenylation sites in the data. Therefore, through the above comparative analysis, we can significantly improve the prediction performance of the model after SMOTE processing.

3.5 The influence of feature selection algorithm on prediction results

After feature extraction and fusion, more protein sequence information is obtained, but it brings more redundant information. To construct an ideal protein S-sulfenylation sites prediction model, seven feature extraction algorithms are fused to get the feature vector ALL of the protein sequence. In the PsePSSM algorithm, we select $\xi = 20$. In EBGW algorithm, we select $L = 21$. At this time, each protein sequence in the training dataset generates 1014-D feature vectors. Then, the LASSO algorithm is used to remove redundant information to determine the best feature vector, and the GTB classifier is used to make a prediction, which is verified by the five-fold cross-validation. The proportions of seven kinds of features among the original feature set and LASSO selected feature set are shown in Fig. 5.

Although dimensionality reduction can effectively remove redundant information in protein sequences, it is hoped that the optimal feature subset selected can improve the prediction performance of the model. We input ALL features vector into GTB classifier. Through the five-fold cross-validation, the overall prediction accuracy, MCC and AUC value are 89.29%, 0.7875 and 0.9506, respectively. The LASSO selected feature set is inputted into the GTB classifier. Through the five-fold cross-validation, the overall prediction accuracy, MCC and AUC value are 92.86%, 0.8631 and 0.9706, respectively. Compared to the unreduced dimension, the overall prediction accuracy and an AUC value of the LASSO selected feature set are increased by 3.57% and 2%, respectively. Therefore, the LASSO algorithm can improve the predictive performance of the model. LASSO is a kind of regularization using $L1$ norm, which helps to reduce the risk of over-fitting and easy to obtain sparse solutions, thus playing a role in feature selection. In constructing a protein S-sulfenylation sites prediction model, the LASSO method reduces noise and eliminates redundant information while retaining important effective features. Therefore, this paper uses the LASSO method to select features while maximizing the optimal feature subset.

As can be seen from Fig. 5, different feature extraction methods play different roles in the construction model. The DC and PsePSSM feature extraction algorithms play a vital role in the construction of protein S-sulfenylation sites prediction models. The percentage of DC and PSAAP in the feature space increased by 2% and 1%, respectively. The contribution rate of the EBGW in the feature space is 5%, and the contribution rate of the PsePSSM in the feature space is 40%. Therefore, the sequence information and evolutionary information of the protein sequence play an

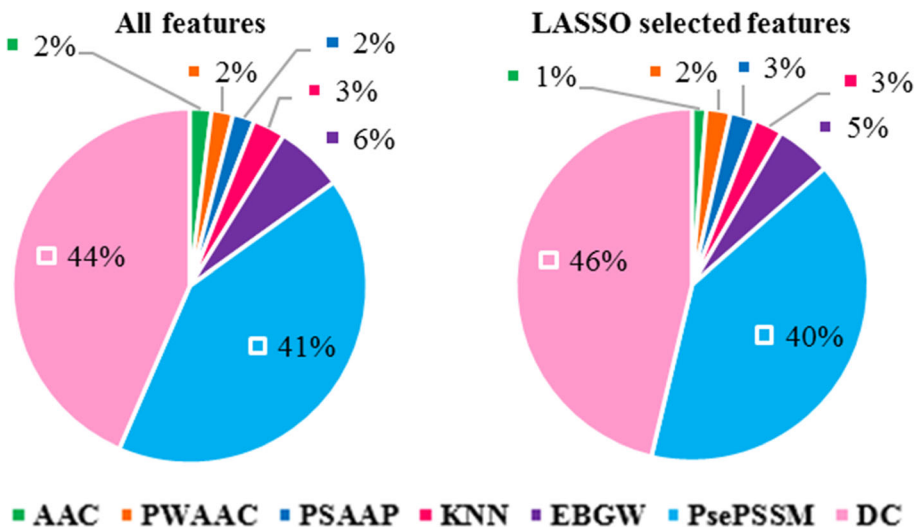

Fig. 5 The proportions of seven kinds of features among the original feature set and LASSO selected feature set. The percentage in the pie chart represents the number and proportion of the seven kinds of features in the original feature set and the LASSO selected feature set

active role in predicting the S-sulfenylation sites of the protein, which helps to improve the predictive performance of the model. Different feature extraction algorithms have different contributions to the prediction of S-sulfenylation sites. The LASSO method performs both automatic variable selection and continuous contraction and selects the relevant variable group.

Considering this, this paper adopts LASSO feature selection algorithm to remove redundant information and select the optimal feature subset ALL (LASSO), which reduces the computational complexity and improves the prediction performance of the model for protein S-sulfenylation sites.

3.6 Selection of classification algorithms

To construct an efficient prediction model, *K* nearest neighbor algorithm (KNN), support vector machine (SVM) [14], decision tree (DT), Naïve Bayes (NB), random forest (RF), AdaBoost [77, 78] and GTB are selected to construct the model. In this paper, we input the reduced feature subset ALL (LASSO) into seven classifiers, respectively. The Euclidean distance is used in *K* nearest neighbor algorithm, and the number of neighboring is 5. The decision tree and Naïve Bayes algorithm use default parameters. The number of iterations for AdaBoost is 20. The support vector machine selects the linear kernel function.

The feature subsets after LASSO dimensionality reduction are input into the *K* nearest neighbor algorithm, decision tree, Naïve Bayes, AdaBoost and GTB, respectively. By five-fold cross-validation, the prediction results are shown in Table 6. To more intuitively analyze the prediction performance of the training datasets on different classifiers, we plot the ACC, MCC, AUC and AUPR values for different classification algorithms, as shown in Fig. 6. In addition, the ROC curve is used to compare the robustness of different prediction models. Figure 7 is the ROC and PR curves on the training set obtained by the seven classification methods.

Table 6 Prediction results of different classification algorithms

Classifier	ACC (%)	Sn (%)	Sp (%)	MCC	AUC	AUPR
KNN	78.95	96.37	61.24	0.6165	0.9203	0.9034
SVM	84.76	81.53	88.04	0.6969	0.9234	0.9026
NB	84.76	79.33	90.28	0.6998	0.9154	0.8813
RF	83.78	81.97	85.61	0.6762	0.9155	0.8947
AdaBoost	88.24	87.79	88.70	0.7649	0.9413	0.9096
DT	89.37	87.79	90.97	0.7878	0.8938	0.8452
GTB	92.86	87.21	98.61	0.8631	0.9706	0.9611

As can be seen from Table 6, for the training dataset GTB algorithm as the model's predictive classification algorithm, the model's ACC, Sn, Sp, MCC, AUC and AUPR reach 92.86%, 87.21%, 98.61%, 0.8631, 0.9706 and 0.9611, respectively. The prediction ACC, Sp, MCC, AUC and AUPR values of the model are higher than other classification algorithms. Using KNN algorithm, the prediction accuracy of the model is 78.69% and 13.91% lower than the GTB algorithm. The overall prediction accuracy of DT, AdaBoost, RF, NB and SVM is 89.57%, 88.09%, 84.02%, 84.82% and 84.76%, respectively. The MCC are 0.7920, 0.7618, 0.6808 and 0.7010, respectively. AUC of GTB are 5.03%, 4.72%, 5.52%, 5.51%, 2.93% and 7.68% higher than KNN, SVM, NB, RF, AdaBoost and DT, respectively. The MCC value of the GTB algorithm is 7.53%–25.16% higher than the other six classification algorithms. The AUPR value of GTB classification algorithm are 5.77%, 5.85%, 7.98%, 6.64%, 5.15% and 11.59% higher than KNN, SVM, NB, RF, AdaBoost and DT, respectively.

It can be seen intuitively from Fig. 6 that the training dataset is related to the changes of ACC, MCC, AUC and AUPR of seven classifiers such as KNN, SVM, NB, RF, AdaBoost, DT and GTB. ACC ranges from 78% to 92%, MCC values from 0.6 to 0.8, AUC values from 0.89 to 0.97, and AUPR values from 0.84 to 0.96. Both the AdaBoost and GTB classifiers achieve good AUC values for the training dataset. As the classification algorithm of DT and GTB, the prediction accuracy and the MCC value increases sequentially. The ACC, MCC, AUC and AUPR value of the SVM is lower than GTB. From the evaluation index ACC, MCC, AUC and AUPR value, the GTB classifier has the highest prediction results. Therefore, the model built with GTB as the classifier has the best predictive performance and the best robustness.

The ROC curve and PR curve are selected to compare the prediction performance of different classifiers. If the ROC and PR curves of one classifier are completely covered by the curves of another classifier, the prediction performance of the latter is better than that of the former. As can be seen from Fig. 7a, for the training dataset, the ROC curve of GTB completely contains the ROC curve corresponding to classifiers KNN, SVM, NB, RF, AdaBoost and DT, and AUC values of GTB are 5.03% and 7.68% higher than KNN and DT, respectively. As can be seen from Fig. 7b, the PR curve of GTB completely contains the corresponding PR curve of classifiers KNN, SVM, NB, RF, AdaBoost and DT, and its AUPR value is 0.9611. The AUPR values of KNN, SVM, NB, RF, AdaBoost and DT are 0.9034, 0.9026, 0.8813, 0.8947, 0.9096 and 0.8452, respectively. The AUPR values of GTB classification algorithm are 5.77%, 5.85%, 7.98%, 6.64%, 5.15% and 11.59% higher than KNN, SVM, NB, RF, AdaBoost and

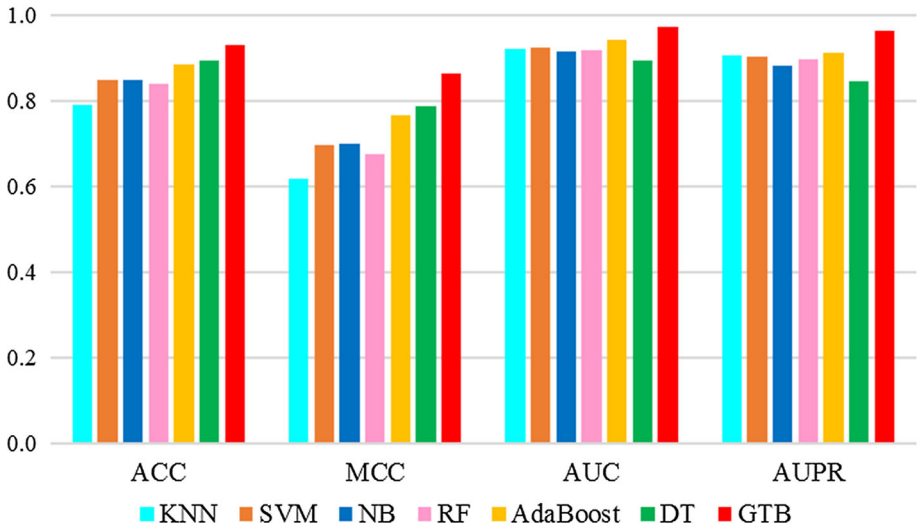

Fig. 6 ACC, MCC, AUC and AUPR values for different classification algorithms

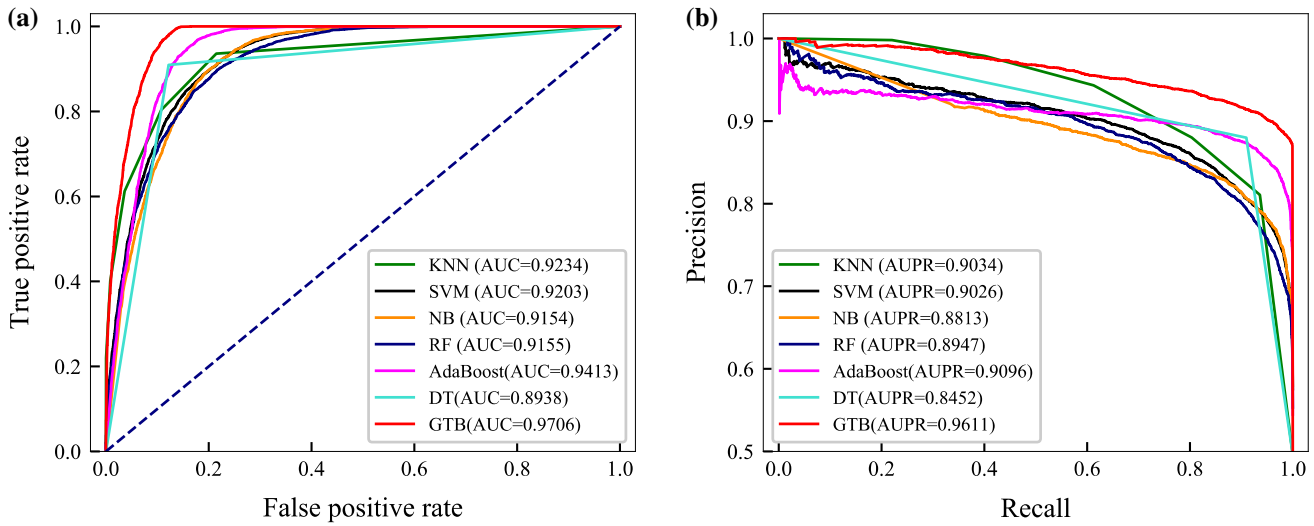

Fig. 7 Prediction results of different classification algorithms on the training dataset. **a** The graph is the ROC curve, and **b** the graph is the PR curve

DT, respectively. In conclusion, the ROC and PR curves of GTB cover the largest area, indicating that the classification algorithm has the best prediction performance and robustness.

By comparing the ACC, AUC, MCC value and ROC and PR curves of the seven classification algorithms on the training dataset, the classification algorithm with higher generalization ability is selected to construct the prediction model. The KNN algorithm is simple and effective, but the calculation amount is relatively large, and the prediction performance of the protein S-sulfenylation sites is poor. The estimated parameters required by the NB algorithm are few, but the independence assumption between attributes is often not established, and there is still room for improvement. The GTB, DT, AdaBoost, and RF are all tree classification algorithms, but there is still a certain gap in predicting performance as a classifier. When GTB trains each regression tree, it predicts the residual of the predicted training data of the previous tree. With such an iteration, the prediction performance is better than the other six classification algorithms. Considering comprehensively, we choose the GTB classification algorithm as the classification algorithm of the model.

3.7 Evaluation of S-sulfenylation sites prediction model using independent test sets

To further test the validity of the prediction model proposed by us, the independent test set constructed by Bui et al. [36] is adopted as the test set, with 216 positive samples and 1418 negative samples. We first combine a variety of feature coding information, including AAC, DC,

EBGW, KNN, PSAAP, PsePSSM and PWAAC, to obtain the feature vector space. According to parameter optimization analysis, the parameters of EBGW and PsePSSM are $L = 21$ and $\xi = 20$, respectively.

To further investigate the composition of the amino acids of residues around the S-sulfenylation sites and the non-S-sulfenylation sites, we use the Two Sample Logos [75] web server (http://www.twosamplelogo.org/cgi-bin/tsl/tsl.cgi) to acquire a comparison of the double sequence identifiers of the independent test set data. The results are shown in Fig. 8.

As can be seen from Fig. 8, there is a significant difference in the composition of residues near the S-sulfenylated and non-S-sulfenylated cysteine in the independent test dataset. The positively charged residue arginine (*R*) appears more frequently at positions − 2, + 4, + 7 and + 10, and the negatively charged residue glutamic acid (E) tends to enrich at position − 4, + 3 and + 5. Lysine (*K*) is significantly more at positions − 6, − 3 and + 6. Arginine acid (*R*) also plays a very important role at downstream of the positive dataset. In the negative dataset, asparagine (*N*) is abundant at positions − 6, − 2 and + 5. Positive data are concentrated at positions − 10, − 9, − 7, + 1, + 8 and + 9 with no significant amino acids. The negative data are concentrated at positions − 10, − 8, − 7, − 4, − 1, + 4, + 6, + 7, + 9 and + 10 with no significant amino acids. Based on these characteristics, it can be inferred that there is a significant difference between the amino acids in the vicinity of the S-sulfenylation and non-S-sulfenylation sites of the independent test dataset.

3.8 Comparison with other methods

For the prediction of protein S-sulfenylation sites, various prediction methods have been proposed. To prove the validity of the proposed model SulSite-GTB, we compare the prediction results of SulSite-GTB with the prediction results of the same dataset. The comparison results of the three methods of the training dataset are shown in Table 7. The comparison results of the seven methods of the independent test set are shown in Table 8. To more intuitively analyze the prediction performance of independent test sets on different prediction models, a histogram of the ACC, Sn, Sp and MCC value is plotted, as shown in Fig. 9.

Table 7 Comparison with other methods on the training dataset

Methods	ACC (%)	Sn (%)	Sp (%)	MCC	AUC
MDD-SOH [36]	69.00	68.00	69.00	0.25	N/A
SulCysSite [41]	79.21	62.90	81.30	0.449	N/A
SulSite-GTB	92.86	87.21	98.61	0.8631	0.9706

N/A not available

Table 8 Comparison of SulSite-GTB with other methods on independent test set

Methods	ACC (%)	Sn (%)	Sp (%)	MCC	AUC
SOHSite [37]	71.00	71.00	71.00	0.2780	N/A
iSulf-Cys [38]	72.41	71.36	72.57	0.3144	N/A
MDD-SOH [36]	71.13	72.00	71.00	0.3000	N/A
SOHPRED [40]	71.45	73.10	71.20	0.3150	N/A
SulCysSite [41]	71.83	76.60	71.10	0.3430	N/A
Sulf_FSVM [43]	79.71	79.81	79.69	0.4461	N/A
SulSite-GTB	88.53	84.10	93.02	0.7740	0.9425

N/A not available

As can be seen from Table 7, the ACC, Sn, Sp, MCC and AUC values of our prediction model SulSite-GTB are 92.86%, 87.21%, 98.61%, 0.8631 and 0.9706, respectively. The ACC, Sn, Sp and MCC value of SulSite-GTB are 13.65%, 24.31%, 17.31% and 41.41% higher than SulCysSite, respectively. The ACC, Sn, Sp and MCC value of SulSite-GTB are 23.86%, 19.21%, 29.61% and 61.31% higher than MDD-SOH, respectively. The ACC value increased from 69.00% to 92.86%, the Sn value increased from 62.90% to 87.21%, the Sp value increased from 69.00% to 98.61%, and the MCC value increased from 0.25 to 0.8631.

From Table 8, for the independent test set the ACC, Sn, Sp, MCC and AUC of SulSite-GTB model are 88.53%, 84.10%, 93.02%, 0.774 and 0.9425, respectively. The ACC value of SulSite-GTB are 17.53%, 16.12%, 17.40%,

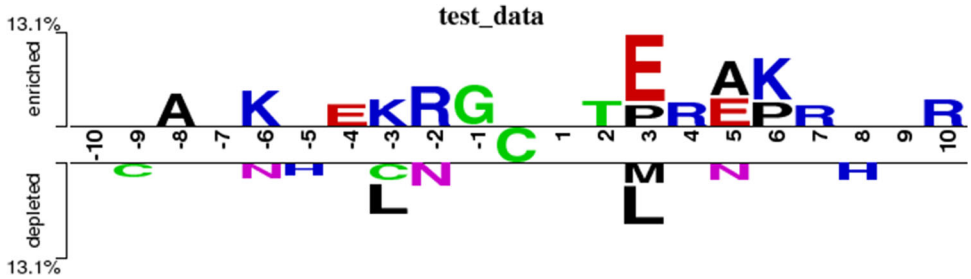

Fig. 8 Comparison of sequence identification of S-sulfenylation and non-S-sulfenylation in the independent test dataset

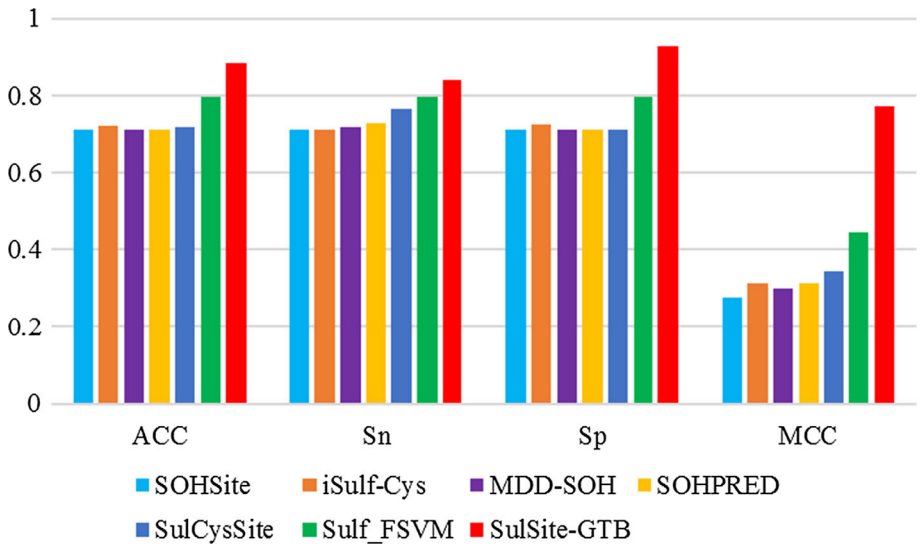

Fig. 9 Comparison of SulSite-GTB with other methods on independent test set

17.08%, 16.70% and 8.82% higher than SOHSite, iSulf-Cys, MDD-SOH, SOHPRED, SulCysSite and Sulf_FSVM, respectively. The MCC values of SulSite-GTB are 32.79%, 43.10%, 45.90%, 47.40%, 45.96% and 49.60% higher than Sulf_FSVM, SulCysSite, SOHPRED, MDD-SOH, iSulf-Cys and SOHSite, respectively. The model SulSite-GTB is 4.29%, 13.33% and 32.79% higher than the Sn, Sp and MCC of the Sulf_FSVM model, respectively. The Sp value of the model SulSite-GTB increased by 13.33–22.02%. Therefore, the model SulSite-GTB has better prediction performance.

It can be seen intuitively from Fig. 9 that for the independent test set, they propose a method for predicting protein S-sulfenylation sites. The ACC, Sn, Sp and MCC values of these methods vary from 0.3 to 0.8. However, the ACC, Sn and Sp value of the prediction model SulSite-GTB established in this paper are all higher than 0.8, and the MCC value is significantly better than the other six methods, which substantially improves the prediction result of the protein S-sulfenylation sites. In summary, the above results fully demonstrate that the predicted model SulSite-GTB can substantially improve the prediction accuracy of protein S-sulfenylation sites, and the prediction results are satisfactory.

4 Discussion

In this study, we propose a new method for predicting protein S-sulfenylation sites, called SulSite-GTB. The overall prediction accuracy of the training set and the independent test set is 92.86% and 88.53%, respectively, and the AUC values are 0.9706 and 0.9425, respectively, and the AUPR values are 0.9611 and 0.9212, respectively.

SulSite-GTB demonstrated good performance on predicting protein S-sulfenylation sites, which is better than the state-of-the-art methods. It is mainly due to the following reasons:

1. AAC, DC, EBGW, KNN, PSAAP, PWAAC and PsePSSM are employed to extract the feature information. AAC extracts the frequency information of amino acids in the protein sequence. DC extracts the frequency information of the amino acid pairs. EBGW extracts physicochemical information of the protein sequence. PsePSSM extracts evolutionary information of protein sequences. KNN extracts clustering information of protein sequences. PWAAC extracts sequence order information around S-sulfenylation sites of proteins. By analyzing the optimal feature subset, we found that the sequence information DC, physicochemical properties EBGW and evolutionary information PsePSSM have a more significant impact on the prediction of protein S-sulfenylation sites, which sufficiently indicates multiple information fusion is of great importance.
2. The SMOTE algorithm is used to process imbalance data, and the LASSO algorithm is applied to remove the feature redundant information to obtain the optimal feature subset. SMOTE is to generate new positive samples that do not exist in the original dataset. The algorithm can balance the sample dataset and make full use of the effective data, which can avoid the model over-fitting and improve the prediction performance of the model. The LASSO algorithm integrates the feature process and the learner training process by introducing $L1$-regularization, eliminating redundant and noise information, and effectively retaining important feature information for identifying S-sulfenylation sites.

3. The optimal feature subset is inputted into the GTB classifier to predict the S-sulfenylation sites. GTB differs from other ensemble learning algorithms in each iteration by predicting the residual of the predicted training data of the previous tree, which is equivalent to weighting the sample of the faulted multiple classifications, thus having good prediction performance, generalization and robustness.

Protein S-sulfenylation sites play an important role in life processes, participates in a variety of cellular activities and is closely related to many human diseases, such as cancer, diabetes and arteriosclerosis. Accurate identification of S-sulfenylation sites is the basis for understanding the structure and function of S-sulfenylation proteins. The SulSite-GTB model can identify cysteine S-sulfenylation sites, making up for the costly and time-consuming defects of laboratory technology, thereby helping biologists understand its molecular function and design of drug targets based on its impact on human diseases.

5 Conclusion

With the growing development of biochip technology and high-throughput sequencing technology, biomedicine has entered the era of big data. Post-translational modification of proteins regulates many critical biological processes in cells and plays a very important role in life. S-sulfenylation is involved in a variety of biological processes including cell signaling, stress response, protein function and signaling. In the face of a large number of protein post-translational modification sites data accumulation, traditional experimental methods cannot meet the needs of modern research, so how to use machine learning methods to predict protein post-translational modification sites has become a hot spot in the field of bioinformatics. In this paper, a new method for protein S-sulfenylation sites prediction, SulSite-GTB, is proposed. First, AAC, DC, EBGW, KNN, PSAAP, PWAAC and PsePSSM are employed to extract the feature information. Secondly, the SMOTE algorithm is used to process imbalance data, and the LASSO algorithm is applied to remove the feature redundant information to obtain the optimal feature subset. Finally, the optimal feature subset is inputted into the GTB classifier to predict the S-sulfenylation sites. The predictive performance of the model is evaluated by the five-fold cross-validation and independent test set methods. On the independent test set, the accuracy is 88.53%, which exceeds 8.82–17.53% of other methods. The results show that the SulSite-GTB method has good performance compared with other methods using the same benchmark dataset. Although the model SulSite-GTB we constructed can effectively improve the prediction accuracy of protein S-sulfenylation sites, there is still room for improvement. Gradient boosting tree classification has high accuracy and robustness to outliers, but it is difficult to train data in parallel due to the dependencies between learners. Zhang et al. [78] proposed a new method of web page classification using Bayesian classifier based on textual and visual contents. The fusion algorithm combines the results of text classifier and image classifier and achieves satisfactory results in the large-scale dataset collected from real fishing cases. Inspired by this, in the next work, we can apply the fusion algorithm to the prediction of post-translation modification sites of proteins, thereby improving the prediction accuracy. Meanwhile, we will systematically study the imbalance of datasets in protein post-translational modification sites, construct a new larger dataset and use deep learning methods to further improve the prediction performance of protein post-translational modification sites.

Acknowledgements This work was supported by the National Nature Science Foundation of China (No. 61863010, 11771188), the Key Research and Development Program of Shandong Province of China (No. 2019GGX101001), and the Natural Science Foundation of Shandong Province of China (No. ZR2018MC007). This work used the Extreme Science and Engineering Discovery Environment, which is supported by the National Science Foundation (No. ACI-1548562).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Matthias M, Jensen ON (2003) Proteomic analysis of post-translational modifications. Nat Biotechnol 21:255–261
2. Wei W, Liu Q, Yi T, Liu L, Li X, Lu C (2009) Oxidative stress, diabetes, and diabetic complications. Hemoglobin 33:370–377
3. Prabhu L, Hartley AV, Martin M, Warsame F, Sun E, Tao L (2015) Role of post-translational modification of the Y box binding protein 1 in human cancers. Genes Dis 2:240–246
4. Paulsen CE, Carroll KS (2013) Cysteine-mediated redox signaling: chemistry, biology, and tools for discovery. Chem Rev 113:4633–4679
5. Paulsen CE, Truong TH, Garcia FJ, Homann A, Gupta V, Leonard SE, Carroll KS (2012) Peroxide-dependent sulfenylation of the EGFR catalytic site enhances kinase activity. Nat Chem Biol 8:57–64
6. Yang J, Gupta V, Carroll KS, Liebler DC (2014) Site-specific mapping and quantification of protein S-sulphenylation in cells. Nat Commun 5:4776
7. Leonard SE, Carroll KS (2011) Chemical 'omics' approaches to understanding protein cysteine oxidation in biology. Curr Opin Chem Biol 15:88–102
8. Poole LB, Nelson KJ (2008) Discovering mechanisms of signaling-mediated cysteine oxidation. Curr Opin Chem Biol 12:18–24

9. Revati W, Jiang Q, Leimiao Y, Erika BS, Bruce K, Poole LB, Eunok P, Tsang AW, Furdui CM (2011) Isoform-specific regulation of Akt by PDGF-induced reactive oxygen species. Proc Natl Acad Sci 108:10550–10555
10. Goedele R, Joris M (2011) Protein sulfenic acid formation: from cellular damage to redox regulation. Free Radic Biol Med 51:314–326
11. Leonard SE, Reddie KG, Carroll KS (2009) Mining the thiol proteome for sulfenic acid modifications reveals new targets for oxidation in cells. ACS Chem Biol 4:783–799
12. Chen Z, Liu XH, Li FY, Li C, Marquez-Lago T, Leier A, Akutsu T, Webb GI, Xu DK, Smith AI, Li L, Chou KC, Song JN (2018) Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. Brief Bioinform 20:2267–2290
13. Weng SL, Kao HJ, Huang CH, Lee TY (2017) MDD-Palm: identification of protein S-palmitoylation sites with substrate motifs based on maximal dependence decomposition. PLoS ONE 12:e0179529
14. Cui XW, Yu ZM, Yu B, Wang MH, Tian BG, Ma Q (2019) UbiSitePred: a novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components. Chemometr Intell Lab Syst 184:28–43
15. Chen YJ, Lu CT, Huang KY, Wu HY, Chen YJ, Lee TY (2015) GSHSite: exploiting an iteratively statistical method to identify S-glutathionylation sites with substrate specificity. PLoS ONE 10:e0118752
16. Xie YB, Luo X, Li Y, Chen L, Ma W, Huang J, Cui J, Zhao Y, Xue Y, Zuo Z (2018) DeepNitro: prediction of protein nitration and nitrosylation sites by deep learning. Genom Proteom Bioinform 16:294–306
17. Wuyun Q, Zheng W, Zhang Y, Ruan J, Hu G (2016) Improved species-specific lysine acetylation site prediction based on a large variety of features set. PLoS ONE 11:e0155370
18. Cai Y, Hu L, Shi X, Xie L, Li Y (2012) Prediction of lysine ubiquitination with mRMR feature selection and analysis. Amino Acids 42:1387–1395
19. Wen PP, Shi SP, Xu HD, Wang LN, Qiu JD (2016) Accurate in silico prediction of species-specific methylation sites based on information gain feature optimization. Bioinformatics 32:3107–3115
20. Zhao XW, Zhao XS, Bao LL, Zhang YG, Dai JY, Yin MH (2017) Glypre: in silico prediction of protein glycation sites by fusing multiple features and support vector machine. Molecules 22:1891
21. Yu JL, Shi SP, Zhang F, Chen GD, Cao M (2019) PredGly: predicting lysine glycation sites for *Homo sapiens* based on XGboost feature optimization. Bioinformatics 35:2749–2756
22. Ning Q, Zhao X, Bao L, Ma Z, Zhao X (2018) Detecting succinylation sites from protein sequences using ensemble support vector machine. BMC Bioinform 19:237
23. Zuo Y, Jia CZ (2017) CarSite: identify carbonylated sites of human proteins based on a one-sided selection resampling method. Mol Biosyst 13:2362–2369
24. Hu J, He X, Yu DJ, Yang XB, Yang JY, Shen HB (2014) A new supervised over-sampling algorithm with application to protein-nucleotide binding residue prediction. PLoS ONE 9:e107676
25. Jia CZ, Zuo Y (2017) S-SulfPred: a sensitive predictor to capture S-sulfenylation sites based on a resampling one-sided selection undersampling-synthetic minority oversampling technique. J Theor Biol 422:84–89
26. Johansen MB, Kiemer L, Brunak S (2006) Analysis and prediction of mammalian protein glycation. Glycobiology 16:844–853
27. Khan YD, Rasool N, Hussain W, Khan SA, Chou KC (2018) iPhosY-PseAAC: identify phosphotyrosine sites by incorporating sequence statistical moments into PseAAC. Mol Biol Rep 45:2501–2509
28. Hou T, Zheng GY, Zhang PY, Jia J, Li J, Xie L, Wei CC, Li YX (2014) LAceP: lysine acetylation site prediction using logistic regression classifiers. PLoS One 9:e89575
29. Li FY, Li C, Marquez-Lago TT, Leier A, Akutsu T, Purcell AW, Smith AI, Lithgow T, Daly RJ, Song J (2018) Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. Bioinformatics 34:4223–4231
30. Li Y, Wang M, Wang H, Tan H, Zhang Z, Webb GI, Song J (2014) Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features. Sci Rep 4:5765
31. Qiu WR, Sun BQ, Tang H, Huang J, Lin H (2017) Identify and analysis crotonylation sites in histone by using support vector machines. Artif Intell Med 83:75–81
32. Qiu WR, Sun BQ, Xiao X, Xu ZC, Jia JH, Chou KC (2017) iKcr-PseEns: identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics 110:239–246
33. Wei L, Xing P, Shi G, Ji ZL, Zou Q (2017) Fast prediction of protein methylation sites using a sequence-based feature selection technique. IEEE/ACM Trans Comput Biol Bioinform 16:1264–1273
34. Luo FL, Wang MH, Liu Y, Zhao XM, Li A (2019) DeepPhos: prediction of protein phosphorylation sites with deep learning. Bioinformatics 33:2766–2773
35. He F, Wang R, Li J, Bao L, Xu D, Zhao X (2018) Large-scale prediction of protein ubiquitination sites using a multimodal deep architecture. BMC Syst Biol 12:109
36. Bui VM, Lu CT, Ho TT, Lee TY (2015) MDD-SOH: exploiting maximal dependence decomposition to identify S-sulfenylation sites with substrate motifs. Bioinformatics 32:165–172
37. Bui VM, Weng SL, Lu CT, Chang TH, Weng TY, Lee TY (2016) SOHSite: incorporating evolutionary information and physicochemical properties to identify protein S-sulfenylation sites. BMC Genom 17:9
38. Xu Y, Ding J, Wu LY (2016) iSulf-Cys: prediction of S-sulfenylation sites in proteins with physicochemical properties of amino acids. PLoS One 11:e0154237
39. Sakka M, Tzortzis G, Mantzaris MD, Bekas N, Kellici TF, Likas A, Galaris D, Gerothanassis IP, Tzakos AG (2016) PRESS: protein S-sulfenylation server. Bioinformatics 32:2710–2712
40. Wang XF, Yan RX, Li JY, Song J (2016) SOHPRED: a new bioinformatics tool for the characterization and prediction of human S-sulfenylation sites. Mol Biosyst 12:2849–2858
41. Hasan MM, Guo D, Kurata H (2017) Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information. Mol Biosyst 13:2545–2550
42. Deng L, Xu XJ, Liu H (2018) PredCSO: an ensemble method for prediction of S-sulfenylation sites in proteins. Mol Omics 14:257–265
43. Ju Z, Wang SY (2018) Prediction of S-sulfenylation sites using mRMR feature selection and fuzzy support vector machine algorithm. J Theor Biol 457:6–13
44. Wang L, Zhang R, Mu Y (2019) Fu-SulfPred: identification of Protein S-sulfenylation Sites by Fusing Forests via Chou's General PseAAC. J Theor Biol 461:51–58
45. Sun MA, Wang Y, Cheng H, Zhang Q, Ge W, Guo D (2012) RedoxDB-a curated database for experimentally verified protein oxidative modification. Bioinformatics 28:2551–2552
46. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659

47. Du XQ, Sun SW, Hu CJ, Yao Y, Yan YT, Zhang YP (2017) DeepPPI: boosting prediction of protein-protein interactions with deep neural networks. J Chem Inf Model 57:1499–1510
48. Manoj B, Raghava GPS (2004) Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem 279:23262–23266
49. Khan A, Majid A, Hayat M (2011) CE-PLoc: an ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition. Comput Biol Chem 35:218–229
50. Zhang ZH, Wang ZH, Zhang ZR, Wang YX (2006) A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS Lett 580:6169–6174
51. Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, Webb GI, Smith AI, Daly RJ, Chou KC (2018) iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 34:2499–2502
52. Tang YR, Chen YZ, Canchaya CA, Zhang ZD (2007) GANN-Phos: a new phosphorylation site predictor based on a genetic algorithm integrated neural network. Protein Eng Des Sel 20:405–412
53. Jones D (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202
54. Yu B, Li S, Qiu WY, Chen C, Chen RX, Wang L, Wang MH, Zhang Y (2017) Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising. Oncotarget 8:107640–107665
55. Yu B, Li S, Qiu WY, Wang MH, Du JW, Zhang YS, Chen X (2018) Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction. BMC Genom 19:478
56. Qiu WY, Li S, Cui XW, Yu ZM, Wang MH, Du JW, Peng YJ, Yu B (2018) Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition. J Theor Biol 450:86–103
57. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
58. Liu TG, Geng XB, Zheng XQ, Li RS, Wang J (2012) Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. Amino Acids 42:2243–2249
59. Shen HB, Chou KC (2007) Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng Des Sel 20:561–567
60. Huang SY, Shi SP, Qiu JD, Liu MC (2015) Using support vector machines to identify protein phosphorylation sites in viruses. J Mol Graph Model 56:84–90
61. Shi SP, Qiu JD, Sun XY, Suo SB, Huang SY, Liang RP (2012) PMeS: prediction of methylation sites based on enhanced feature encoding scheme. PLoS One 7:e38772
62. Shi SP, Qiu JD, Sun XY, Suo SB, Huang SY, Liang RP (2012) A method to distinguish between lysine acetylation and lysine methylation from protein sequences. J Theor Biol 310:223–230
63. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
64. Wang XY, Yu B, Ma AJ, Chen C, Liu BQ, Ma Q (2019) Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. Bioinformatics 35:2395–2402
65. Shi H, Liu SM, Chen JQ, Li X, Ma Q, Yu B (2019) Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. Genomics 111:1839–1852
66. Yu B, Qiu WY, Chen C, Ma AJ, Jiang J, Zhou HY, Ma Q (2020) SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. Bioinformatics 36:1074–1081
67. Kang CZ, Huo YH, Xin LH, Tian BG, Yu B (2019) Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. J Theor Biol 463:77–91
68. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc B 58:267–288
69. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232
70. Liu Y, Gu Y, Nguyen JC, Li H, Zhang J, Gao Y, Huang Y (2017) Symptom severity classification with gradient tree boosting. J Biomed Inform 75:105–111
71. Pan Y, Liu D, Deng L (2017) Accurate prediction of functional effects for variants by combining gradient tree boosting with optimal neighborhood properties. PLoS One 12:e0179314
72. Fan C, Liu D, Huang R, Chen Z, Deng L (2016) PredRSA: a gradient boosted regression trees approach for predicting protein solvent accessibility. BMC Bioinform 17:8
73. Yu B, Li S, Chen C, Xu JM, Qiu WY, Wu X, Chen RX (2017) Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition. Chemometr Intell Lab 167:102–112
74. Chen C, Zhang QM, Ma Q, Yu B (2019) LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion. Chemometr Intell Lab Syst 191:54–64
75. Vladimir V, Iakoucheva LM, Predrag R (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics 22:1536–1537
76. Yu B, Lou LF, Li S, Zhang YS, Qiu WY, Wu X, Wang MH, Tian BG (2017) Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising. J Mol Graph Model 76:260–273
77. Zhu J, Zou H, Rosset S, Hastie T (2006) Multi-class adaboost. Stat Interface 2:349–360
78. Zhang H, Liu G, Chow TW, Liu W (2011) Textual and visual content-based anti-phishing: a Bayesian approach. IEEE Trans Neural Netw 22:1532–1546