IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 22, 2020, accepted May 3, 2020, date of publication May 6, 2020, date of current version May 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2992672

PCSPred_SC: Prediction of Protein Citrullination Sites Using an Effective Sequence-Based Combined Method

LINA ZHANG[1], JINGUI CHEN[1], CHENGJIN ZHANG[1], RUI GAO[2], AND RUNTAO YANG[1]
[1]School of Mechanical, Electrical and Information Engineering, Shandong University at Weihai, Weihai 264209, China
[2]School of Control Science and Engineering, Shandong University, Jinan 250100, China

Corresponding author: Runtao Yang (yrt@sdu.edu.cn)

This work was supported in part by the China Postdoctoral Science Foundation under Grant 2018M630778, and in part by the National Natural Science Foundation of China under Grant U1806202, Grant 61533011, Grant 61573213, Grant 61673245, and Grant 61603214.

ABSTRACT As one of post-translational modifications (PTMs), protein citrullination is crucial in a diverse array of cellular processes and implicated in a slew of human pathology. Therefore, accurate identification of protein citrullination sites (PCSs) is urgently needed to illuminate the reaction details and the complex pathogenesis related to the protein citrullination. In view of the limitations of the existing PCS predictors, this study proposes a novel and powerful sequence-based combined method named PCSPred_SC to further enhance the prediction performance. Various feature extraction methods are developed to mine sequence-derived biological information. Under the feature space, the predictive capabilities of different prediction algorithms, over-sampling methods, and feature selection methods are respectively explored. Experimental results indicate that the over-sampling methods are effective to solve the imbalanced dataset problem and the feature selection methods are significant in removing irrelevant and redundant features. On the same dataset using 10-fold cross validation, PCSPred_SC constructed by the combination of support vector machine (SVM), Adasyn, and t-distributed stochastic neighbor embedding (t-SNE) achieves much more outstanding performance than the competing methods, while reducing the number of features used for this task remarkably. It is anticipated that the proposed method will provide significant information to broaden our knowledge of citrullination-related biological processes.

INDEX TERMS Citrullination, prediction algorithm, over-sampling, feature selection.

I. INTRODUCTION

Post-translational modifications (PTMs) can increase the diversity of protein functions to maintain physiological homeostasis [1]. As one of critical PTMs, protein citrullination illustrated in Figure 1 is a hydrolytic reaction converting positively charged arginine into neutrally charged citrulline [2]. Mediated by the calcium-dependent peptidyl arginine deiminases (PADs) [3], citrullination can alter total charge and hydrogen bonding with consequent effects on the target protein's molecular conformation, biochemical activities, immunogenicity, and interactions with proteins or nucleic acids [4].

The existing PAD isozymes (PAD 1-4 and 6) exhibit a tissue specific expression [5]. Under normal circumstances, PAD1 and PAD3 are mainly expressed in the skin and hair follicles. They participate in the terminal differentiation of keratinocytes by catalyzing the citrullination of pro-filaggrin [6]. PAD2 that may function within the epidermal growth factor signaling pathway to regulate cell

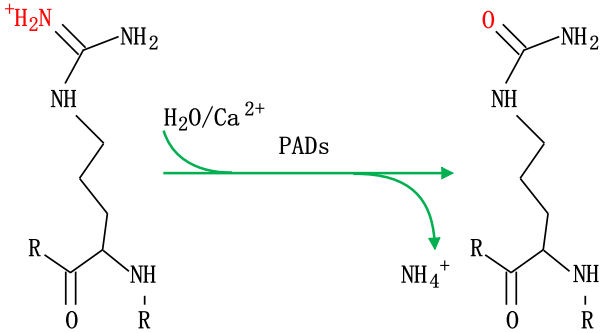


FIGURE 1. The process of protein citrullination mediated by the calcium-dependent PADs.

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

TABLE 1. Known substrates that are targeted by the individual isozymes of the PAD family of isozymes [10].

Isozyme	Substrates
PAD1	Keratin, filaggrin
PAD2	Myelin basic protein, vimentin, actin, histones
PAD3	Filaggrin, trichohyalin, apoptosis-inducing factor, vimentin
PAD4	Histones, ING4, p300, p21, nucleophosmin, nuclear lamin C
PAD6	None known

migration is principally distributed in skeletal muscle, brain, pancreas and spleen [7]. PAD4 involved in gene expression and protein localization can be primarily detected in neutrophils and other myeloid derived cells [8]. PAD6 closely related with embryo development is largely located in oocytes and embryonic stem cells [9]. As listed in Table 1, the PAD isozymes also show specificity for their targeted substrates.

At physiological concentrations of calcium, protein citrullination is crucial in a diverse array of cellular processes including proliferation, differentiation, apoptosis, myelinization, neutrophil extracellular trap (NET) formation, gene expression regulation, and skin homeostasis [5], [11]–[13]. For instance, citrullinated keratin and histones have important effects on skin protection and gene regulation [14]; Citrullinated fibronectin can regulate the function of synovial fibroblasts [15]; Citrullinated vimentin and its antibody can induce osteoclast differentiation and subsequent bone resorption [16]; Citrullinated calreticulin on the cell surface can enhance its role in signaling pathways [17]. Although T. Goulas *et al.* shed light on a general regulatory mechanism of citrullination [18], the exact factors that lead to citrullination in vivo remain largely elusive [13]. To illuminate the reaction details of citrullination, development of novel strategies to comprehensively identify protein citrullination sites (PCSs) is urgently needed.

Recently, accumulating evidence has indicated that dysregulations of PADs in citrullination are involved in a slew of human pathology [10], [13], [19]–[21]. As reported, abnormally elevated protein citrullination followed by the production of anti-citrullinated protein antibodies (ACPAs) was detected in patients with rheumatoid arthritis (RA) [22]. Citrullinated histone H3, a biomarker of NET formation, is independently associated with the occurrence of venous thromboembolism in cancer patients [23]. PAD1-mediated citrullination is positively correlated with human triple negative breast cancer [24]. Overexpressing PAD2 has been implicated in the onset and progression of human malignant cancers [12], [20], [25], [26], whereas downregulation of PAD2 is observed in the pathogenesis of colorectal cancer [27]. Additionally, PAD4 has been linked to a wide range of inflammatory autoimmune diseases including arthritis, colitis and multiple sclerosis [28], [29]. Given the strong evidence linking dysregulated citrullination to human diseases, autoantibodies targeting citrullinated proteins have been used as promising diagnostic markers [5], [12], [28]. However, the pathological roles that PAD-mediated citrullination play in these diseases are still to be discerned [10]. Given this background, accurate identification of PCSs is required to broaden our knowledge of PAD's substrate specificity and clarify the critical effect of citrullination on substrate's functions, which will ultimately have diagnostic or prognostic value in diverse citrullination related diseases [30].

At present, a series of experimental methods have been developed to detect PCSs [31]. S.M. Hensen *et al.* proposed a robust and sensitive antibody-independent strategy to visualize the modified citrullines through western blot analysis [32]. Using the ionization characteristics of citrulline residues, detection of citrulline by mass spectrometry (MS) is a widely adopted technique. However, the abundance of citrulline peptides is too low to produce high-quality MS/MS fingerprints, and related non-citrulline fragments are easily deleted [33]. Furthermore, as citrullination results in only 1 dalton change in mass, ion signals of a citrullinated peptide in a MS are always difficult to detect [34]. Therefore, only a handful of PAD substrates are known due to the technical challenges associated with experimental methods [22].

With the advances in sequencing technologies, cost-effective computational methods have been proposed to accelerate the discovery of PCSs. By incorporating multiple sequence information such as amino acid composition, position-specific scoring matrix (PSSM) conservation scores, amino acid factors and disorder scores, Q. Zhang *et al.* employed a random forest classifier together with the maximum Relevance Minimum Redundancy (mRMR)-incremental feature selection (IFS) method to predict PCSs [35]. However, the sensitivity achieved by the predictor is as low as 0.603 due to the unsolved class imbalance problem in the dataset. Stimulated by the pseudo amino acid composition (PseAAC) approach [36], a sequence-based predictor called CKSAAP_CitrSite was proposed to improve the prediction performance by coupling support vector machine (SVM) with the composition of k-spaced amino acid pairs (CKSAAP) selected by F-score [37]. Likewise, CKSAAP_CitrSite did not give a solution to the class imbalance problem. In addition, as the feature extraction strategy is based on a single technique, the intrinsic biological properties of protein citrullination are not fully considered, which may limit the prediction performance of CKSAAP_CitrSite.

The aforementioned methods have made certain contributions to stimulating the development of PCS detection. But there is still room for improvement, particularly in terms of sensitivity. In view of the limitations of the above-mentioned methods, this study proposes a novel and powerful method named PCSPred_SC for identifying PCSs using an effective sequence-based combined method. Firstly, different feature extraction methods including binary encoding (BE), position specific amino acid propensity (PSAAP), pseudo amino acid composition (PseAAC) [36] and physicochemical properties (PP) [38], [39] are adopted to convert peptides into numeric feature vectors. Secondly, under the complete feature space, the PCS predictors are respectively constructed by various prediction algorithms, including naïve bayes (NB), logistic

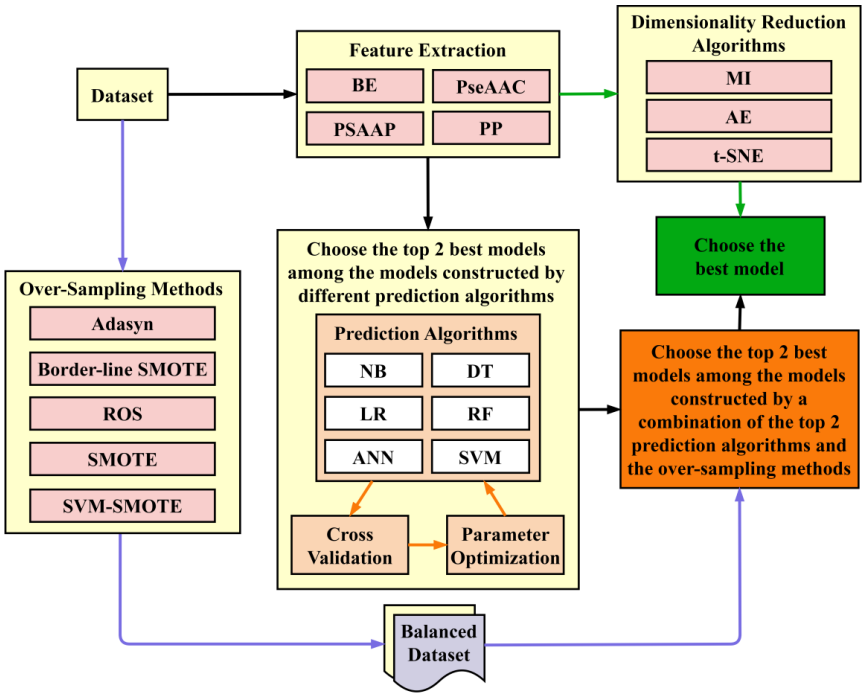


FIGURE 2. The overall workflow of the proposed method.

regression (LR), artificial neural network (ANN), decision tree (DT), random forest (RF), and support vector machine (SVM). Thirdly, the effects of the over-sampling methods including random over-sampling (ROS), synthetic minority over-sampling technique (SMOTE) [40], Border-line SMOTE [41], SVM-SMOTE [42] and Adasyn [43] are systematically explored using the top 2 prediction algorithms that achieve the best performance. Finally, to determine the best prediction model, different feature selection methods including mutual information (MI) [44], autoencoder (AE) [45], and t-distributed stochastic neighbor embedding (t-SNE) [46] are respectively incorporated into the top 2 models constructed by a combination of the prediction algorithm and over-sampling methods. Compared with exiting methods, experimental results demonstrate that the proposed method achieves a superior performance in terms of various performance measures. A summary of the computational framework of our method is displayed in Figure 2.

II. MATERIALS AND METHODS

A. DATASET

To make fair comparisons with previous studies, we use the same trainning dataset introduced by Q. Zhang *et al.* [35]. The dataset was generated by scanning across the protein sequence within a window of size 21 centered at citrullination or non-citrullination sites. If less than 10 upstream or downstream residues flanked the central site, the missing positions would be filled with a dummy residue 'X'. As a result, the dataset included 116 experimentally annotated PCSs and 232 non-annotated PCSs.

To further reliably estimate the predictive ability of the proposed method, we construct an independent test dataset as follows. The citrullinated proteins are collected from Universal Resource of Protein (UniProt, available at https://www.uniprot.org/) by searching the keyword of 'citrulline' in the field of 'Modified residue'. If the number of flanking amino acids is less than 10, the missing positions are expanded with a special residue 'X'. Then, the peptides within a window of size 21 centered at the citrullination site are extracted. Among all the retrieved sequences, only experimentally identified and reviewed citrullination sites are kept. Furthermore, all duplicate samples and the samples included in the training dataset are removed. As a result, a positive dataset including 138 samples with citrullination sites is obtained. We then randomly select 150 non-citrullination sites from the citrullinated proteins to construct a representative negative dataset. The same strict filtering criteria, as mentioned above, are applied to the negative dataset. Thus, the independent test dataset has a total of $138 + 150 = 288$ peptide samples.

B. FEATURE EXTRACTION

For constructing a robust and reliable predictor, it is a crucial step to transform the input sequence into a set of numerical attributes that could really reflect the intrinsic correlation with the desired target [47]. To avoid the bias of using single descriptor, integrating complementary information from different types of protein feature representations has become a new trend of feature design [48], [49]. In this study, we explore four types of quantitative feature

descriptors, including binary encoding (BE), position specific amino acid propensity (PSAAP), pseudo amino acid composition (PseAAC), and physicochemical properties (PP). The detailed feature extraction processes are explained in the following subsections.

1) BINARY ENCODING

20 amino acids plus the aforementioned gap-filling residue "X" are ordered as ACDEFGHIKLMNPQRSTVWYX. According to the alphabetical order, the values of $j = 1, 2, \cdots, 21$ denote different kinds of amino acids. We encode amino acid j at each position using a 21-dimensional binary vector $\{a_1, a_2, \cdots, a_{21}\}$, where $a_j = 1$ and $a_i (i \neq j) = 0$.

2) POSITION SPECIFIC AMINO ACID PROPENSITY

The position specific amino acid propensity (PSAAP) would be employed to measure the amino acid preferences in different positions flanking the known PCSs. Given a peptide P in the dataset, its most straightforward expression is

$$P = R_1, R_2, \cdots, R_{21}, \tag{1}$$

where R_i represents the i-th residue of the peptide P. The detailed procedure of PSAAP is as follows. Firstly, the amino acid compositions of the j-th position for the positive dataset and the negative dataset are respectively calculated and denoted as

$$(A^+_{1,j}, A^+_{2,j}, \cdots, A^+_{21,j}) \quad j = 1, 2, \cdots, 21, \tag{2}$$

$$(A^-_{1,j}, A^-_{2,j}, \cdots, A^-_{21,j}) \quad j = 1, 2, \cdots, 21. \tag{3}$$

Then, a score $z_{i,j} = A^+_{i,j} - A^-_{i,j}$ is computed to indicate the propensity of the i-th amino acid in the j-th position of the peptide centered at PCSs. Finally, a 21 dimensional vector for every peptide can be easily read out from the PSAAP matrix $\mathbb{Z} = (z_{i,j})$, where the vector's i-th element μ_i is denoted as

$$\mu_i = \begin{cases} z_{1,i} & R_i = A \\ z_{2,i} & R_i = C \\ \vdots & \vdots \\ z_{21,i} & R_i = X \end{cases} \quad i = 1, 2, \cdots, 21. \tag{4}$$

3) PSEUDO AMINO ACID COMPOSITION

To avoid losing sequence order information hidden in protein sequences, the pseudo amino acid composition (PseAAC) proposed by KC Chou [36] is introduced to comprehensively incorporate the occurrences and physicochemical properties of amino acids. Ever since then, the concept of PseAAC has been penetrated into various areas of computational proteomics [47], [50], [51].

Considering the peptide given in Equation (1), the sequence-order correlation factor is defined as

$$\theta = \frac{1}{L-1} \sum_{i=1}^{L-1} [M(R_{i+1}) - M(R_i)]^2, \tag{5}$$

where $M(R_i)$ indicates the normalized side-chain mass of the amino acid R_i and can be subjected to a standard conversion as described by the following equation:

$$M(i) = \frac{M^0(i) - \sum_{i=1}^{20} \frac{M^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20} [M^0(i) - \sum_{i=1}^{20} \frac{M^0(i)}{20}]^2}{20}}}, \tag{6}$$

where $M^0(i)$ is the original side-chain mass of the i-th amino acid in alphabetical order.

Then, the peptide given in Equation (1) is represented as

$$V = \{v_1, v_2, \cdots, v_{21}\}, \tag{7}$$

where the components are given by

$$v_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + \omega\theta} & 1 \leq u \leq 20 \\ \dfrac{\omega\theta}{\sum_{i=1}^{20} f_i + \omega\theta} & u = 21, \end{cases} \tag{8}$$

and $f_u (u = 1, 2, \cdots, 20)$ is the normalized occurrence frequency of the 20 amino acids in the peptide sequence P; Without loss of generality, the weight factor ω is set to be 0.05. In this representation, the first 20 descriptors depict the components of its basic amino acid composition and the last descriptor reflects sequence order information.

4) PHYSICOCHEMICAL PROPERTIES

Several studies have indicated that the physicochemical properties (PP) of residues determine its interactions with the others [38], [39]. In this study, 13 physicochemical properties closely related to the behavior of the protein interfaces, including positively charged, negatively charged, neutral charged, polarity, non polarity, hydrophobicity, hydrophilicity, secondary structure (helix), secondary structure (strands), secondary structure (coil), solvent accessibility (buried), solvent accessibility (exposed), and solvent accessibility (intermediate), are extracted from the web server named Pfeature (https://webs.iiitd.edu.in/raghava/pfeature/). Then, the average values of the amino acid's each physicochemical property along peptide samples are calculated.

C. OVER-SAMPLING METHODS

As the dataset indicated in Section II.A, the number of peptide chains without PCSs is twice that of peptide chains with PCSs. In other words, the imbalanced dataset problem exists in the benchmark dataset, which would lead to most of the incoming data labeled as the majority class by traditional machine learning algorithms [59]. In this study, over-sampling methods including random over-sampling (ROS), synthetic minority over-sampling technique (SMOTE), Border-line SMOTE, SVM-SMOTE, and Adasyn are respectively employed to balance the positive and negative training samples. The ROS method replicates randomly selected samples within the minority set; The SMOTE

method generates novel synthetic samples through performing the interpolation algorithm between each minority class sample and its k minority class nearest neighbors [40]; The Border-line SMOTE method only over-samples or focuses on the borderline minority samples by identifying noise samples, danger samples, and safe samples [41]; The SVM-SMOTE method generates artificial minority instances at the boundary of majority class and minority class with SVM trained to predict future instances [42]; The Adasyn method generates more minority class samples that are harder to learn [43].

D. FEATURE SELECTION

Evidently, there always exist noisy, irrelevant, and redundant features in the integrated feature space, which can potentially cause the curse of dimensionality, over fitting, and the increase of the computation complexity [60]. That is to say, not all of these candidate features facilitate the prediction of PCSs. Therefore, mutual information, autoencoder, and t-distributed stochastic neighbor embedding described in detail below are respectively employed to select the informative features.

1) MUTUAL INFORMATION

In a nonlinear context, the mutual information (MI) is widely used as the criterion to measure the amount of information shared between different variables [44]. Suppose the set of the values of the i-th feature F_i and the set of the class labels are respectively denoted as V_i and C, the MI of V_i and C is defined as

$$MI\left(V_i, C\right) = \sum_{c \in C} \sum_{v \in V_i} p\left(v, c\right) \log \frac{p\left(v, c\right)}{p\left(v\right) p\left(c\right)}. \tag{9}$$

From the perspective of information gain, MI represents the amount by which the uncertainty of C is reduced due to the introduction of F_i. Greater MI means that the feature F_i is more beneficial to distinguish the elements in C.

2) AUTOENCODER

Implemented with unsupervised learning, autoencoder (AE) is a derivative of ANNs to reconstruct the input data at its output layer [45]. If the number of the neurons in the hidden layer is fewer than that of the input layer, dimensionality reduction of the original input patterns can be achieved by deriving features from the hidden layer. The AE learns the optimal weights connecting neurons through the backpropagation algorithm.

3) t-Distributed STOCHASTIC NEIGHBOR EMBEDDING

By matching distances between high-dimensional and low-dimensional spaces, t-distributed stochastic neighbor embedding (t-SNE) is a dimensionality reduction algorithm retaining the original clustering [46]. The whole procedure of the t-SNE is given in the following steps. (i) Calculate ''unscaled'' similarity scores between the high-dimensional points using a ''t-distribution'' and then scale them. (ii) Construct the similarity matrix with each element representing the similarity score. (iii) Create an initial set of low-dimensional points. (iv) Iteratively update the low-dimensional points to minimize the Kullback-Leibler divergence.

E. PERFORMANCE MEASURES

To evaluate the prediction performance of PCS predictors, the widely used performance measures including sensitivity (Sn), specificity (Sp), accuracy (Acc), Matthew's correlation coefficient (MCC), and area under the receiver operating characteristic curve (AUC), are calculated. The first 4 performance measures are defined as follows:

$$Sn = \frac{TP}{TP + FN}, \tag{10}$$

$$Sp = \frac{TN}{TN + FP}, \tag{11}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \tag{12}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}, \tag{13}$$

where TP, FP, TN and FN represent the numbers of true positives, false positives, true negatives, and false negatives, respectively. For the imbalanced dataset problem, there is a preference for a low Sn and a high Sp. A high Sp often means a high Acc [38]. Therefore, Acc is not an appropriate measure for the performance evaluation. To achieve a comprehensive and stable performance, the MCC reflecting a trade-off between Sn and Sp is employed as the main measure to construct the PCS predictor and compare it with existing methods.

To further evaluate the performance of our method, the receiver operating characteristics (ROC) curve is plotted with the true positive rate (i.e. Sn) as a function of the false positive rate (i.e. 1-Sp) for varying decision thresholds [61]. The AUC, a reliable measure for the prediction performance, is also calculated.

10-fold cross validation [62] is adopted in this study to calculate the above-mentioned performance evaluation indexes. That is, the benchmark dataset is randomly partitioned into 10 data subsets with approximately equal size. One subset is retained for testing and all others form the training dataset. This process is repeated 10 times to test each subset. Finally, the average performance measures over the 10 folds are calculated as the final result of evaluation.

III. RESULTS AND DISCUSSIONS

A. COMPARISON OF DIFFERENT PREDICTION ALGORITHMS

The classification performance is data sensitive and algorithm dependent. Hence, the effect of different algorithms given in the Section II.C for identifying PCSs is examined using the complete feature space without over-sampling methods. In these experiments, we tune the ideal parameters for each algorithm under the 10-fold cross validation.

TABLE 2. Performance of different prediction algorithms using the complete feature space without over-sampling methods.

Prediction Algorithm	Sn	Sp	Acc	MCC
NB	0.690	0.526	0.580	0.204
LR	0.690	0.841	0.790	0.529
ANN	0.655	0.853	0.787	0.516
DT	0.664	0.784	0.744	0.439
RF	0.509	0.953	0.805	0.542
SVM	0.414	0.987	0.796	0.534

As listed in Tables 2, the Acc achieved by RF is 0.805, which is 0.009-0.225 higher than those achieved by the other five algorithms. The closest competitor of RF in terms of Acc and MCC is the SVM. The Sp obtained by SVM has the largest Sp of 0.987, which is 0.461, 0.146, 0.134, 0.203, and 0.034 higher than that obtained by NB, LR, ANN, DT and RF, respectively. Among these algorithms, RF achieves the best MCC of 0.542, and SVM achieves the second best MCC of 0.534. These results indicate that SVM and RF attain much more outstanding performance for PCS prediction. It's also worth noting that most of the algorithms yields much higher Sp than Sn due to the imbalanced dataset problem. Therefore, SVM and RF are respectively selected as the prediction algorithm to balance the dataset using the over-sampling methods.

B. THE CHOICE OF OVER-SAMPLING METHODS

In previous experiments, the PCS predictors are constructed by the imbalanced dataset given in Section II.A. To alleviate the class imbalance problem, different over-sampling methods to balance the dataset are adopted in this study. The results in Table 3 summarizes the performance of the combinations of RF or SVM with each of the over-sampling method. Specifically, a combination of SVM and ROS (SVM + ROS) achieves the highest Sn; A combination of SVM and Adasyn (SVM + Adasyn) outperforms the other predictors in terms of two critical measures, Acc and MCC with values of 0.934 and 0.850, respectively. These values are 0.02 and 0.046 higher than those obtained by the closest competitor, a combination of SVM and SVM-SMOTE (SVM + SVM-SMOTE). Additionally, the second best prediction performance is achieved by SVM + SVM-SMOTE with a Sn of 0.793, a Sp of 0.974, a Acc of 0.914, and a MCC of 0.804. Therefore, SVM + Adasyn and SVM + SVM-SMOTE as respectively chosen as the basic predictor to conduct the feature selection processes.

TABLE 3. Performance of the combinations of RF or SVM with different over-sampling methods.

Prediction Algorithm	Over-Sampling Method	Sn	Sp	Acc	MCC
RF	Adasyn	0.750	0.884	0.839	0.637
	Border-line SMOTE	0.621	0.888	0.799	0.533
	ROS	0.810	0.892	0.865	0.698
	SMOTE	0.629	0.914	0.819	0.579
	SVM-SMOTE	0.716	0.892	0.833	0.619
SVM	Adasyn	0.853	0.974	0.934	0.850
	Border-line SMOTE	0.741	0.974	0.897	0.765
	ROS	0.879	0.815	0.850	0.696
	SMOTE	0.750	0.974	0.899	0.771
	SVM-SMOTE	0.793	0.974	0.914	0.804

C. ADDED VALUE OF OVER-SAMPLING METHODS

To provide insights in the added value of over-sampling methods, the prediction results without and with over-sampling methods respectively given in Table 2 and Table 3 are compared. Obviously, no matter what the prediction algorithm is, the predictors with over-sampling methods perform significantly better than the variants without over-sampling methods. As listed in Table 3, all the 5 MCCs achieved by SVM combined with over-sampling methods are higher than 0.69 and 4 of them are higher than 0.76, while the MCC achieved by SVM without over-sampling methods is only 0.534. Similar comparison results can be obtained for the RF. In addition, the Sns achieved by RF and SVM without over-sampling methods are less than 0.52, and there is a relatively large gap between Sn and Sp. On the contrary, the Sns achieved by RF and SVM with over-sampling methods are higher than 0.62, while keeping the comparable Sp and Acc. These results highlights the incremental value of the over-sampling methods on enhancing the PCS predictors' reliability and performance.

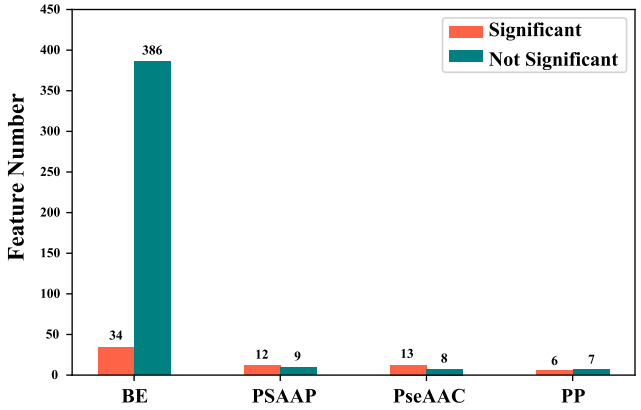


FIGURE 3. The statistical significance test results between the actual peptide chains and the generated peptide chains on the complete feature space.

To further validate the effectiveness of the over-sampling methods, the statistical significance between the actual peptide chains and the generated peptide chains on the complete feature space is assessed by the paired t-test with $\alpha = 0.05$. As shown in Figure 3, for the majority of features, there is no significant difference between the actual peptide chains and the generated peptide chains.

D. PERFORMANCE COMPARISONS OF FEATURE SELECTION METHODS

The feature selection methods employed in this study can be categorized into the filter algorithm (MI) and the projection algorithms (AE and t-SNE). For the filter algorithm, features are ranked according to their weights given by MI. Then, to select the optimal feature subset, MCCs corresponding to varying top-ranking features are calculated. For the projection algorithms, the prediction results of the feature spaces with different dimensions mapped by AE or t-SNE are

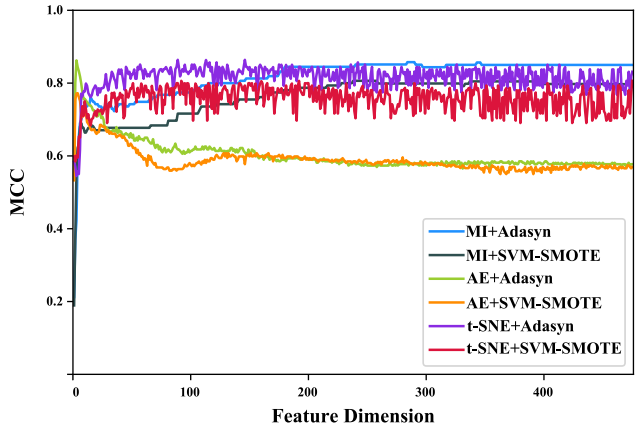


FIGURE 4. The values of MCC against feature subsets selected by different feature selection methods with Adasyn or SVM-SMOTE.

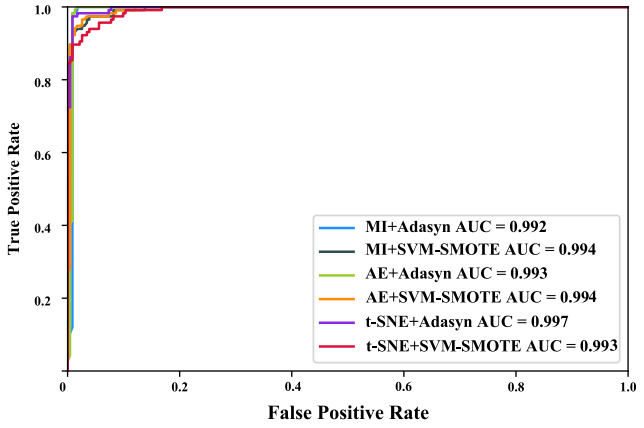


FIGURE 5. ROC curves of different feature selection methods combined with Adasyn or SVM-SMOTE.

evaluated to determine the optimal dimension of feature vector. Taking SVM as the prediction algorithm, Figure 4 illustrates the relations between the MCCs and the feature subsets selected by different feature selection methods with Adasyn or SVM-SMOTE. From the curves in Figure 4, the increasing number of features does not guarantee the better prediction performances for each feature selection method, as they may have a higher possibility of being correlated or redundancy.

Table 4 provides the prediction performance of the models built with the optimal feature set for each feature selection method. The feature dimensions of the optimal feature sets in Table 4 are respectively the values of the x-coordinate when the corresponding curves in Figure 4 reach their maximums. As shown in Figure 4, the model trained with a combination of AE and Adasyn yields the highest MCC with the feature dimension being 90. As given in Table 4, the model trained with a combination of t-SNE and Adasyn yields the highest Sn and AUC with the feature dimension being 3. According to the results of Figure 4, the feature dimension will be set up to 90. In the case of the comparable performance achieved by different models, we tend to select the t-SNE + Adasyn with 3 features to significantly reduce the computational cost and the risk of overfitting. The 3 potentially important features incorporate the combinatorial information of all the features. Therefore, we should analyze the correlations of all the features and protein citrullination sites. In this study, we explore four types of quantitative feature descriptors, including binary encoding (BE), position specific amino acid propensity (PSAAP), pseudo amino acid composition (PseAAC), and physicochemical properties (PP). The BE and the PSAAP measures the amino acid preferences in different positions flanking the known citrullination sites; The PseAAC incorporates the order information hidden in protein sequences; The PP of residues is known to be important for protein interactions as it is associated with protein folding, interior packing, catalytic mechanism. These features may provide some clues for uncovering the mechanisms of protein citrullinations. Furthermore, the classification boundary of PCSs and non-PCSs in the feature space obtained by t-SNE + Adasyn is clearly visible in Figure 6. Therefore, the SVM, Adasyn, and t-SNE are respectively employed as the prediction algorithm, the over-sampling method, and the feature selection method to construct our final PCS predictor, PCSPred_SC.

E. EFFECTIVENESS OF THE FEATURE SELECTION METHODS

Feature selection is a crucial step for constructing a robust prediction model. To evaluate the effectiveness of the feature selection method, the prediction performance on the original feature set without feature selection is compared to that on the optimal feature subset with feature selection. As listed in Table 3 and Table 4, the SVM + Adasyn with MI is superior to the SVM + Adasyn without MI in terms of Acc and MCC increasing from 0.934 and 0.850 to 0.937 and 0.858, respectively. Similar conclusions can be conducted for SVM + Adasyn with AE or t-SNE. Except that the Acc

TABLE 4. Performance comparisons of different feature selection methods with Adasyn or SVM-SMOTE.

Prediction Algorithm	Feature Selection Method	Over-sampling Method	Feature Dimension	Sn	Sp	Acc	MCC
SVM	MI	Adasyn	285	0.836	0.987	0.937	0.858
		SVM-SMOTE	239	0.767	0.987	0.914	0.806
	t-SNE	Adasyn	3	0.948	0.931	0.937	0.862
		SVM-SMOTE	4	0.836	0.931	0.899	0.772
	AE	Adasyn	90	0.845	0.987	0.940	0.864
		SVM-SMOTE	93	0.767	0.987	0.914	0.806

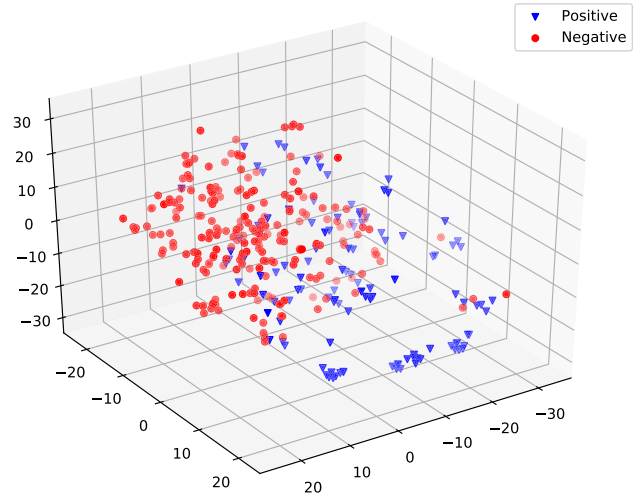


FIGURE 6. The distributions of PCSs and non-PCSs in the feature space obtained by t-SNE + Adasyn.

and MCC achieved by SVM + SVM-SMOTE with t-SNE is lower than those achieved by SVM + SVM-SMOTE without t-SNE, all the models with feature selection in Table 4 outperforms the models without feature selection in Table 3. These results indicate that the feature selection methods adopted in this study are effective to remove irrelevant and redundant features from the original feature space.

F. PERFORMANCE COMPARISONS UNDER DIFFERENT VALIDATION METHODS

After the predictor is completely trained using the training set, the independent testing is performed using the independent test set. In the leave-one-out cross validation, each sequence in the dataset is in turn singled out as the independent test sample and the remaining samples train the predictor. The 10-fold cross validation, the leave-one-out cross validation and the independent dataset test are respectively conducted 20 times and the corresponding performance measures are averaged to avoid over-fitting. The results in Table 5 shows that the prediction performance under the 10-fold cross validation, the leave-one-out cross validation and the independent dataset test is exactly similar, indicating the robustness and the excellent generalization ability of the proposed method.

TABLE 5. The prediction performance of PCSPred_SC under different validation methods.

Validation Method	Sn	Sp	Acc	MCC	AUC
10-Fold Cross Validation	0.948	0.931	0.937	0.862	0.997
Leave-One-Out Cross Validation	0.957	0.931	0.940	0.869	0.998
Independent Dataset Test	0.928	0.933	0.931	0.861	0.995

G. PERFORMANCE COMPARISONS WITH EXISTING METHODS

To gain insights into the efficiency of the proposed PCSPred_SC, we make comparisons with the competing prediction methods, Q. Zhang *et al.*'s method [35] and CKSAAP_CitrSite [37] by the 10-fold cross validation. For PCSPred_SC in Table 6, the performance measures are calculated by the prediction results of the samples in the original dataset, and not including the prediction results of the samples generated by the over-sampling method. That is to say, the data used to compare the prediction performance of the proposed method with other methods is not changed. Therefore, the comparisons are relatively fair. As listed in Table 6 where the best results are highlighted in bold, all performance measures except Sp achieved by PCSPred_SC is superior to those of the competing prediction methods. Specifically, PCSPred_SC achieves the highest MCC, followed by CKSAAP_CitrSite with MCC = 0.753 and Q. Zhang *et al.*'s method with MCC = 0.598. The Acc yielded by PCSPred_SC is 0.937, which is respectively 0.079 and 0.043 higher than Q. Zhang *et al.*'s method and CKSAAP_CitrSite. The high Sps achieved by Q. Zhang *et al.*'s method and CKSAAP_CitrSite with values of 0.943 and 0.953 are notably accompanied with extremely low Sn with values of 0.603 and 0.776, respectively. On the contrary, PCSPred_SC achieves a pretty high Sn of 0.948. The excellent performance of PCSPred_SC is also reflected in the value of AUC approaching to 1. Most importantly, PCSPred_SC just employs 3 features to yield the outstanding performance, followed by Q. Zhang *et al.*'s method with 44 features and CKSAAP_CitrSite with 250 features. Overall, PCSPred_SC significantly enhances the PCS prediction performance and at the same time reduces the number of features used for this task remarkably.

There are some possible factors accounting for the competitive performance of PCSPred_SC. Firstly, the feature extraction methods can capture the characteristics of PCSs, leading to more discriminative power; Secondly, the imbalanced dataset problem is solved by the over-sampling methods; Thirdly, the feature selection methods are effective to remove irrelevant and redundant features; Lastly, the combined method integrates the consistency of prediction algorithms, over-sampling methods, and feature selection methods.

Generally, overfittng occurs under the following 3 cases: (i) high-dimensional features containing noise; (ii) overtraining; (iii) insufficient training data. To reduce the influence of the overfitting problem, we have adopted feature selection methods to map the high dimensional feature space to a low dimensional feature space, while filtering out the redundant and noisy information. In addition, the traditional prediction algorithms are employed to implement the classification with some default parameters to prevent overtraining. Therefore, the insufficient training data adopted in this study and previous studies is the potential factor that may cause our model overfitting. Recent breakthrough of proteomic techniques has resulted in a rapid growth of newly discovered protein sequences. In the future work, expanding the benchmark dataset for citrullination site prediction to avoid overfitting will be an important research direction.

TABLE 6. Performance comparisons with the existing methods by the 10-fold cross validation.

Method	Reference	Feature Dimension	Sn	Sp	Acc	MCC	AUC
Q. Zhang et al.[a]	[35]	44	0.603	0.943	0.858	0.598	—[b]
CKSAAP_CitrSite[a]	[37]	250	0.776	**0.953**	0.894	0.753	0.941
PCSPred_SC	This study	**3**	**0.948**	0.931	**0.937**	**0.862**	**0.997**

[a] The corresponding prediction results are cited from the reference [37].
[b] "—" means the corresponding value is not available in the reference [35].

IV. CONCLUSIONS

In view of the significant roles of PCSs on numerous biological events and human diseases, a novel and powerful sequence-based PCS prediction method named PCSPred_SC is proposed with hybrid features integrating BE, PSAAP, PseAAC, and PP. Under the complete feature space, the PCS predictors are respectively constructed by various prediction algorithms. To solve the imbalanced dataset problem, several over-sampling methods are systemically explored. For the irrelevant and redundant features in the feature space, the feature selection methods are adopted to further enhance prediction performance. Experimental results indicate that the combination of SVM, Adasyn, and t-SNE attains much more outstanding performance for PCS prediction. When performed on the training dataset using the 10-fold cross validation, PCSPred_SC achieves excellent performance with a Sn of 0.948, a Sp of 0.931, a Acc of 0.937, a MCC of 0.862 and a AUC of 0.997, which is far better than the competing methods. Furthermore, PCSPred_SC can significantly reduce the computational and space cost by just employing 3 features. In the future work, a wider range of segmented-based feature extraction methods will be integrated into PCSPred_SC to further improve the performance. Additionally, we will construct deep learning based framework to solve deficiencies of the traditional hand-crafted features.

REFERENCES

[1] H. Härmä, N. Tong-Ochoa, A. J. van Adrichem, I. Jelesarov, K. Wennerberg, and K. Kopra, "Toward universal protein post-translational modification detection in high throughput format," *Chem. Commun.*, vol. 54, no. 23, pp. 2910–2913, 2018.

[2] M. K. Verheul, P. A. van Veelen, M. A. M. van Delft, A. de Ru, G. M. C. Janssen, T. Rispens, R. E. M. Toes, and L. A. Trouw, "Pitfalls in the detection of citrullination and carbamylation," *Autoimmunity Rev.*, vol. 17, no. 2, pp. 136–141, Feb. 2018.

[3] S. Wang and Y. Wang, "Peptidylarginine deiminases in citrullination, gene regulation, health and pathogenesis," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mech.*, vol. 1829, no. 10, pp. 1126–1135, Oct. 2013.

[4] J. Fuhrmann, K. W. Clancy, and P. R. Thompson, "Chemical biology of protein arginine modifications in epigenetic regulation," *Chem. Rev.*, vol. 115, no. 11, pp. 5413–5461, Jun. 2015.

[5] A. Muth, V. Subramanian, E. Beaumont, M. Nagar, P. Kerry, P. McEwan, H. Srinath, K. Clancy, S. Parelkar, and P. R. Thompson, "Development of a selective inhibitor of protein arginine deiminase 2," *J. Med. Chem.*, vol. 60, no. 7, pp. 3198–3211, 2017.

[6] A. Ishida-Yamamoto, T. Senshu, R. A. J. Eady, H. Takahashi, H. Shimizu, M. Akiyama, and H. Iizuka, "Sequential reorganization of cornified cell keratin filaments involving filaggrin-mediated compaction and keratin 1 deimination," *J. Investigative Dermatol.*, vol. 118, no. 2, pp. 282–287, Feb. 2002.

[7] S. Horibata, K. E. Rogers, D. Sadegh, L. J. Anguish, J. L. McElwee, P. Shah, P. R. Thompson, and S. A. Coonrod, "Role of peptidylarginine deiminase 2 (PAD2) in mammary carcinoma cell migration," *BMC Cancer*, vol. 17, no. 1, p. 378, Dec. 2017.

[8] V. V. Nemmara, V. Subramanian, A. Muth, S. Mondal, A. J. Salinger, A. J. Maurais, R. Tilvawala, E. Weerapana, and P. R. Thompson, "The development of benzimidazole-based clickable probes for the efficient labeling of cellular protein arginine deiminases (PADs)," *ACS Chem. Biol.*, vol. 13, no. 3, pp. 712–722, Mar. 2018.

[9] L. Y. Guo, "Research progress on correlation of PAD-mediated citrullination with malignant tumors," *Prog. Mod. Biomed.*, vol. 15, no. 30, pp. 5977–5981, 2015.

[10] E. Witalison, P. Thompson, and L. Hofseth, "Protein arginine deiminases and associated citrullination: Physiological functions and diseases associated with dysregulation," *Current Drug Targets*, vol. 16, no. 7, pp. 700–710, Jul. 2015.

[11] A. Steckel and G. Schlosser, "Citrulline effect is a characteristic feature of deiminated peptides in tandem mass spectrometry," *J. Amer. Soc. Mass Spectrometry*, vol. 30, no. 9, pp. 1586–1591, Sep. 2019.

[12] L. Wang, G. Song, X. Zhang, T. Feng, J. Pan, W. Chen, M. Yang, X. Bai, Y. Pang, J. Yu, J. Han, and B. Han, "PADI2-mediated citrullination promotes prostate cancer progression," *Cancer Res.*, vol. 77, no. 21, pp. 5755–5768, Nov. 2017.

[13] C. Zhao, B. Ling, L. Dong, and Y. Liu, "Theoretical insights into the protonation states of active site cysteine and citrullination mechanism of *Porphyromonasgingivalis* peptidylarginine deiminase," *Proteins, Struct., Function, Bioinf.*, vol. 85, no. 8, pp. 1518–1528, Aug. 2017.

[14] Z. Baka, B. György, P. Géher, E. I. Buzás, A. Falus, and G. Nagy, "Citrullination under physiological and pathological conditions," *Joint Bone Spine*, vol. 79, no. 5, pp. 431–436, Oct. 2012.

[15] L. Harlow, I. O. Rosas, B. R. Gochuico, T. R. Mikuls, P. F. Dellaripa, C. V. Oddis, and D. P. Ascherman, "Identification of citrullinated Hsp90 isoforms as novel autoantigens in rheumatoid arthritis-associated interstitial lung disease," *Arthritis Rheumatism*, vol. 65, no. 4, pp. 869–879, Apr. 2013.

[16] K. Van Steendam, K. Tilleman, M. De Ceuleneer, F. De Keyser, D. Elewaut, and D. Deforce, "Citrullinated vimentin as an important antigen in immune complexes from synovial fluid of rheumatoid arthritis patients with antibodies against citrullinated proteins," *Arthritis Res. Therapy*, vol. 12, no. 4, p. R132, 2010.

[17] S. Ling, E. N. Cline, T. S. Haug, D. A. Fox, and J. Holoshitz, "Citrullinated calreticulin potentiates rheumatoid arthritis shared epitope signaling," *Arthritis Rheumatism*, vol. 65, no. 3, pp. 618–626, Mar. 2013.

[18] T. Goulas, D. Mizgalska, I. Garcia-Ferrer, T. Kantyka, T. Guevara, B. Szmigielski, A. Sroka, C. Millán, I. Usón, F. Veillard, B. Potempa, P. Mydel, M. Solà, J. Potempa, and F. X. Gomis-Rüth, "Structure and mechanism of a bacterial host-protein citrullinating virulence factor, *Porphyromonasgingivalis* peptidylarginine deiminase," *Sci. Rep.*, vol. 5, no. 1, Dec. 2015, Art. no. 11969.

[19] Y. Zheng, G. Zhao, B. Xu, C. Liu, C. Li, X. Zhang, and X. Chang, "PADI4 has genetic susceptibility to gastric carcinoma and upregulates CXCR2, KRT14 and TNF-α expression levels," *Oncotarget*, vol. 7, no. 38, pp. 62159–62176, 2016.

[20] H. Wang, B. Xu, X. Zhang, Y. Zheng, Y. Zhao, and X. Chang, "PADI2 gene confers susceptibility to breast cancer and plays tumorigenic role via ACSL4, BINC3 and CA9 signaling," *Cancer Cell Int.*, vol. 16, no. 1, Dec. 2016.

[21] E. E. Witalison, X. Cui, C. P. Causey, P. R. Thompson, and L. J. Hofseth, "Molecular targeting of protein arginine deiminases to suppress colitis and prevent colon cancer," *Oncotarget*, vol. 6, no. 34, pp. 36053–36062, Nov. 2015.

[22] R. Tilvawala, S. H. Nguyen, A. J. Maurais, V. V. Nemmara, M. Nagar, A. J. Salinger, S. Nagpal, E. Weerapana, and P. R. Thompson, "The rheumatoid arthritis-associated citrullinome," *Cell Chem. Biol.*, vol. 25, no. 6, pp. 691.e6–704.e6, Jun. 2018.

[23] L.-M. Mauracher, F. Posch, K. Martinod, E. Grilz, T. Däullary, L. Hell, C. Brostjan, C. Zielinski, C. Ay, D. D. Wagner, I. Pabinger, and J. Thaler, "Citrullinated histone H3, a biomarker of neutrophil extracellular trap formation, predicts the risk of venous thromboembolism in cancer patients," *J. Thrombosis Haemostasis*, vol. 16, no. 3, pp. 508–518, Mar. 2018.

[24] H. Qin, X. Liu, F. Li, L. Miao, T. Li, B. Xu, X. An, A. Muth, P. R. Thompson, S. A. Coonrod, and X. Zhang, "PAD1 promotes epithelial-mesenchymal transition and metastasis in triple-negative breast cancer cells by regulating MEK1-ERK1/2-MMP2 signaling," *Cancer Lett.*, vol. 409, pp. 30–41, Nov. 2017.

[25] B. D. Cherrington, X. Zhang, J. L. McElwee, E. Morency, L. J. Anguish, and S. A. Coonrod, "Potential role for PAD2 in gene regulation in breast cancer cells," *PLoS ONE*, vol. 7, no. 7, 2012, Art. no. e41242.

[26] N. Cantariño, M. T. Fernández-Figueras, V. Valero, E. Musulén, R. Malinverni, I. Granada, S. J. Goldie, J. Martín-Caballero, J. Douet, S.-V. Forcales, and M. Buschbeck, "A cellular model reflecting the phenotypic heterogeneity of mutant HRAS driven squamous cell carcinoma," *Int. J. Cancer*, vol. 139, pp. 1106–1116, Sep. 2016.

[27] N. Cantarino, E. Musulen, V. Valero, M. A. Peinado, M. Perucho, V. Moreno, S.-V. Forcales, J. Douet, and M. Buschbeck, "Downregulation of the deiminase PADI2 is an early event in colorectal carcinogenesis and indicates poor prognosis," *Mol. Cancer Res.*, vol. 14, no. 9, pp. 841–848, Sep. 2016.

[28] S. Aratani, H. Fujita, N. Yagishita, Y. Yamano, Y. Okubo, K. Nishioka, and T. Nakajima, "Inhibitory effects of ubiquitination of synoviolin by PADI4," *Mol. Med. Rep.*, vol. 16, no. 6, pp. 9203–9209, Dec. 2017.

[29] M. A. Shelef, J. Sokolove, L. J. Lahey, C. A. Wagner, E. K. Sackmann, T. F. Warner, Y. Wang, D. J. Beebe, W. H. Robinson, and A. Huttenlocher, "Peptidylarginine deiminase 4 contributes to tumor necrosis factor α-induced inflammatory arthritis," *Arthritis Rheumatol.*, vol. 66, no. 6, pp. 1482–1491, 2014.

[30] M. Sharma, D. Damgaard, L. Senolt, B. Svensson, A. C. B. Jensen, C. H. Nielsen, and P. Hägglund, "Identification of citrullination sites specific for peptidylarginine deiminase 2 (PAD2) and PAD4 in fibrinogen from synovial fluid of patients with rheumatoid arthritis," *Ann. Rheumatic Diseases*, vol. 76, no. 2, p. 501, 2017.

[31] S. M. M. Hensen and G. J. M. Pruijn, "Methods for the detection of peptidylarginine deiminase (PAD) activity and protein citrullination," *Mol. Cellular Proteomics*, vol. 13, no. 2, pp. 388–396, Feb. 2014.

[32] S. Hensen, W. Boelens, K. Bonger, R. van Cruchten, F. van Delft, and G. Pruijn, "Phenylglyoxal-based visualization of citrullinated proteins on western blots," *Molecules*, vol. 20, no. 4, pp. 6592–6600, 2015.

[33] M. Hermansson, K. Artemenko, E. Ossipova, H. Eriksson, J. Lengqvist, D. Makrygiannakis, A. I. Catrina, A. P. Nicholas, L. Klareskog, M. Savitski, R. A. Zubarev, and P.-J. Jakobsson, "MS analysis of rheumatoid arthritic synovial tissue identifies specific citrullination sites on fibrinogen," *Proteomics Clin. Appl.*, vol. 4, no. 5, pp. 511–518, 2010.

[34] E. Shin and S. Cha, "*In situ* probing citrullinated sites in a peptide by reactive desorption electrospray ionization mass spectrometry," *Bull. Korean Chem. Soc.*, vol. 39, no. 1, pp. 40–44, Jan. 2018.

[35] Q. Zhang, X. Sun, K. Feng, S. Wang, Y.-H. Zhang, S. Wang, L. Lu, and Y.-D. Cai, "Predicting citrullination sites in protein sequences using mRMR method and random forest algorithm," *Combinat. Chem. High Throughput Screening*, vol. 20, no. 2, pp. 164–173, Jun. 2017.

[36] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, Mar. 2011.

[37] Z. Ju and S.-Y. Wang, "Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition," *Gene*, vol. 664, pp. 78–83, Jul. 2018.

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

[39] O. Bedoya and I. Tischer, "Remote homology detection incorporating the context of physicochemical properties," *Comput. Biol. Med.*, vol. 45, pp. 43–50, Feb. 2014.

[40] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[41] H. Han, W. Wen-Yuan, and M. Bing-Huan, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput.*, Aug. 2005, pp. 878–887.

[42] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," *Int. J. Knowl. Eng. Soft Data Paradigms*, vol. 3, no. 1, pp. 4–21, 2009.

[43] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 1322–1328.

[44] R.-N. Ma, B. Wang, G. Wang, X.-C. Guo, and W.-B. Liu, "Evaluation method for node importance in communication network based on mutual information," *Acta Electron. Sinica*, vol. 45, no. 3, pp. 747–752, 2017.

[45] Z. Aydin, O. Kaynar, and Y. Görmez, "Dimensionality reduction for protein secondary structure and solvent accesibility prediction," *J. Bioinf. Comput. Biol.*, vol. 16, no. 5, Oct. 2018, Art. no. 1850020.

[46] W. Zhu, Z. T. Webb, K. Mao, and J. Romagnoli, "A deep learning approach for process data visualization using t-distributed stochastic neighbor embedding," *Ind. Eng. Chem. Res.*, vol. 58, no. 22, pp. 9564–9575, Jun. 2019.

[47] M. Arif, M. Hayat, and Z. Jan, "IMem-2LSAAC: A two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into chou's pseudo amino acid composition," *J. Theor. Biol.*, vol. 442, pp. 11–21, Apr. 2018.

[48] Y. Liang and S. Zhang, "Prediction of apoptosis Protein's subcellular localization by fusing two different descriptors based on evolutionary information," *Acta Biotheoretica*, vol. 66, no. 1, pp. 61–78, Mar. 2018.

[49] C. Zhang, P. L. Freddolino, and Y. Zhang, "COFACTOR: Improved protein function prediction by combining structure, sequence and protein-protein interaction information," *Nucleic Acids Res.*, vol. 45, no. W1, pp. W291–W299, Jul. 2017.

[50] X. Cheng, W.-Z. Lin, X. Xiao, and K.-C. Chou, "pLoc_bal-mAnimal: Predict subcellular localization of animal proteins by balancing training dataset and PseAAC," *Bioinformatics*, vol. 35, no. 3, pp. 398–406, 2019.

[51] W. Zhang, X. Yue, G. Tang, W. Wu, F. Huang, and X. Zhang, "SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions," *PLOS Comput. Biol.*, vol. 14, no. 12, 2018, Art. no. e1006616.

[52] L. Deng and Z. Chen, "An integrated framework for functional annotation of protein structural domains," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 4, pp. 902–913, Jul. 2015.

[53] P. Breheny and J. Huang, "Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors," *Statist. Comput.*, vol. 25, no. 2, pp. 173–187, Mar. 2015.

[54] I. Ahmed, P. Witbooi, and A. Christoffels, "Prediction of human-Bacillus anthracis protein–protein interactions using multi-layer neural network," *Bioinformatics*, vol. 34, no. 12, pp. 4159–4164, 2018.

[55] B. C. Yavuz, N. Yurtay, and O. Ozkan, "Prediction of protein secondary structure with clonal selection algorithm and multilayer perceptron," *IEEE Access*, vol. 6, pp. 45256–45261, 2018.

[56] M. Tayefi, H. Esmaeili, M. S. Karimian, A. A. Zadeh, M. Ebrahimi, M. Safarian, M. Nematy, S. M. R. Parizadeh, G. A. Ferns, and M. Ghayour-Mobarhan, "The application of a decision tree to establish the parameters associated with hypertension," *Comput. Methods Programs Biomed.*, vol. 139, pp. 83–91, Feb. 2017.

[57] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

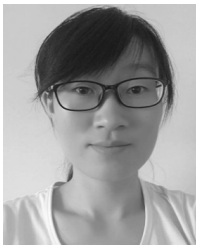
[58] S. Daberdaku and C. Ferrari, "Exploring the potential of 3D Zernike descriptors and SVM for protein–protein interface prediction," *BMC Bioinf.*, vol. 19, no. 1, p. 35, 2018.

[59] R. Sharma, M. Bayarjargal, T. Tsunoda, A. Patil, and A. Sharma, "MoRFPred-plus: Computational identification of MoRFs in protein sequences using physicochemical properties and HMM profiles," *J. Theor. Biol.*, vol. 437, pp. 9–16, Jan. 2018.

[60] Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, G. I. Webb, A. I. Smith, R. J. Daly, K.-C. Chou, and J. Song, "iFeature: A Python package and Web server for features extraction and selection from protein and peptide sequences," *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, Jul. 2018.

[61] L. Zhang, G. Yu, D. Xia, and J. Wang, "Protein–protein interactions prediction based on ensemble deep neural networks," *Neurocomputing*, vol. 324, pp. 10–19, Jan. 2019.

[62] H. Hu, L. Zhang, H. Ai, H. Zhang, Y. Fan, Q. Zhao, and H. Liu, "HLPI-Ensemble: Prediction of human lncRNA-protein interactions based on ensemble strategy," *RNA Biol.*, vol. 15, no. 6, pp. 797–806, 2018.



LINA ZHANG was born in Zibo, Shandong, China, in 1987. She received the B.S. degree in engineering from the School of Information and Control Engineering, China University of Petroleum, in 2010, the M.S. and Ph.D. degrees from the School of Control Science and Engineering, Shandong University, in 2012 and 2017, respectively. She is currently a Lecturer with the School of Mechanical, Electrical and Information Engineering, Shandong University at Weihai. Her current research interests include bioinformatics, system biology, and the mathematical modeling of molecular biology.

JINGUI CHEN was born in Shangrao, Jiangxi, China, in 1995. She received the B.S. degree in engineering from the School of Physical Science and Technology, Shenyang Normal University, in 2017. She is currently pursuing the master's degree with the School of Control Science and Engineering, Shandong University. Her research interests include bioinformatics and machine learning.

CHENGJIN ZHANG was born in Laiwu, Shandong, China, in 1962. He received the M.S. degree from the Shandong University of Science and Technology, in 1992, and the Ph.D. degree from Northeastern University, in 1997. He is currently a Professor with the School of Mechanical, Electrical and Information Engineering, Shandong University at Weihai. His research interests include control theory and applications, intelligent robot control, and bioinformatics.

RUI GAO received the Ph.D. degree in applied mathematics from Shandong University, in 2003. He is currently a Professor with the School of Control Science and Engineering, Shandong University. His current research interests include hybrid dynamical systems, optimal control theory, mathematical modeling of molecular biology, and bioinformatics.

RUNTAO YANG was born in Dongying, Shandong, China, in 1989. He received the B.S. degree in engineering from the School of Mechanical, Electrical and Information Engineering, Shandong University at Weihai, in 2011, and the Ph.D. degree in control science and engineering from Shandong University, in 2016. He is currently a Lecturer with the School of Mechanical, Electrical and Information Engineering, Shandong University at Weihai. His research interests include swarm intelligence robotics, bioinformatics, and system biology.