IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 15, 2019, accepted September 24, 2019, date of publication September 30, 2019,
date of current version October 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2944411

DeepGly: A Deep Learning Framework With Recurrent and Convolutional Neural Networks to Identify Protein Glycation Sites From Imbalanced Data

JINGUI CHEN[1], RUNTAO YANG[1], CHENGJIN ZHANG[1], LINA ZHANG[1], AND QIAN ZHANG[2]
[1]School of Mechanical, Electrical and Information Engineering, Shandong University at Weihai, Weihai 264209, China
[2]Heze Institute of Science and Technology Information, Heze 274015, 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 61573213, Grant 61673245, and Grant 61603214.

ABSTRACT As an unavoidable non-enzymatic reaction between proteins and reducing sugars, glycation can decline antioxidant defense mechanisms, damage cellular organelles, and form advanced glycation end products (AGEs), thereby resulting in a series of destructive physiological diseases. Identification and analysis of protein glycation sites will be beneficial to understand the complex pathogenesis related to the glycation. In this paper, a new glycation site predictor, DeepGly, is proposed based on a deep learning framework with a recurrent neural network (RNN) and a convolutional neural network (CNN). Firstly, for the class imbalance problem in the benchmark dataset, Long Short-Term Memory (LSTM) RNNs are designed to generate artificial peptides with glycation sites to form a balanced dataset. Then, the peptides in the balanced dataset are cleaved into a series of biological words, and continuous distribution representation is employed to transform the biological words into digital vectors. Finally, the digital vectors are input into the CNN with participations of the plurality and multiple convolution kernels to automatically extract various features, pooling layers to perform feature selection, and a softmax function to classify peptides. On the same datasets using 10-fold cross validation test, the prediction performance of DeepGly is far superior to that of existing methods, which indicates that the proposed method can be used as an ideal choice for protein glycation site prediction and also has a certain promotion effect on other related fields.

INDEX TERMS Glycation, recurrent neural network, continuous distributed representation, convolutional neural network.

I. INTRODUCTION

Glycation, first described by LC Maillard in 1912, is one of the most important post-translational modifications (PTMs) that involves a series of complex reactions as follows [1]. The formyl and ketone having a carbonyl group are produced by the beta-oxidation or the peroxidation reduction of sugar. The oxygen atom of the carbonyl group is negatively charged. Under high glucose condition, it can undergo non-enzymatic glycation reactions with nucleophilic groups in biomolecules to form advanced glycation end products (AGEs). The long-term accumulation of AGEs in the human body will trigger the following two major cellular effects: (i) Intermolecular bonds or crosslinking between extracellular and intracellular proteins occur, and physiological characteristics of extracellular matrix proteins are changed. (ii) The interactions between AGEs and the cell surface receptor for AGE (RAGE) happen, activating complex signaling pathways that ultimately lead to the production of pro-inflammatory mediators and reactive oxygen species [2], [3]. Studies [4]–[8] have shown that these variations are closely related to the pathogenesis of many diseases such as diabetes [9], [10], nephritis [11], [12], atherosclerosis [13], [14], cataract [15], [16], Alzheimer's disease [17], [18], etc. As listed in Table 1, the formations of typical AGE compounds are mainly associated with lysine. In other words, the majority of glycation

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

TABLE 1. Typical AGE compounds and their precursors.

AGE compound	Precursor
Nε-carboxymethyl-lysine (CML)	Lysine, glyoxal
Nε-carboxyethyl-lysine (CEL)	Lysine, methylglyoxal
N-fructosyl-lysine	Lysine
Pyrraline	Lysine, 3-deoxyglucosone
Glucosepane	Lysine
Imidazolium dilysine (IDL)	Lysine
Alkyl formyl glycosyl pyrroles (AFGP)	Lysine
Arginine-lysine imidazole (ALI)	Lysine, arginine
Glyoxal lysine dimer (GOLD)	Lysine, glyoxal
Methylglyoal lysine dimer (MOLD)	Lysine, methylglyoxal
Crossline	Lysine
Pentosidine	Lysine, arginine
Argpyrimidine	Arginine, methylglyoxal
Vesperlysine	Lysine

reactions occur in lysine. Therefore, identification of lysine glycation sites is particularly crucial to understanding the pathogenesis and providing a theoretical basis for curing the diseases.

Generally, it is time-consuming and labor-intensive to perform functional annotation of protein binding sites by conventional experimental methods. In recent years, machine learning theory has provided new ideas for protein glycation site prediction. Through statistical analysis, Johansen *et al.* indicated that acidic amino acids catalyze the glycations of nearby lysine. They established a neural network-based glycation site predictor ''GlyNN'' with the amino acid composition around the glycation site and the positional information of lysine as input [4]. After that, taking support vector machine (SVM) as the prediction algorithm, Liu *et al.* developed an improved predictor ''PreGly'' by integrating different feature extraction strategies, including the frequency of amino acid appearance, amino acid factor and k-spaced amino acid pair [5]. In 2016, Xu *et al.* investigated the role of sequence information and position-specific amino acid propensity (PSAAP) in glycation site prediction, and constructed the predictor ''Gly-PseAAC'' based on the Compendium of a database of protein lysine modifications (CPLM). The prediction results demonstrated that PSAAP could distinguish whether lysine in the peptide chain undergoes glycation reactions [6]. In 2017, Zhao *et al.* applied multiple feature information to encode peptide chains for glycation site prediction, including positional scoring functions, secondary structures, AAindex and k-spaced amino acid pairs. A new predictor ''Glypre'' [7] was developed with different window sizes and a two-step feature selection. In 2018, Islam *et al.* constructed a SVM based glycation site predictor ''iProtGly-SS'' by extracting amino acid compositions, secondary structures and polarities from peptide chains, and then employing the forward feature selection method to obtain the optimal feature set [8].

Although the methods discussed above have their own advantages and did promote the understanding on protein glycations, there is still room for significant improvements. Firstly, glycation site prediction is a standard imbalanced learning problem where the number of negative samples in the benchmark dataset is much larger than that of positive samples. Thus, it could lead to biased prediction [19]. There are three solutions [20] available for this class imbalance problem. The first is data-level technology, which modifies the training set to fit the standard learning algorithm. The second is the algorithm-level approaches that modifies existing algorithms to mitigate their bias towards majority groups. The third is the hybrid methods that combines the aforementioned methods to extract their strengths and reduce their weaknesses. In addition, the data-level technology can be divided into two types. One is to develop undersampling algorithms to delete negative samples, which may lose part of the sample information. Another is to develop oversampling algorithms to add positive samples such as the Synthetic Minority Oversampling Technique (SMOTE) [21], [22], which may not work well in high-dimensional data. To deal with the imbalanced dataset problem, a LSTM RNNs model is constructed in this study to oversample peptide chains with glycation sites.

In recent years, deep learning has been proven extremely effective in object detection and recognition [23], managing smart homes [24], signals diagnosis [25]. More specifically, based on the fuzzy theory and a convolutional neural network, a novel two stage model for detection of important features from images is proposed in [23]; A hierarchical structure consisting two types of neural networks is proposed in [24]; A Radial Basis Probabilistic Neural Network (RBPNN) trained by a novel method to preserve its generalization is proposed in [25]. As a common deep learning method, Recurrent Neural Networks (RNN) [26] can predict the next possible word distribution is that the hidden layer records the information in the front of the sequence and adds it to the current input to affect the output. It widely used in various tasks of NLP such as text generation [27], word segmentation [28], translation [29], etc. However, due to gradient disappearance or gradient explosion, traditional RNN structures are difficult to handle long-term dependencies in sequences. Bypassing these problems, Long Short-Term Memory (LSTM) cell can make the model remember long-term information [30]. Protein sequences can be regarded as a special genetic language, which is highly similar with natural language. Thus, LSTM RNNs [31] will be designed in this study to generate peptide chains with lysines.

In order to simplify the process of feature extraction, the ProtVec [32] extracted from the Skip-gram network will be employed in this study to transform sparse peptide chain features into a denser representation (continuous distribution representation) [33], [34]. Continuous distribution representation can capture the semantic and syntactic relationships between words in a sentence [35]. The peptide chains encoded by continuous distribution representation have been successfully applied in the fields of protein family classification, protein visualization, protein structure prediction, disordered protein identification and protein-protein interaction prediction [32].

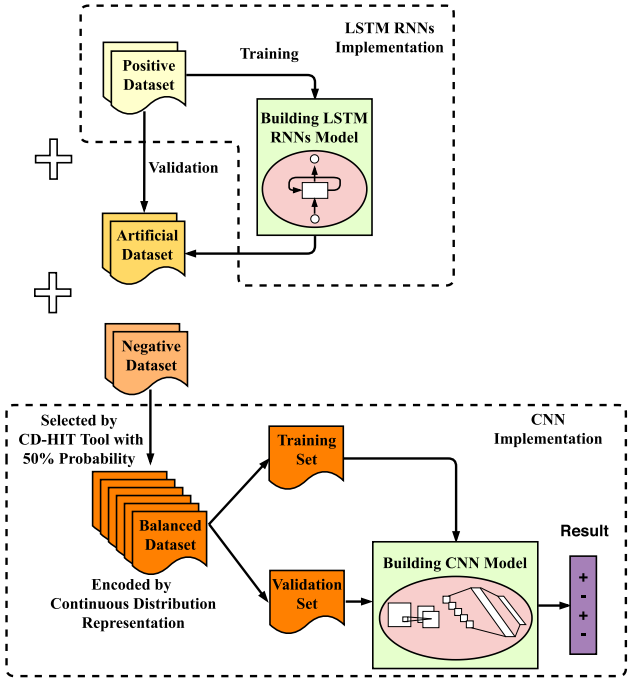


FIGURE 1. The overall workflow of the present method. Firstly, the LSTM RNNs model trained with the positive dataset is used to sample peptide chains at an appropriate sampling temperature. To validate the effectiveness of the generated peptide chains, we present systematic comparisons between the positive samples and the generated peptide chains. Then, the artificial samples generated by the LSTM RNNs are added to the dataset, and the redundancy of new dataset is removed by the CD-HIT with a threshold of 50%. Finally, the CNN is constructed to predict glycation sites with the samples in the balanced dataset encoded by continuous distribution representation as input.

The major limitation of the prediction capabilities of the earlier methods is that the strategies to extract traditional hand-crafted features are highly complicated and subjective. Convolution Neural Networks (CNN) [36] have the abilities to extract advanced features from the input data, and learn the nonlinear mapping from a specific input to a specific output [23], [37]. Kim applied CNN for the text classification and achieved a high accuracy [38]. Recently, deep learning methods have become a new paradigm for PTM predictions. Fu *et al.* applied CNN to predict the ubiquitination sites of proteins and achieved excellent success [39]. The results indicated that CNN could learn some instinct information from input data without using complex feature extraction and feature selection methods.

The establishment process of the model proposed in this study for glycation site prediction is shown in Figure 1. Firstly, the LSTM RNNs model trained with the positive dataset is used to sample peptide chains at an appropriate sampling temperature. To validate the effectiveness of the generated peptide chains, we present systematic comparisons between the positive samples and the generated peptide chains. Then, the artificial samples generated by the LSTM RNNs are added to the dataset, and the redundancy of new dataset is removed by the CD-HIT [40] with a threshold of 50%. Finally, the CNN is constructed to predict glycation sites with the samples in the balanced dataset encoded by continuous distribution representation as input.

II. MATERIALS AND METHODS

A. DATASETS

Three datasets, respectively named dataset A, dataset B, and dataset C, are introduced in this study for glycation site prediction. Among the three datasets, dataset A and dataset B respectively proposed in references [6] and [4] are employed to make a fair comparison between our proposed method and previous studies, while dataset C is employed to construct our glycation site predictor. These datasets will be described in detail below.

1) DATASET A

The dataset is derived from a comprehensive database of CPLM [41] available at http://cplm.biocuckoo.org/. CPLM encompasses 12 different experimentally identified protein lysine modifications. For the glycation site prediction, the dataset widely used in previous studies [6]–[8] contains 323 positive samples and 2046 negative samples extracted from 72 different proteins. If the sequence similarity is $>40\%$, the homologous peptides from the dataset are removed by using CD-HIT. Due to the excessive number of negative samples, some are randomly removed from the dataset. Finally, the dataset A contains 223 positive and 446 negative samples.

2) DATASET B

By searching more than 400 papers, Johansen *et al.* [4] obtained a lysine glycation site dataset. After manual inspection, it was found that part of the lysines were in propeptides and signal peptides, and some remained lysines were unconfirmed or controversial. To avoid confusing the prediction algorithm, the glycation sites mentioned above were masked out. As a result, the final dataset referred to dataset B in this paper was finally made up of 89 positive samples and 126 negative samples, and widely used in subsequent research.

3) DATASET C

The samples in dataset C is the union of the samples in dataset A before CD-HIT applied and the samples in dataset B. Previous methods were trained on relatively small datasets, which may affect the prediction performance. For this reason, dataset C will be employed to construct our glycation site predictor. To avoid redundancy and homology bias, the peptide chains greater than 50% sequence identity are removed by the CD-HIT. The final dataset contains 155 positive samples and 674 negative samples.

Suppose that the window size is η, a norm sample in the dataset is a peptide chain with $2\eta+1$ amino acids, and defined as follows.

$$P = A_{-\eta}A_{-}(\eta-1)\cdots A_{-2}A_{-1}KA_1A_2\ldots A_{(\eta-1)}A_{\eta}, \quad (1)$$

where $A_{-\eta}$ represents the η-th upstream amino acid of the center amino acid K, while A_{η} represents the

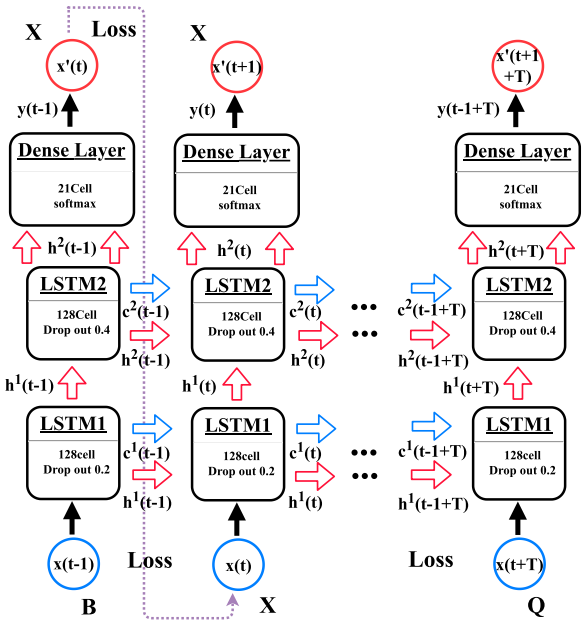


FIGURE 2. The architecture of the proposed LSTM RNNs to generate peptide chains with glycation sites.

η-th downstream amino acid of the center amino acid K. If the number of flanking amino acids is less than η, the missing positions are expanded with a special residue 'X'. The selected value of the window size η has a certain impact on the glycation site prediction performance. After validations of the prediction performance with different window sizes, it is observed that the predictor performs best for $\eta = 24$. Thus, the window size is chosen as 24.

B. ARCHITECTURE OF THE PROPOSED LSTM RNNS

As indicated in Section II.A, the number of peptide chains with glycation sites in the dataset C is much smaller than that of peptide chains without glycation sites. To deal with the imbalanced dataset problem, LSTM RNNs are proposed to generate peptide chains with glycation sites. Subsequently, the generated peptide chains are added to the dataset C to form a balanced dataset.

LSTM consisting of forgetting gates, input gates and output gates is an improved RNN that can enhance network prediction performance by learning long-term dependency information in training sequences [42]. Before building the LSTM network architecture as illustrated in Figure 2 for generating artificial positive samples, each input sample is added with a begin token 'B' at the head of the sample, and the corresponding One-Hot vector is ''00000000000000000000''. The order of amino acids is 'A', 'R', 'N', 'D', 'C', 'Q', 'E', 'G', 'H', 'I', 'L', 'K', 'M', 'F', 'P', 'S', 'T', 'W', 'Y', 'V', 'X', wherein the corresponding code for lysine (K) is ''00000000000010000000000''.

In Figure 2, for each amino acid residue $x_{(t)}$, LSTM RNNs can predict the next amino acid residue along the generated peptide chain. To depict the differences between the

Original Sequence

GKDLMEKVKSPELQAEAKSYFEKSKEQLTPLIKKAGTELVN

Biological Words

GKD KDL DLM LME MEK EKV KVK VKS KSP SPE PEL ... ...

FIGURE 3. The biological words extracted from an original sequence.

generated peptide chains with the actual peptide chains, the cross entropy loss calculated as in Equation 2 is adopted to obtain the optimal network weight.

$$L = -\sum_{i=1}^{K} y_i \log\left(\hat{y}_i\right), \tag{2}$$

where $\hat{y}_i$ and y_i respectively denote the i-th amino acid residues of the generated peptide chains and the actual peptide chains. K denotes the length of the peptide chain. Experiments are conducted to evaluate the performances of the multiscale LSTM RNNs with different hyper-parameters. The optimal hyper-parameters that can minimize the loss function are tuned using an automated grid search procedure [43], [44].

During generation, to control sequence variabilities, a temperature factor [45] is introduced in the softmax function, and defined as

$$P\left(y_i\right) = \frac{\exp\left(y_i/T\right)}{\sum_{j=1}^{K} \exp\left(y_i/T\right)}, \tag{3}$$

where y_i denotes the one-hot vector of the i-th amino acid residues of the peptide chains; K indicates the length of the peptide chain; T represents the sampling temperature. The value of the temperature factor can directly affect the diversity of the generated peptide chains. The higher the value of the temperature factor, the more diverse the generated peptide chains.

C. CONTINUOUS DISTRIBUTION REPRESENTATION

Derived from the NLP field, the concept of continuous distribution representation is proposed to map words from an original space to a new multidimensional space by the Skip-gram model or the Continuous Bag-Of-Words (CBOW) model [33], [34]. In order to increase the speed of calculation, the negative sampling technique is used for optimization [46].

In view of the obvious similarities between peptide chains and natural languages, each peptide chain is cut into biological words consisting of three adjacent amino acids as shown in Figure 3. The ProtVec [32] adopted in this study is a feature matrix extracted from biological words by continuous distribution representation, which can be formally represented as

$$ProtVec = \begin{bmatrix} BWVec(1) \\ BWVec(2) \\ \vdots \\ BWVec(N) \end{bmatrix}, \tag{4}$$

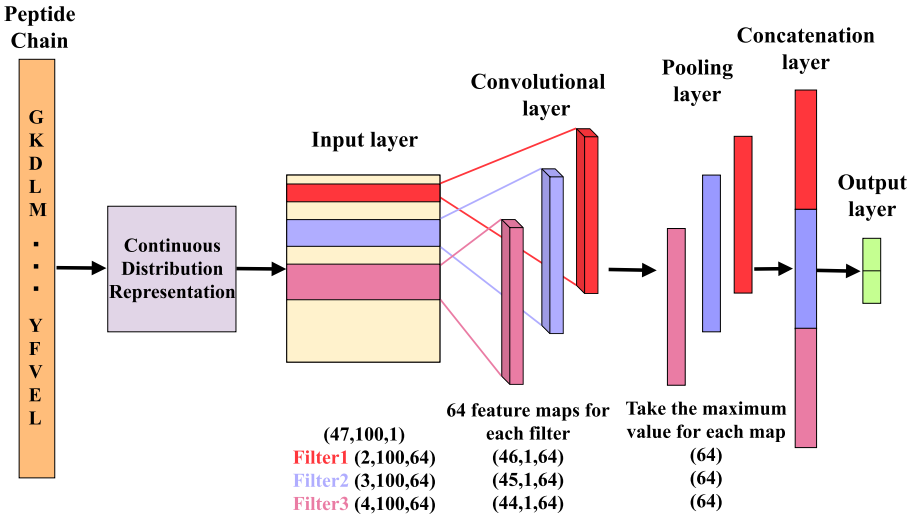


FIGURE 4. The architecture of the proposed CNN. For each sample, the input is a matrix with a shape of $47 * 100$. As the sizes of the selected convolution kernels are $2 * 100$, $3 * 100$, $4 * 100$, respectively, and 64 convolution kernels are operated in each layer, the shapes of the obtained feature maps in convolutional layer are $46 * 64$, $45 * 64$, $44 * 64$, respectively. After the maximum pooling operation, 64 features are obtained in the pooling layer for each feature map, and these features are concatenated to input the output layer.

where $BWVec(i)$ represents the continuous distribution representation of the i-th biological word along a given peptide chain.

D. CONVOLUTIONAL NEURAL NETWORK CONSTRUCTION

The task of image processing has some similarities to peptide classification. In the image processing, an image is usually represented as a pixel matrix. In the peptide classification, a peptide is also transformed into a matrix by continuous distribution representation. In both cases, we are trying to recognize an object within a large context. This suggests CNN that has successful applied in image processing can be adapted to work for peptide classification. Using the feature vectors encoded by continuous distribution representation as input, a CNN will be constructed for the first time to identify glycation sites.

As shown in Figure 4, for a given peptide chain, the input of the CNN is a two-dimensional array encoded by continuous distribution representation. As the feature map of the biological word cannot be divided, the convolution kernel in the convolution layer has the same length as the feature map of the biological word.

Suppose the kernel size is $m * n$ and $x_{i:i+m-1}$ denotes the i-th row to $i+m-1$-th row of the input data, a new feature $conv_{j,i}$ can be obtained by convolving the kernel with $x_{i:i+m-1}$ as follows.

$$conv_{j,i} = f\left(W_j \otimes x_{i:i+m-1} + b\right), \tag{5}$$

where $\otimes$ represents the convolution operator; j is the number of kernels; W_j denotes the parameters of kernel; b is the bias term; and f denotes the rectification linear unit (ReLU) activation function [47]. Through performing the above operations, feature maps is generated for different convolution kernels, and then filtered by the max-pooling method in the pooling layer. The advantage of this method is that it preserves the most important part of each feature map and greatly reduces the computational complexity of the model. Finally, the pooling results for each feature map are merged in the concatenation layer. The softmax function is adopted to generate the model output.

Due to the large number of hyper-parameters in the process of CNN construction, it is impractical to search optimal values of all parameters by the method of exhaustion. Thus, the proposed CNN is firstly constructed with hyper-parameters chosen from a wide range, and then the search range around the best performing parameter is narrowed gradually. After many iterations, the hyper-parameters that yield the minimum softmax cross entropy loss are selected as the optimal hyper-parameters.

E. PERFORMANCE EVALUATION INDEXES

The performance of glycation site predictors is systematically measured by the following indexes, sensitivity (SN), specificity (SP), accuracy (ACC), Matthews correlation coefficient (MCC), and area under the receiver operating characteristic curve (AUC). The first 4 indexes are defined as follows.

$$SN = \frac{TP}{TP + FN}, \tag{6}$$

$$SP = \frac{TN}{TN + FP}, \tag{7}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}, \tag{8}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}, \tag{9}$$

TABLE 2. Hyper-parameters of LSTM RNNS.

Hyper-parameter	Value
Input Length	50
Batch size Full	batch
Layers	2
Dropout	(0.2, 0.4)
LSTM blocks	128
Fully connected layer units	21
Regularization	L2
Learning rate	0.01
Optimizer	Adam

where *TP*, *FP*, *TN* and *FN* represent the numbers of true positive, false positive, true negative, and false negative, respectively.

Receiver operating characteristic (*ROC*) curve is a graphical plot of the true positive rate versus the false positive rate under different discrimination thresholds [48]. Across all possible decision thresholds, *AUC* summarizes a model's performance with values ranging from 0 to 1. The higher the *AUC* value, the better the prediction performance.

III. RESULTS AND DISCUSSIONS

A. THE PERFORMANCE OF LSTM RNNS

After many rounds of iterative hyper-parameter selection, the optimal hyper-parameters of the LSTM RNNs for artificial peptide chain generation are listed in Table 2. Note that *Adam* is chosen as the optimizer to compute different and adaptive learning rates for each parameter using a batch size of full batch for an initial learning rate of 0.01. The loss function is penalized with a $L2$-norm of the model parameters to prevent overfitting.

With the hyper-parameters given in Table 2, the proposed LSTM RNNs tend to be completely stable after 2000 iterations. But at this point the network has a risk of overfitting. In order to avoid this, the training of the network is terminated at 1000 iterations. The sampling process starts with the character 'B', and ends with the last amino acid of the peptide chain to be sampled. Furthermore, to make the artificial samples consisting of glycation sites with a great probability, the amino acids in the centre positions of the sampled peptide chains are fixed to lysine (K).

Compared with the SMOTE [21] algorithm, LSTM RNNs are not only suitable for dealing with high-dimensional data but also can generate visualized peptide chains rather than a bunch of abstract feature vectors, which lays the foundation for filtering generated samples by using existing glycation site predictors. To decrease sequence redundancy, the peptide chains generated at different temperatures are selected by the CD-HIT tool [40] with a threshold of 50%. The glycation site predictors GlyNN [4] and Gly-PseAAC [6] are employed to determine whether the generated peptide chain contains the glycation site. If the generated peptide chain is predicted to contain the glycation site simultaneously by the webservers of the two predictors, it will be labeled as an artificial positive sample.

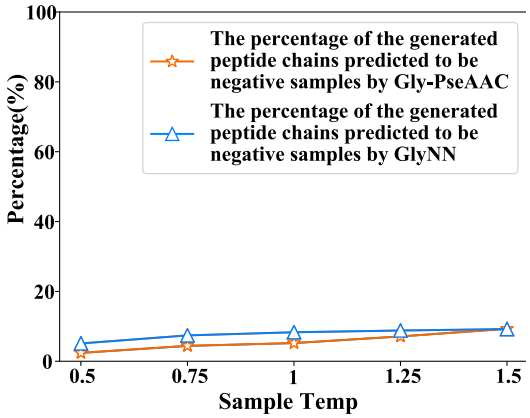


FIGURE 5. The percentage of the generated peptide chains at various temperatures predicted to be negative samples by GlyNN and Gly-PseAAC.

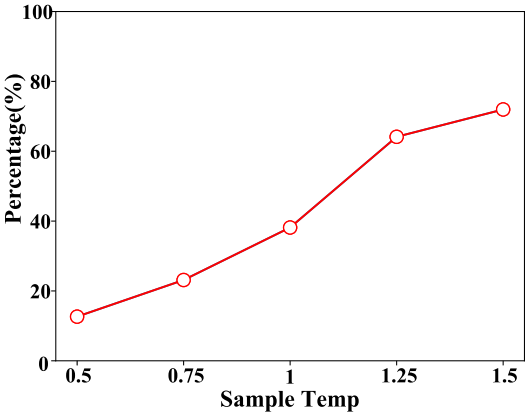


FIGURE 6. The percentage of the generated peptide chains at various temperatures selected by the CD-HIT tool.

B. ANALYSIS OF THE GENERATED PEPTIDE CHAINS

As can be seen from Figure 5, at any temperature, the percentage of the generated peptide chains predicted to be negative samples by Gly-PseAAC and GlyNN differs slightly. Figure 6 shows that when the temperature factor is 1.25, the percentage of the generated peptide chains selected by the CD-HIT tool with sequence identity less than 50% tends to converge. Thus, the peptide chain sampling by the LSTM RNNs is performed at this temperature.

Previous research [1], [4], [8] has found that glycation reactions are strongly influenced by physicochemical properties of neighboring residues that surround the glycation site. A positively charged amino acid close to the glycation site in the primary or tertiary (3D) structure often triggers the glycation reactions. As reported, the hydrophobicity of the native amino acids is adopted to develop feature extraction methods and achieves satisfactory results for glycation site prediction. Furthermore, the biological function of a protein is dependent on its amino acid compositions [49]. To validate the effectiveness of the generated peptide chains, here we present systematic comparisons between the actual peptide chains given in the dataset C and the generated peptide chains on amino acid composition, global charge and hydrophobicity [50].

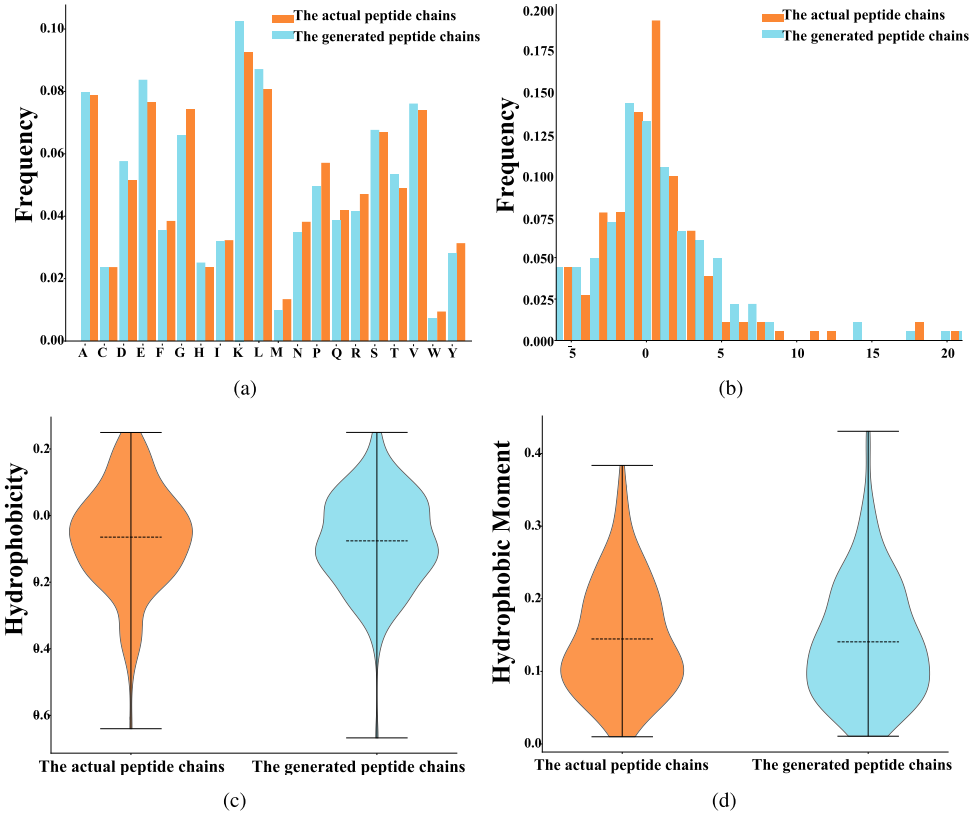


FIGURE 7. Comparisons between the actual peptide chains and the generated peptide chains on amino acid composition, global charge, hydrophobicity and hydrophobic moment. (a) The overall frequencies of the 20 native amino acids for the actual peptide chains and the generated peptide chains. (b) The relative frequency of global charged amino acids at each position around the actual peptide chains and the generated peptide chains. (c) The violin plots for the hydrophobicity distributions of the actual peptide chains and the generated peptide chains. (d) The violin plots for the hydrophobic moment distributions of the actual peptide chains and the generated peptide chains.

TABLE 3. The statistical significance test results between the generated sample and the original sample.

Characteristic	P-value
Amino acid composition	0.500
Global charge	0.071
Hydrophobicity	0.251
Hydrophobic moment	0.326

Figure 7(a) clearly shows that the overall frequencies of the 20 native amino acids for the actual peptide chains and the generated peptide chains are almost identical. The relative frequency of global charged amino acids at each position around the actual peptide chains and the generated peptide chains as displayed in Figure 7(b) approximately obey the same distribution. In Figure 7(c) and Figure 7(d), the actual peptide chains and the generated peptide chains have similar shapes for the distributions of amino acid hydrophobicity. These results indicate that the generated peptide chains effectively match the actual peptide chains, which will be beneficial to enhance the dataset quality, and then improve the performance of glycation site prediction.

TABLE 4. Hyper-parameters of the proposed CNN.

Hyper-parameter	Value
Input Length	4700
Batch size	Full batch
Convolution blocks	([2,3,4], 64, ReLU)
Fully connected layer units	128
Cutoff	0.5
Regularization	L2
Learning rate	0.01 with decay rate 0.95
Optimizer	Adam

Next, we use the paired t-test with the significance level of $alpha = 0.05$ to perform a statistical significance test [51] between the actual peptide chains and the generated peptide chains on amino acid composition, global charge, hydrophobicity and hydrophobic moment. The results in the table below indicates that there is no significant difference between the actual peptide chains and the generated peptide chains.

C. EFFECTIVENESS OF THE LSTM RNNS

After many rounds of iterative hyper-parameter selection, the optimal hyper-parameters of the proposed CNN for glycation site prediction are listed in Table 4.

TABLE 5. Prediction results of the proposed CNN with and without the LSTM RNNs on the dataset C.

Method	k-Fold Cross-Validation	ACC (%)	SP (%)	SN (%)	AUC	MCC
Without LSTM RNNs	4 - Fold Cross-Validation	81.3±0.3	100	0	0.504±0.2	Undefined*
	6 - Fold Cross-Validation	81.3±0.1	100	0	0.507±0.4	Undefined*
	8 - Fold Cross-Validation	81.2±0.2	100	0	0.509±0.2	Undefined*
	10- Fold Cross-Validation	81.1±0.1	100	0	0.500±0.1	Undefined*
With LSTM RNNs	4 - Fold Cross-Validation	83.1±1.1	84.8±2.1	81.5±0.6	0.870±1.2	0.665±2.2
	6 - Fold Cross-Validation	89.8±0.6	88.9±0.6	90.8±1.1	0.921±0.4	0.800±1.3
	8 - Fold Cross-Validation	90.0±0.5	87.9±1.2	92.4±0.3	0.910±1.3	0.802±0.9
	10- Fold Cross-Validation	91.8±0.3	91.1±0.5	92.6±0.4	0.944±0.5	0.838±0.6

*The value of MCC cannot be calculated as the denominator in Equation 15 equals to 0.

Three methods are commonly used to evaluate the performance of a predictor, namely k-fold cross-validation test, leave-one-out cross-validation (LOOCV) test, and independent dataset test [52]. The LOOCV test is supposed to be the most rigorous one that can always yield a unique result for a given benchmark dataset [53]. However, to reduce the computational complexity, the k-fold cross-validation test is adopted to access the performance of the proposed predictor DeepGly. During the process of the k-fold cross-validation test, the dataset is divided into k subsets with roughly equal size. Each subset is taken as a testing set in turn and the other $k-1$ subsets are taken to train the predictor [54], [55]. The average performance measures over the k folds are used for performance evaluation. To avoid sampling bias, the above procedure is repeated 50 times.

To provide insights in the effectiveness of the LSTM RNNs, the prediction results of the proposed CNN with and without the LSTM RNNs on the dataset C are compared based on different k-fold cross-validations. As listed in Table 5, the proposed glycation site predictor DeepGly with LSTM RNNs and CNN yields better performance than the variant using only CNN for all k-fold crossvalidations. Taking 10-fold cross-validation as example, the LSTM RNNs improves the ACC, SN, and AUC from 81.1%, 0,and 0.500 to 91.8%, 92.6%, 0.944, respectively. Similar conclusions can be conducted for other k-fold cross-validations. It is worth noting that the SP achieved by the predictor without LSTM RNNs is as high as 100%, and the sensitivity as low as 0. This phenomenon demonstrates that the imbalanced dataset problem will lead to most of samples classified as the majority class. In Table 5, the standard deviation of each evaluation index is relatively small, which indicates that the proposed method is hardly affected by the random division of samples.

Additionally, *ROC* curves with LSTM RNNs and without LSTM RNNs for different k-fold cross-validations are depicted to further validate the effect of LSTM RNNs on the prediction performance. As shown in Figure 8, *ROC* curves without LSTM RNNs for different k-fold cross-validations fluctuation up and down around the straight line $y = x$, while *ROC* curves with LSTM RNNs are far from and above the straight line $y = x$. These results indicate that LSTM RNNs can select informative and representative data subset to achieve a relatively good prediction effect for the imbalanced dataset.

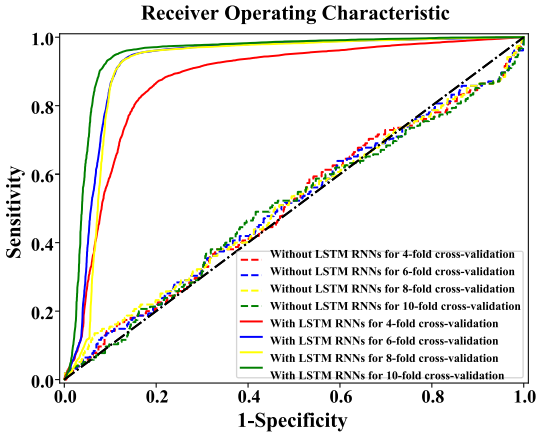


FIGURE 8. ***ROC*** **curves with LSTM RNNs and without LSTM RNNs for different k-fold cross-validations.**

D. COMPARISON BETWEEN THE PROPOSED CNN AND SUPPORT VECTOR MACHINE

In recent years, support vector machine (SVM) has been successfully applied in the field of glycation site prediction [5]–[8]. To demonstrate the powerful capacity of the proposed CNN, its prediction performance is compared with SVM using the same feature extraction method proposed in this study. To improve prediction performance of the SVM, all features are ranked according to their weights calculated by the minimum redundancy maximum correlation (mRMR) algorithm [56]. In classification, adding a new feature will simultaneously introduce useful information and redundant information. When the useful information is more than the redundant information, the prediction accuracy will be improved. On the contrary, a decrease in the prediction accuracy will happen. When the useful information and the useless information are almost the same, the prediction accuracy will behave constant. As shown in Figure 9, the feature set corresponding to top-ranking k features that gives the best prediction accuracy is chosen as the input of the SVM. Table 6 summarizes the prediction results of the proposed CNN and the SVM on dataset A and dataset B using 10-fold

TABLE 6. Prediction results of the proposed CNN and the SVM on dataset A and dataset B using 10-fold cross-validation.

Dataset	Classification Model	ACC (%)	SP (%)	SN (%)	AUC
Dataset A	CNN	90.5	96.4	78.9	0.766
	SVM	76.6	76.2	100	0.233
Dataset B	CNN	92.1	97.7	84.4	0.838
	SVM	65.2	60.3	100	0.405

TABLE 7. Performance comparisons of DeepGly and existing methods on the dataset A and dataset B.

Dataset	Method	ACC (%)	SP (%)	SN (%)	AUC	MCC
Dataset A	Gly-PseAAC[6]	68.9	74.0	58.7	0.726	0.320
	iProtGly-SS[8]	81.6	60.1	92.4	0.592	0.562
	DeepGly	90.5	96.4	78.9	0.911	0.766
Dataset B	Gly-PseAAC[6]	68.1	80.2	56.1	0.771	0.380
	PreGly[7]	85.5	95.9	71.1	——	0.700
	iProtGly-SS[8]	93.6	93.4	93.7	0.977	0.878
	DeepGly	91.8	91.1	92.6	0.944	0.838

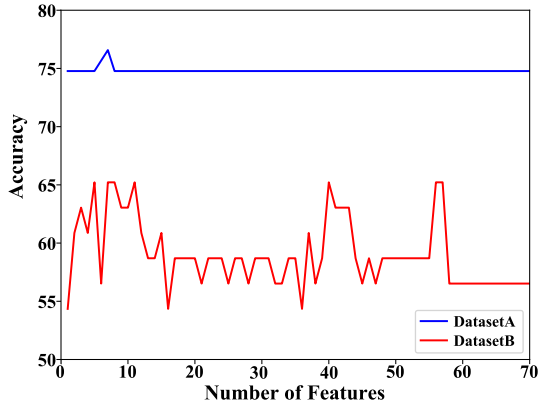


FIGURE 9. Prediction accuracy of SVM against top *k* features ranked by the mRMR algorithm on dataset A and dataset B.

cross-validation. It can be seen that the *ACC*, *SP*, *MCC* of the proposed CNN on dataset A and dataset B are significantly higher than those of the SVM, which highlights the superiority of the proposed CNN to capture complex patterns around glycation sites.

E. PERFORMANCE COMPARISONS WITH EXISTING METHODS

Various kinds of computational methods have been proposed for predicting protein glycation sites on the dataset A and dataset B. To verify the effectiveness of the proposed method DeepGly, the prediction results of DeepGly and existing methods that has been introduced in Section 1 are compared on the dataset A and dataset B. As listed in Table 7, the *ACC* yielded by Gly-PseAAC and iProtGly-SS on the dataset A are 68.9% and 81.6%, respectively, much lower than the *ACC* achieved by DeepGly. Similar conclusions can be obtained for the other performance evaluation indexes on the dataset A. Considering the *SP*, none of the other methods can perform better than DeepGly on the dataset B. In addition, all the performance evaluation indexes achieved by DeepGly on the dataset B are higher than those of existing methods except iProtGly-SS. The prediction performance of DeepGly is slightly inferior to that of iProtGly-SS, probably due to the fact that deep neural networks are better at handling large data sets while the size of dataset B is small. These observations indicate that DeepGly generally outperforms existing methods integrating multiple sources of hand-designed features.

The competitive performance of DeepGly stems from the following factors. (i) Representing peptide chains by continuous distribution representation can effectively extract the semantic information of peptide chains. (ii) The proposed LSTM-RNNs is very efficient for the imbalanced dataset problem. (iii) CNN can automatically mine the hidden high-level discriminative features, thereby reducing the process of the feature extraction and feature selection to capture complex patterns around glycation sites.

IV. CONCLUSION

As an important type of post-translational modifications, protein glycation sites are closely tied to many human diseases. Therefore, correctly identifying glycation sites could provide important clues for discovering the pathogenesis of the relevant diseases. In this study, a deep learning framework called DeepGly with recurrent and convolutional neural networks has been proposed to identify glycation sites. For the imbalanced dataset problem, LSTM RNNs that can remember the long short-term information of sequences are constructed to generate peptide chains with glycation sites. Driven by the obvious analogy between protein sequences and natural languages, the concept of continuous distribution representation is adopted to encode peptide chains. In view of the powerful ability to extract advanced features, CNN is selected as the prediction algorithm of glycation sites. This study represents the first application of deep learning in glycation site prediction. Experimental results have demonstrated the possibility of implementing LSTM RNNs to balance the dataset, the feasibility of applying continuous distribution

representation to encode peptide chains, and the effectiveness of establishing CNN to classify peptide chains. Compared with existing methods, the proposed method achieves superior performance. However, the deep learning model is prone to overfitting. To reduce the influence of this problem, the network scale of the CNN model constructed in this paper is relatively small. In addition, the scales of the datasets used in previous methods in this field are all small, and there is no standard dataset to test the degree of overfitting. In the future, the over-fitting problem of deep learning will be an important research direction. Furthermore, new and informative feature sets will be integrated into DeepGly, and different network architectures such as generative adversarial networks will be constructed to further improve the performance.

ACKNOWLEDGMENT

The author would like to thank CPLM for supplying glycation data applied in this study.

REFERENCES

[1] S. Ahmad, M. S. Khan, F. Akhter, M. S. Khan, A. Khan, J. M. Ashraf, R. P. Pandey, and U. Shahab, "Glycoxidation of biological macromolecules: A critical approach to halt the menace of glycation," *Glycobiology*, vol. 24, pp. 979–990, Jun. 2014.

[2] M. Fournet and F. Bonté, and A Desmoulière, "Glycation damage: A possible hub for major pathophysiological disorders and aging," *Aging and Disease*, vol. 9, pp. 880–900, Oct. 2018.

[3] A. Soboleva, M. Vikhnina, T. Grishina, and A. Frolov, "Probing protein glycation by chromatography and mass spectrometry: Analysis of glycation adducts," *Int. J. Mol. Sci.*, vol. 18, pp. 2557–2570, Nov. 2017.

[4] M. B. Johansen, L. Kiemer, and S. Brunak, "Analysis and prediction of mammalian protein glycation," *Glycobiology*, vol. 16, no. 9, pp. 844–853, Sep. 2006.

[5] Y. Liu, W. Gu, W. Zhang, and J. Wang, "Predict and analyze protein glycation sites with the mRMR and IFS methods," *BioMed Res. Int.*, vol. 12, Dec. 2015, Art. no. 561547.

[6] Y. Xu, L. Li, J. Ding, L.-Y. Wu, G. Mai, and F. Zhou, "Gly-PseAAC: Identifying protein lysine glycation through sequences," *Gene*, vol. 602, pp. 1–7, Feb. 2017.

[7] X. Zhao, X. Zhao, L. Bao, Y. Zhang, J. Bai, and M. Yin, "Glypre: In silico prediction of protein glycation sites by fusing multiple features and support vector machine," *Molecules*, vol. 22, pp. 1891–1906, Nov. 2017.

[8] M. M. Islam, S. Saha, M. M. Rahman, S. Shatabda, D. M. Farid, and A. Dehzangi, "iProtGly-SS: Identifying protein glycation sites using sequence and structure based features," *Proteins, Struct., Function, Bioinf.*, vol. 86, no. 7, pp. 777–789, Jul. 2018.

[9] V. P. Singh, A. Bali, N. Singh, and A. S. Jaggi, "Advanced glycation end products and diabetic complications," *Korean J. Physiol. Pharmacol.*, vol. 18, pp. 1–14, Feb. 2014.

[10] S. Y. Rhee and Y. S. Kim, "The role of advanced glycation end products in diabetic vascular complications," *Diabetes Metabolism J.*, vol. 42, pp. 188–195, Jun. 2018.

[11] P. J. Saulnier, K. M. Wheelock, S. Howell, E. J. Weil, S. K. Tanamas, W. C. Knowler, K. V. Lemley, M. Mauer, B. Yee, R. G. Nelson, and P. J. Beisswenger, "Advanced glycation end products predict loss of renal function and correlate with lesions of diabetic kidney disease in American Indians with type 2 diabetes," *Diabetes*, vol. 65, pp. 3744–3753, Dec. 2016.

[12] L. Lan, F. Han, X. Lang, and J. Chen, "Monocyte chemotactic protein-1, fractalkine, and receptor for advanced glycation end products in different pathological types of lupus nephritis and their value in different treatment prognoses," *PLoS ONE*, vol. 11, Jul. 2016, Art. no. e0159964.

[13] M. Kosmopoulos, D. Drekolias, P. D. Zavras, C. Piperi, and A. G. Papavassiliou, "Impact of advanced glycation end products (AGEs) signaling in coronary artery disease," *Biochim. Biophys. Acta (BBA)-Mol. Basis Disease*, vol. 1865, pp. 611–619, Mar. 2019.

[14] R. López-Díez, A. Shekhtman, R. Ramasamy, and A. M. Schmidt, "Cellular mechanisms and consequences of glycation in atherosclerosis and obesity," *Biochim. Biophys. Acta (BBA)-Mol. Basis Disease*, vol. 1862, no. 12, pp. 2244–2252, Dec. 2016.

[15] L. A. Moemen, A. M. Mahmoud, A. M. Mostafa, F. Ghaleb, M. A. Aziz, M. A. Abdelhamid, M. Y. Farrag, and I. A. Fahmy, "The relation between advanced glycation end products and cataractogensis in diabetics," *World J. Med. Sci.*, vol. 10, no. 4, pp. 368–374, 2014.

[16] T. Holm, C. T. Raghavan, R. Nahomi, R. H Nagaraj, and L. Kessel, "Effects of photobleaching on selected advanced glycation end products in the human lens," *BMC Res. Notes*, vol. 8, Jan. 2015, Art. no. 5.

[17] S. Y. Ko, H.-A. Ko, K.-H. Chu, T.-M. Shieh, T.-C. Chi, H.-I. Chen, W.-C. Chang, and S.-S. Chang, "The possible mechanism of advanced glycation end products (AGEs) for Alzheimer's disease," *PLoS ONE*, vol. 10, pp. 210–225, Nov. 2015.

[18] H. M. Vicente, O. M. El-Agnaf, and T. F. Outeiro, "Glycation in parkinson's disease and Alzheimer's disease," *Movement Disorders*, vol. 31, no. 6, pp. 782–790, Jun. 2016.

[19] 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, pp. 398–406, Feb. 2019.

[20] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, p. 27, Dec. 2019.

[21] 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.

[22] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, pp. 1–16, Mar. 2013.

[23] M. Woźniak and D. Połap, "Object detection and recognition via clustered features," *Neurocomputing*, vol. 320, pp. 76–84, Dec. 2018.

[24] A. Winnicka, K. Kesik, D. Polap, M. Wozniak, and Z. Marszalek, "A multi-agent gamification system for managing smart homes," *Sensors*, vol. 19, p. 1249, Jan. 2019.

[25] F. Beritelli, G. Capizzi, G. L. Sciuto, C. Napoli, and M. Woźniak, "A novel training method to preserve generalization of RBPNN classifiers applied to ECG signals diagnosis," *Neural Netw.*, vol. 108, pp. 331–338, Dec. 2018.

[26] T. Mikolov, M. Karafiát, L. Burget, J. H. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 1–4.

[27] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu, "Texygen: A benchmarking platform for text generation models," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 1097–1100.

[28] Y. S. Yao and Z. Huang, "Bi-directional LSTM recurrent neural network for Chinese word segmentation," in *Proc. Int. Conf. Neural Inf. Process.*, 2016, pp. 345–353.

[29] X. Huang, H. Tan, G. Lin, and Y. Tian, "A LSTM-based bidirectional translation model for optimizing rare words and terminologies," in *Proc. Int. Conf. Artif. Intell. Big Data*, May 2018, pp. 185–189.

[30] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," in *Proc. Comput. Sci.*, 2015, pp. 1–38.

[31] A. T. Müller, J. A. Hiss and G. Schneider, "Recurrent neural network model for constructive peptide design," *J. Chem. Inf. Model.*, vol. 58, pp. 472–479, Jan. 2018.

[32] E. Asgari and M. R. K. Mofrad, "Continuous distributed representation of biological sequences for deep proteomics and genomics," *PLoS ONE*, vol. 10, pp. 1–15, Nov. 2015.

[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

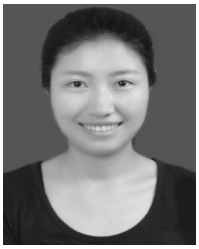
[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Represent.*, 2013, pp. 1–12.

[35] N. Poerner, B. Roth, and H. Schütze, "Interpretable textual neuron representations for NLP," in *Proc. Assoc. Comput. Linguistics*, 2018, pp. 325–327.

[36] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[37] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," in *Proc. Int. Conf. Commun. Signal Process.*, Apr. 2017, pp. 1–6.

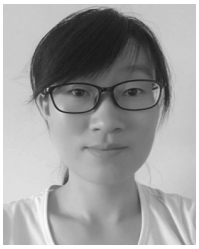
[38] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1746–1751.
[39] H. L. Fu, Y. Yang, X. Wang, H. Wang, and Y. Xu, "DeepUbi: A deep learning framework for prediction of ubiquitination sites in proteins," *BMC Bioinf.*, vol. 20, pp. 37–45, Feb. 2019.
[40] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006.
[41] Z. Liu, Y. Wang, T. Gao, Z. Pan, H. Cheng, Q. Yang, Z. Cheng, A. Guo, J. Ren, and Y. Xue, "CPLM: A database of protein lysine modifications," *Nucleic Acids Res.*, vol. 42, no. 1, pp. 531–536, Jan. 2014
[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.
[43] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.
[44] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, "Long short term memory hyperparameter optimization for a neural network based emotion recognition framework," *IEEE Access*, vol. 6, pp. 49325–49339, 2018.
[45] P. Reverdy and N. E. Leonard, "Parameter estimation in softmax decision-making models with linear objective functions," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 1, pp. 54–67, Jan. 2016.
[46] Y. Goldberg and O. Levy, "Word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method," Feb. 2014, *arXiv:1402.3722*. [Online]. Available: https://arxiv.org/abs/1402.3722
[47] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.
[48] Y. Li, L.-P. Li, L. Wang, C.-Q. Yu, Z. Wang, and Z.-H. You, "An ensemble classifier to predict protein–protein interactions by combining PSSM-based evolutionary information with local binary pattern model," *Int. J. Mol. Sci.*, vol. 20, p. 3511, Jul. 2019.
[49] S. M. Cascarina and E. D. Ross, "Proteome-scale relationships between local amino acid composition and protein fates and functions," *PLoS Comput. Biol.*, vol. 14, Sep. 2018, Art. no. e1006256.
[50] A. T. Müller, G. Gabernet, J. A. Hiss, and G. Schneider, "modlAMP: Python for antimicrobial peptides," *Bioinformatics*, vol. 33, no. 1, pp. 2753–2755, Sep. 2017.
[51] R. Dror, G. Baumer, S. Shlomov, and R. Reichart, "The hitchhiker's guide to testing statistical significance in natural language processing," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, Jul. 2018, pp. 1383–1392.
[52] A. Schoenrock, D. Burnside, H. Moteshareie, S. Pitre, M. Hooshyar, J. R. Green, A. Golshani, F. Dehne, and A. Wong, "Evolution of protein-protein interaction networks in yeast," *PLoS ONE*, vol. 12, Mar. 2017, Art. no. e0171920.
[53] M. Balachandran, T. H. Shin, and L. Gwang, "PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine," *Frontiers Microbiol.*, vol. 9, pp. 476–488, Mar. 2018.
[54] J. Chen, H. Xu, P.-A. He, Q. Dai, and Y. Yao, "A multiple information fusion method for predicting subcellular locations of two different types of bacterial protein simultaneously," *Biosystems*, vol. 139, pp. 37–45, Jan. 2016.
[55] H. Wang and X. Hu, "Accurate prediction of nuclear receptors with conjoint triad feature," *BMC Bioinformatics*, vol. 16, pp. 402–415, Dec. 2015.
[56] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.



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 current research interests include bioinformatics and machine learning.

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 current research interests include swarm intelligence robotics, bioinformatics, and system biology.

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 current research interests include control theory and applications, intelligent robot control, and bioinformatics.



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. degree from the School of Control Science and Engineering, Shandong University, in 2012, and the Ph.D. degree in control science and engineering from Shandong University, in 2017. 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 mathematical modeling of molecular biology.

QIAN ZHANG was born in Heze, 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 M.S. degree in control science and engineering from Shandong University, in 2014. He is currently an Assistant Research Fellow with the Heze Institute of Science and Technology Information. His current research interests include information analysis, assessment, and prediction.