Computers in Biology and Medicine 184 (2025) 109328

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

WGAN-GP_Glu: A semi-supervised model based on double generator-Wasserstein GAN with gradient penalty algorithm for glutarylation site identification

Qiao Ning [a,b,c,d,*], Zedong Qi [a]

[a] Information Science and Technology, Dalian Maritime University, Dalian, Liaoning, China
[b] The School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
[c] Neusoft Education Technology Group, Dalian, China
[d] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, 130012, China

ARTICLE INFO

Keywords:
Semi-supervised model
Negative sampling
Multiple generators
Glutarylation site identification

ABSTRACT

As an important post-translational modification, glutarylation plays a crucial role in a variety of cellular functions. Recently, diverse computational methods for glutarylation site identification have been proposed. However, the class imbalance problem due to data noise and uncertainty of non-glutarylation sites remains a great challenge. In this article, we propose a novel semi-supervised learning algorithm, called WGAN-GP_Glu, for identifying reliable non-glutarylation lysine sites from those without glutarylation annotation. WGAN-GP_Glu method is a multi-module framework algorithm, which mainly includes a reliable negative sample selection module, a deep feature extraction module, and a glutarylation site prediction module. In reliable negative sample selection module, we design an improved method of Wasserstein GAN with Gradient Penalty (WGAN-GP), named ReliableWGAN-GP, including three parts, two generators G1, G2 and a discriminator D, which can select reliable non-glutarylation samples from a great number of unlabeled samples. Generator G1 is utilized to generate noise data from unlabeled samples. For generator G2, both the positive sample and the noise data are used as inputs to improve the discriminant capability of discriminator D. Then, convolutional neural network and bidirectional long short-term memory network combined with attention mechanism are utilized to extract deep features for glutarylation samples and reliable non-glutarylation samples. Finally, a glutarylation site prediction module based on the three-layer fully connected layer is designed to make class predictions for samples. The sensitivity, specificity, accuracy and Matthew correlation coefficient of WGAN-GP_Glu on the independent test data set reach 90.58 %, 95.82 %, 94.44 % and 0.8645, respectively, which surpassed the existing methods for glutarylation sites prediction. Therefore, WGAN-GP_Glu can serve as a powerful tool in identifying glutarylation sites and the ReliableWGAN-GP algorithm is effective in selecting reliable negative samples. The data and code are available at https://github.com/xbbxhbc/WGAN-GP_Glu.git.

1. Introduction

As the main active component of organisms, biological macromolecules are carriers of biological information, mainly including proteins, nucleic acids and high relative molecular weight hydrocarbons. Among them, protein is the material basis of all life and an important part of the body's cells. Proteins are composed of biomolecules called amino acids, which are composed of amine (-amino), carboxyl (-carboxy) functional group, and hydrocarbon R group. In protein formation, messenger RNA (mRNA) and transfer RNA (tRNA) generate peptide chains through the action of ribosomes. At this time, the peptide chains are precursor proteins, which do not have protein activity and require post-processing and post-translation modification (PTM) to become functional proteins [1]. Protein post-translational modification sites (PTMs) are related to diversiform biological processes [2], which are significant in regulating functions such as protein activity, localization and protein interaction in a variety of biological and physiological interactions. Therefore, the accurate prediction of protein PTMs is the main direction of current research.

Lysine is an essential amino acid, that enhances immunity and

* Corresponding author. The School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China.
E-mail address: ningq669@dlmu.edu.cn (Q. Ning).

https://doi.org/10.1016/j.compbiomed.2024.109328
Received 2 January 2024; Received in revised form 22 October 2024; Accepted 23 October 2024
Available online 14 November 2024
0010-4825/© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

improves central nervous system function [3]. It is a basic amino acid, whose side chain e-NH2 is extremely unstable, resulting in strong nucleophilic activity, so it is very prone to post-translational modifications, including acetylation [4], propionation [5], sulfonylation [6] and methylation [7]. Recently, glutarylation, a new lysine acylation discovered in eukaryotic and prokaryotic cells [8], plays a significant role in chromatin structure, regulating nucleosome assembly, DNA damage repair, gene expression and cell cycle. However, traditional sequencing technology to identify glutarylation sites has many disadvantages [9,10], such as expensive biological materials and long time consumption. Therefore, computational technology to develop new prediction methods is an effective way to quickly identify potential modification sites. The Glutpred [11] is based on multi-feature extraction and biased support vector machine algorithms for glutarylation sites identification. Dou et al. designed the iGlu Adaboost model based on features including k-spaced amino acid pair composition (CKSAAP) and enhanced amino acid composition to identify lysine glutarylation site modifications [12]. Chi2-based increment feature selection strategy (IFS) was applied to select the optimal feature set, and AdaBoost was used as the classifier [13]. The RF-GlutarySite prediction method [14] combined features extracted by the FEPS web server with binary encoding (Binary encoding), amino acid factor (AAF) and amino acid index (AAindex) as new feature representation. Recently, Arafat et al. proposed BiPepGlut, a machine learning method based on the concept of dipeptide-based evolution method for feature extraction [15].

Despite many efforts to date, the performance for lysine glutarylation site prediction remains to be improved. The main challenge in the task of predicting glutarylation sites is data imbalance. The label of the glutarylated lysine samples is confirmed while the label of the other lysine sites is not available [16]. In this case, positive-unlabeled learning (PU) methods are proposed for model construction with positive samples and unlabeled samples [17]. Therefore, a semi-supervised model based on the WGAN-GP algorithm [18] is proposed for glutarylation site identification, named WGAN-GP_Glu, whose framework is shown in Fig. 1.

First, glutarylation site samples and unlabeled samples are characterized and the ReliableWGAN-GP algorithm is proposed for reliable negative samples selection, which consists of two generators and one discriminator. Then, a deep feature extraction layer is built for deep feature extraction based on convolutional neural network (CNN), bidirectional long-short term memory network (Bi-LSTM), and attention mechanism. Next, the fully connected network is built for glutarylation site prediction. Finally, we evaluated the model in various aspects. The contributions are summarized as follows.

1. We propose an improved method of WGAN-GP named ReliableWGAN-GP, including two generators G1, G2 and discriminator D, to solve the data imbalance problem when modeling.
2. The WGAN-GP_Glu is a semi-supervised learning algorithm integrating the ReliableWGAN-GP algorithm and multi-view feature encoding scheme, greatly improving the prediction performance of glutarylation sites.
3. CNN and Bi-LSTM combined with attention mechanism are utilized to extract deep features and direction-dependent features for glutarylation samples and reliable non-glutarylation samples.

2. Methods

2.1. Data preparation

The Protein Lysine modification Database (PLMD) [19] is a protein post-translational modification database that contains up to 20 types of PTMs on lysine. Glutarylation proteins are collected from PLMD and after removing proteins with more than 40 % identity by the CD-HIT program [20,21], and we obtain 187 glutarylation proteins with 644 glutarylation sites. Then, 20 proteins are randomly selected as independent test set and the rest proteins as training dataset. Finally, the independent test set includes 54 glutarylation sites and 122 non-glutarylation annotation lysine sites, and the training dataset

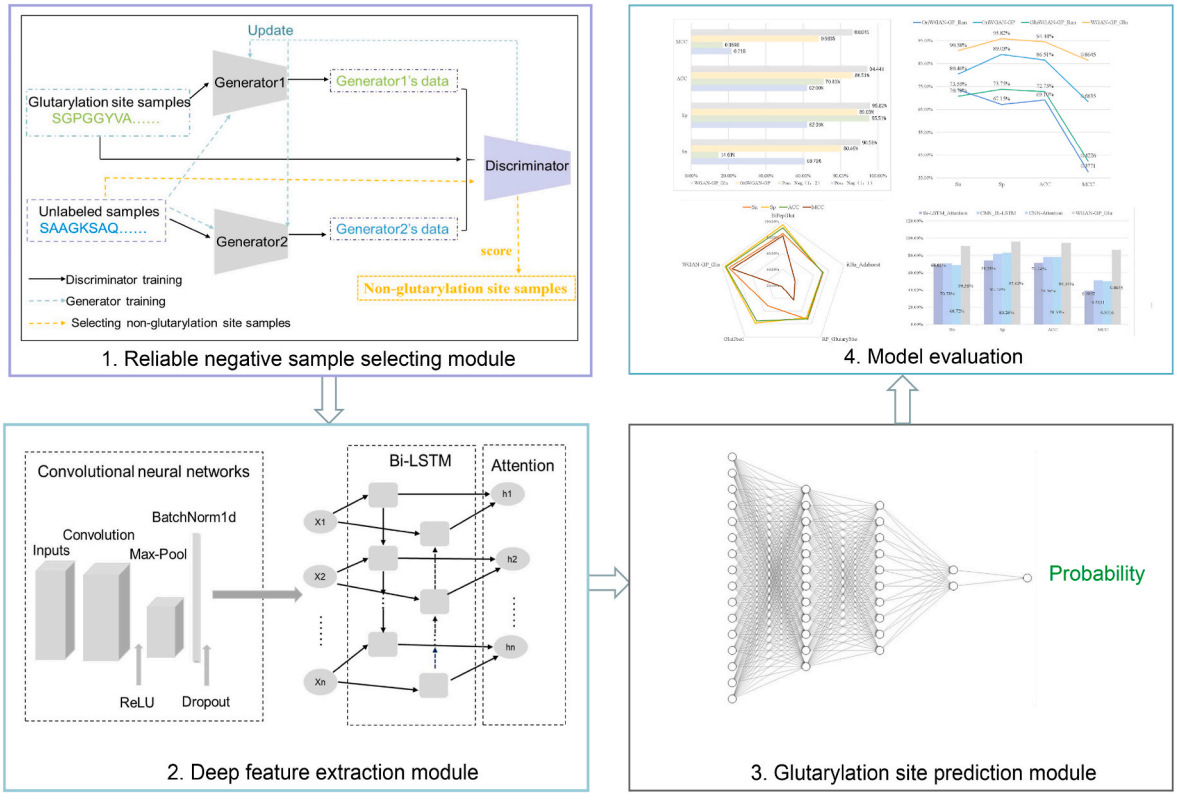


Fig. 1. Flowchart of WGAN-GP_Glu method.

contains 590 glutarylation sites and 3498 non-glutarylation annotation lysine sites. According to the compositional difference around glutarylation lysine sites compared to non-glutarylation annotation lysine sites generated by Two Sample Logo [22] shown in Fig. 2, we adopt 35 as the window size, which is consistent with [11]. To ensure the uniformity of the length of each sample, fill in 'X' for the vacant position [11].

2.2. Multi-view feature encoding scheme

Suitable feature coding scheme will significantly impact on the prediction results. Therefore, we extract features from the perspective of multiple views in order to comprehensively characterize glutarylation sites, including amino acid composition, composition of k-spaced amino acid pairs, BLOSUM62, amino acid factor.

2.2.1. Amino acid composition (AAC)

Amino acid composition (AAC) counts the frequency of occurrence of each amino acid type (i.e. "ACDEFGHIKLMNPQRSTVWY") in samples [23]:

$$f(t)=\frac{N(a)}{L}, a \in \{A, C, D, \ldots, Y\} \tag{1}$$

where $N(a)$ is the number of a-type amino acids, and L is the length of a peptide sequence.

2.2.2. Composition of k-spaced amino acid pairs (CKSAAP)

CKSAAP calculates the frequency of K-spacing amino acid pairs in peptides [24], reflecting short linear motif information in fragments. An amino acid pair with an interval of k means that they are separated by any k amino acids. In this study, $k = 0$ is adopted, encoding each sample into a 400-dimensional CKSAAP feature vector.

2.2.3. BLOSUM62

In this study, the BLOSUM62 matrix represents protein first-order sequence information as the basic feature set [25]. Each residue in the sample is represented by a $m \times n$ matrix, where n represents the peptide length and $m = 20$, representing 20 amino acids.

2.2.4. Amino acid factor (AAF)

The AAindex database records a variety of physicochemical properties for amino acids. Through multivariate statistical analysis, five types of physicochemical properties are selected from the AAIndex database: secondary structure, molecular volume, the attributes reflected the polarity, electrostatic charge and codon diversity [25], which are called amino acid factors (AAF).

2.3. Feature selection

Accurate identification of modification sites often relies on limited individual characteristics. Multi-perspective features are often redundant or even noisy, which can lead to negative effects. Therefore, in this study, the incremental feature selection (IFS) algorithm combined with the Gini Index is employed to select the optimal feature set [26]. At each round of IFS, the feature with the highest Gini Index-score was added to the model and was deleted from the feature set until the feature set was empty. In the end, feature subset with the highest performance was selected.

2.4. ReliableWGAN-GP algorithm for selecting reliable negative samples

In the prediction of glutarylation sites, due to the limitations of biological experimental techniques, non-glutarylation site data cannot be determined, so only reliable positive samples are available, but negative samples are indefinite, named unlabeled samples. Facing this problem, traditional supervised learning methods are often limited by model generalization ability. Therefore, it is necessary to develop a semi-supervised learning approach to improve the performance with unlabeled samples.

In this study, we propose ReliableWGAN-GP algorithm to filter reliable negative samples from samples without glutarylation annotations and at the same time address the imbalance problem. ReliableWGAN-GP mainly includes two generators G1, G2 and discriminator D according to the Wasserstein GAN with Gradient Penalty algorithm.

2.4.1. Generative adversarial network

The essence of generative adversarial network (GAN) is to generate a dynamic minimax game between generative network and discriminative network [27]. With a database, the task of the discriminator is to determine whether a sample belongs to it, while the purpose of the generator is to generate control samples according to the distribution of data in the database. The training goal of discriminant network D is to improve the ability to distinguish between real data and control samples, while the training goal of generative network G is to generate control samples that are difficult to distinguish. This translates mathematically to:

$$\min_G \max_D V(D,G) = E_{x\sim P_{data}}[logD(x)] + E_{z\sim P_z}[log(1-D(G(z)))] \tag{2}$$

Although GAN performs very well, it is limited by the instability of the loss function and the convergence difficulty.

2.4.2. Wasserstein GAN with gradient penalty

Wasserstein GAN with Gradient Penalty (WGAN-GP) can solve the learning instability problem of GAN [18]. Different from the GAN, WGAN-GP utilizes gradient penalty to achieve uniform weight distribution and fully exploit the learning ability of neural networks. WGAN-GP optimizes the Wasserstein distance between real distribution P_r and generator distribution P_g, providing a better gradient and more stable convergence for the training stage. The Wasserstein distance can be calculated as follows:

$$W\left(p_r, p_g\right) = \sup_{\|f\|_L \le 1} E_{x\sim P_r}[f(x)] - E_{x\sim P_g}[f(x)] \tag{3}$$

The generator G and discriminator D optimization functions of WGAN-GP are defined as:

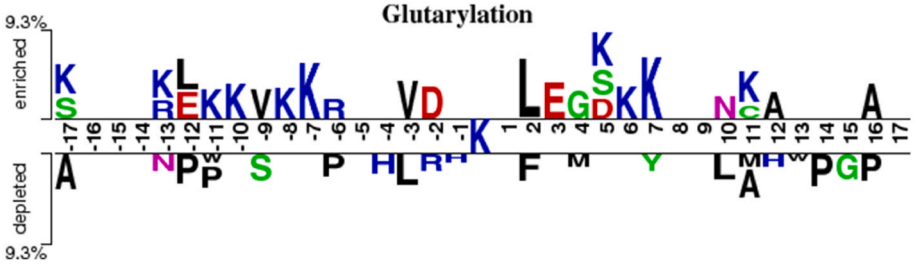


Fig. 2. The difference of amino acid composition between glutarylation site and non-glutarylation site, generated by Two Sample Logos.

$$G = \underset{G}{argmin} E_{\tilde{x}\sim P_g}[D(\tilde{x})] - E_{x\sim P_r}[D(x)] \quad (4)$$

$$L = \underset{D}{argmax} E_{\tilde{x}\sim P_g}[D(\tilde{x})] - E_{x\sim P_r}[D(x)] + \lambda E_{\hat{x}\sim P_g}\left[\left(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\right)^2\right] \quad (5)$$

For the discriminator D, its gradient penalty term is:

$$\lambda\lambda E_{\hat{x}\sim P_{\hat{x}}}\left[\left(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\right)^2\right] \quad (6)$$

where $\tilde{x} = G(x)$, λ represents the penalty term coefficient, $\hat{x} = \varepsilon x + (1-\varepsilon)\tilde{x}, 0 < \varepsilon < 1$, and $P_{\hat{x}}$ represents the distribution of sampled data.

2.4.3. ReliableWGAN-GP algorithm

The aim of PU learning is to take full advantage of labeled data and unlabeled data. In identification of glutarylation sites, the vast majority of samples are unlabeled, and selecting reliable negative samples from unlabeled data based on the existing positive samples is urgently needed. Therefore, ReliableWGAN-GP, an improved WGAN-GP method for reliable negative samples selection, is proposed. The architecture of ReliableWGAN-GP algorithm is shown in Fig. 3. ReliableWGAN-GP is made up of two generators G1, G2 and a discriminator D. The structure of G1 and G2 is the same, as shown in Fig. 4, including 5 fully connected layers. The composition of the discriminator is similar to that of the generator, shown in Fig. 5, including four layers of full connection layer and a sigmoid activation output layer. The full connection layer of four layers adopts the ReLU activation function. The ReliableWGAN-GP algorithm employs an iterative loop training generator and a discriminator, and includes two main training phases.

1 Discriminator training phase. Firstly, the positive samples, the samples generated from generator G1, and the unlabeled samples generated from generator G2 are input into the discriminator D for training learning. The purpose of the positive samples and the samples from the generator G1 is to improve the discrimination ability of discriminator D, while the generated unlabeled samples from generator G2 is used as noise data to improve the generalization ability of discriminator D.
2 Generator training phase. A large number of unlabeled datasets are utilized to train the generator G1, generating noise data. Then, both positive samples and the noise data are input into Generator G2 to generate new data. Then, the generated samples are identified by the discriminator D. The difference between the generated data and the original data is calculated according to the loss function to optimize parameters of the generator. The difference between the discriminant results and real labels is used to optimize the discriminator parameters.

Through the iterative cycle training of the above two stages, the generator and the discriminator fight against each other and constantly improve the generalization ability of the discriminator. Finally, the generator and discriminator are trained and the discriminator is used for reliable negative samples selection from unlabeled samples.

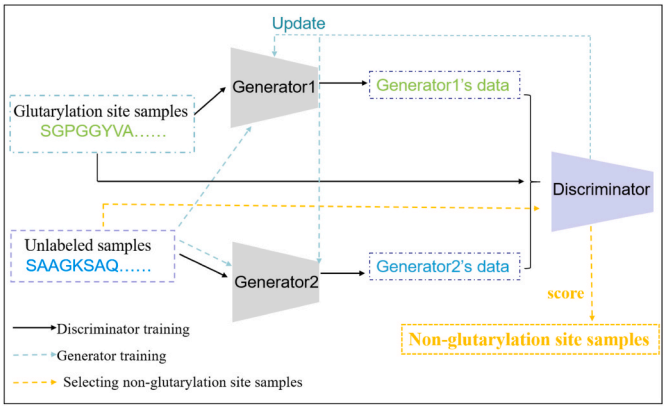


Fig. 3. The ReliableWGAN-GP algorithm selecting reliable negative samples.

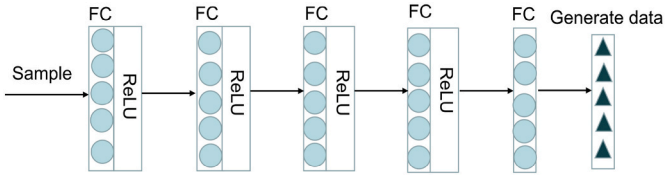


Fig. 4. The architecture of the generator network.

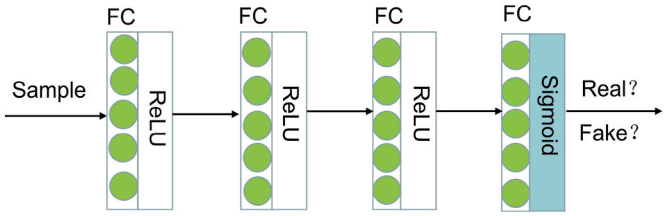


Fig. 5. The architecture of the discriminator network.

2.5. Deep feature extraction module

To better capture protein sequence information, the features of reliable negative samples and positive samples are deeply extracted, and the deep feature extraction module mainly includes the convolutional neural network layer [28], Bi-LSTM layer [29], and attention layer [30,31], shown in Fig. 6. The CNN layer extracts the hidden information of the feature sequence. Bi-LSTM learns the sequence information of the feature, better extracting the dependency of the feature sequence. The attention layer identifies the key information of the feature information. The CNN layer is made up of the convolutional layer, the pooling layer, the BatchNorm1d layer and the Dropout layer. ReLU is employed as the activation function, and a pooling layer is added to minimize feature redundancy. At the same time, the pooling layer is immediately followed by the BatchNorm1d layer for normalization, and the Dropout layer can avoid overfitting. A hidden layer is added after the convolutional neural network layer to learn the sequence information of the features, and then an attention layer is added after the Bi-LSTM layer to identify the key information from features.

2.6. Glutarylation site prediction layer

Finally, the WGAN-GP_Glu model adopts three fully connected layers to predict glutarylation sites. The binary cross-entropy loss function is employed to measure the difference between the true label and the predicted label:

$$L(w) = -\sum_{i=1}^{N} y_i \log(y_i') + (1-y_i)\log(1-y_i') \quad (7)$$

where y_i is the true label and y_i' is the predicted label of the WGAN-GP_Glu model.

3. Results and analysis

3.1. Performance evaluation strategies

In this study, four metrics are utilized for performance evaluation, including sensitivity (Sn), specificity (Sp), accuracy (ACC), precision and Matthew's correlation coefficient (MCC), which are defined as

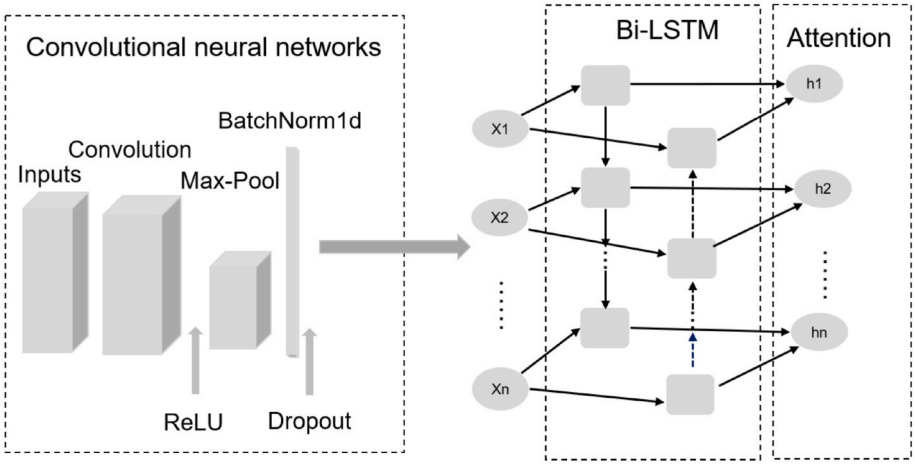


Fig. 6. Schematic diagram of deep feature extraction module.

follows:

$$S_n = \frac{TP}{TP + FN} \tag{8}$$

$$S_p = \frac{TN}{TN + FP} \tag{9}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP})}} \tag{12}$$

in which *TP*, *FP*, *TN* and *FN* represent the number of true positive samples, false positive samples, true negative samples and false negative samples, respectively.

3.2. Comparison with state-of-the-art models

To further evaluate the performance of the WGAN-GP_Glu model, it is compared with four state-of-the-art methods, including BiPepGlut, IGlu Adaboost, RFGlutarySite and GlutPred on the same independent test set. In these comparative models, reliable negative samples in training stage are selected with ReliableWGAN-GP from unlabeled samples, and then depth feature extraction is carried out through the depth feature extraction module. Therefore, all these methods are trained on the similar dataset. From Fig. 7, it can be observed that WGAN-GP_Glu based on the PU learning algorithm has the highest value among all metrics, which indicates that WGAN-GP_Glu is effective and feasible in glutarylation site prediction.

3.3. Analysis of WGAN-GP_Glu

In PU learning, the categories of samples are divided into positive samples and unlabeled samples, with no specific label, may be positive or negative. Therefore, we proposed WGAN-GP_Glu for reliable negative samples selection from unlabeled samples. To further verify the effectiveness of the WGAN-GP_Glu, we randomly select negative samples with 1 and 2 times the number of positive samples (positive: negative = 1:1, positive: negative = 1:2) from unlabeled samples to construct comparison models.

At the same time, we also compares another PU learning algorithm OriWGAN-GP (original WGAN-GP), which selects reliable negative samples from unlabeled samples based on the original WGAN-GP. First,

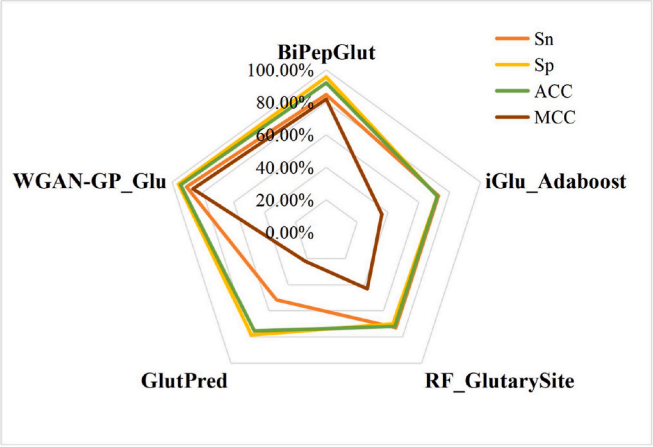


Fig. 7. Performance comparison between WGAN-GP_Glu and other State-of-the-Art models.

the discriminator is trained and learned by using positive samples as the real data set, and randomly selected unlabeled samples are input into the generator as a fake data set, and the generator does not participate in training at this stage. The second is the generator training stage. Generative adversarial training of generators and discriminators is carried out using a large number of unlabeled datasets, (1) putting unlabeled data into the generator, (2) generating new data with the generator, (3) putting generated new data into the discriminator, (4) generating the distribution gap between new data and input data according to the loss function calculation, and (5) optimizing the parameters of the generator. Through the above two stages of training in turn, the trained generator and discriminator are finally obtained.

It can be observed from Fig. 8 that Sn, Sp, ACC, Precision and MCC values of the model with Pos:Neg (1:1) on the independent test set are 60.79 %, 62.09 %, 62 %, 44.0 % and 0.216, respectively. For the model with Pos:Neg (1:2), Sn, Sp, ACC, Precision and MCC are 14.61 %, 95.51 %, 70.83 %, 61.54 % and 0.1698, respectively. The extremely unbalanced Sn and Sp values indicate that when the number of negative samples is much higher than the number of positive samples, the model learns too much negative sample information, resulting in obvious classification bias. However, for OriWGAN-GP and WGAN-GP_Glu, which utilize reliable negative samples from unlabeled samples, Sn, Sp, ACC, Precision and MCC are significantly improved. Compared with OriWGAN-GP, the Sn, Sp, ACC, Precision and MCC values of WGAN-GP_Glu increase by 10.12 %, 6.79 %, 7.93 %, 16.95 % and 0.181. The

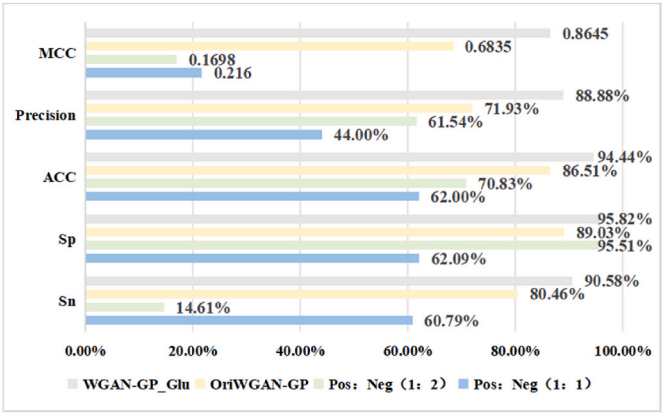


Fig. 8. WGAN-GP_Glu algorithm analysis.

Receiver Operator Characteristic (ROC) curve is shown in Fig. 9(a). The AUC values of WGAN-GP_Glu, Pos:Neg (1:1), Pos:Neg (1:2) and OriWGAN-GP are 0.9515, 0.67, 0.5733 and 0.8645, respectively, and WGAN-GP_Glu achieves the highest AUC value. The Precision Recall (PR) curve in Fig. 9(b) clearly shows that WGAN-GP_Glu outperforms other sample selection methods.

3.4. *ReliableWGAN-GP method analysis*

To better verify the validity of the reliable negative samples determined by the ReliableWGAN-GP algorithm in the WGAN-GP_Glu model, we compares OriWGAN-GP, OriWGAN-GP_Ran and WGAN-GP_Glu_Ran. WGAN-GP_Glu_Ran is mainly composed of generators G1, G2 and discriminator D, including two stages, one is the discriminator training stage. Firstly, the positive samples, the generated positive samples from generator G1 and the random data are input to the generator G2 as noise, and the false data generated by G2 are input to the discriminator D for training learning. The second is the generator training stage. A large number of random data sets of the same dimension are used to generate adversarial training for G1 generator, G2 generator and discriminator D. The random data are inputted into the generator, generating more real data. We input the generated new data to the discriminator, and calculate the distribution gap between the new data and the input data according to the loss function, and optimize the parameters of the generator. For OriWGAN-GP_Ran, which consists of a generator and a discriminator, random data of the same dimension is input to the generator as noise data to generate false data to improve the generalization ability of the discriminator.

From Table 1, it can be seen that when random data is input to the generator as noise, the Sn, Sp, ACC, Precision, and MCC of OriWGAN-GP_Ran are 63.16 %, 74.78 %, 70.83 %, 52.31 % and 0.3633, respectively, and the Sn, Sp, ACC, Precision, and MCC of WGAN-GP_Glu_Ran are 70.78 %, 73.75 %, 53.52 % and 71.75 %, respectively and 0.4226, which are both lower than the corresponding unlabeled samples entered into the generator as noise data OriWGAN-GP and WGAN-GP_Glu. Therefore, it can be seen that the unlabeled samples are input into the generator as noise data, which can improve the generalization ability of the discriminator and identify more reliable negative samples, in addition, the Sn, Sp, ACC, Precision, and MCC of WGAN-GP_Glu are 90.58 %, 95.82 %, 94.44 %, 88.88 %, and 0.8645 higher than that of OriWGAN-GP of 80.46 %, 89.03 %, 86.51 %, 71.93 % and 0.6835. The ROC curve is shown in Fig. 10 (a), with AUC values of WGAN-GP_Glu, WGAN-GP_Glu_Ran, OriWGAN-GP_Ran and OriWGAN-GP of 0.9515, 0.7481, 0.6860 and 0.8645, respectively, with WGAN-GP_Glu obtaining the highest AUC value. Comparative results indicate that the ReliableWGAN-GP can improve the ability of the model, so the ReliableWGAN-GP method is feasible and effective in glutarylation site prediction. Besides, the Precision Recall (PR) curve is shown in Fig. 10 (b), and it is clear that WGAN-GP_Glu is superior to other methods.

3.5. *The effect of iteration rounds and penalty term coefficients on the results*

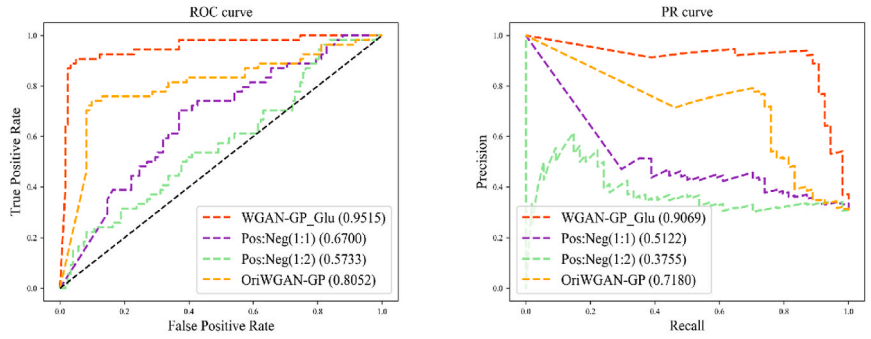
In the selection of reliable negative samples by ReliableWGAN-GP in the WGAN-GP_Glu model, the number of adversarial cycle iterations and the coefficient of penalty term are crucial, and the difference between the number of countercycle iterations and the coefficient of penalty terms will affect the quality of selected reliable negative samples. Therefore, it is significant to determine the optimal number of cycle iterations and penalty term coefficient, and the MCC value is used as a metric in this paper. The default penalty coefficient of WGAN-

GP is 10, on the basis of which the penalty coefficient increases by 2 and decreases by 2, respectively. Besides, we test the model without penalty term, where the coefficient of penalty term is 0.

From Fig. 11 can be seen that the MCC values of different penalty term coefficients are in the process of adversarial cycle iteration 0 to 1000 in the process of increasing process, reaching a maximum after 1000 rounds, and then gradually decreasing in the process of adversarial cycle iteration rounds 1000 to 2000. When the penalty term coefficient is 10 and the number of adversarial cycle iterations is about 1000, the MCC value is 0.8645, which is higher than 0.5829, 0.6831 and 0.1965 for penalty coefficients 8, 12 and 0. When the penalty term coefficient is

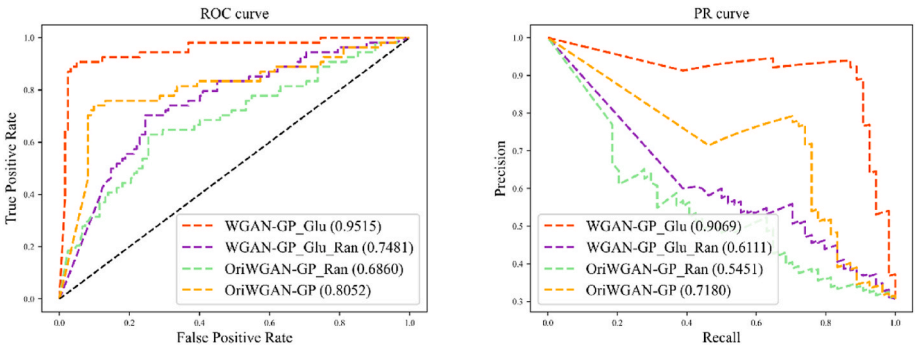
Table 1
ReliableWGAN-GP algorithm analysis.

Models	Sn	Sp	ACC	Precision	MCC
OriWGAN-GP_Ran	63.16 %	74.78 %	70.83 %	52.31 %	0.3633
OriWGAN-GP	80.46 %	89.03 %	86.51 %	71.92 %	0.6835
GluWGAN-GP_Ran	70.78 %	73.75 %	71.75 %	53.52 %	0.4226
WGAN-GP_Glu	90.58 %	95.82 %	94.44 %	88.88 %	0.8645



(a) ROC curves for different negative sample selection strategies. (b) PR curves for different negative sample selection strategies.

Fig. 9. ROC curves and PR curves for different negative sample selection strategies.



(a) ROC curves for different methods (b) PR curves for different methods

Fig. 10. ROC curves and PR curves for the ReliableWGAN-GP algorithm and other negative sampling methods.

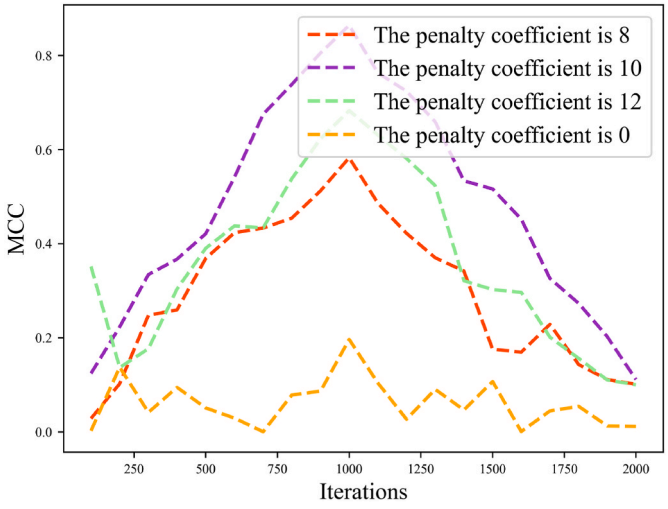


Fig. 11. The trend of MCC values of different number of adversarial cycle iterations with different penalty term coefficient.

0, the value of the MCC for each number of rounds is much lower than the other penalty term coefficients, so adding the penalty term is critical to the model prediction.

3.6. Analysis of model structure

After selecting reliable negative samples by the ReliableWGAN-GP method, the deep feature extraction module is utilized to extract deep features. To analyze the influence of different parts in deep feature extraction module, we construct various model structures and test them on the independent test set, namely Bi-LSTM Attention, CNN Bi-LSTM, CNN-Attention. The results of the different model structures are shown in Table 2. Compared with WGAN-GP_Glu, the structure of Bi-LSTM Attention model decreased from 90.58 %, 95.82 %, 94.44 % and 0.8645–68.82 %, 74.24 %, 71.14 % and 0.3892 without CNN layer. Therefore, the convolutional neural network layer is crucial for extracting the hidden information of feature sequences. At the same time, the Sn, Sp, ACC and MCC of the CNN_Bi-LSTM model without attention mechanism layer are 70.73 %, 81.73 %, 78.3 % and 0.5131, respectively, which indicates that the attention mechanism in WGAN-GP_Glu focus on key information in protein. In addition, compared with WGAN-GP_Glu, the Sn, Sp, ACC and MCC of CNN-Attention model decreased from 90.58 %, 95.82 %, 94.44 % and 0.8645–68.72 %, 83.26 %, 78.33 % and 0.5016. Bi-LSTM in the WGAN-GP_Glu model can consider the context information of the protein, improving the performance of the WGAN-GP_Glu. The results from Fig. 12 show that the different modules of the deep feature extraction module play a crucial role in identifying glutarylation sites.

Table 2
Performance results of different model structures.

Models	Sn	Sp	ACC	MCC
Bi-LSTM_Attention	68.82 %	74.24 %	71.14 %	0.3892
CNN_Bi-LSTM	70.73 %	81.73 %	78.30 %	0.5131
CNN-Attention	68.72 %	83.26 %	78.33 %	0.5016
WGAN-GP_Glu	90.58 %	95.82 %	94.44 %	0.8645

4. Conclusion

In this article, we propose a novel semi-supervised learning algorithm, called WGAN-GP_Glu, for identifying reliable non-glutarylation sites from those without glutarylation annotation. WGAN-GP_Glu uses CKSAAP, AAC, AAF, BLOSUM62 multi-view feature coding and IFS incremental feature selection. The WGAN-GP_Glu prediction method mainly includes a reliable negative sample selection module, a deep feature extraction module, and a glutarylation site prediction module. In the Reliable Negative Sample Selection Module, the ReliableWGAN-GP method is proposed for reliable negative samples selection, and the ReliableWGAN-GP method improves WGAN-GP, that is, a generator and unlabeled samples are re-added to the original WGAN-GP as noise data input to the generator, improving the discrimination ability of the discriminator. The deep feature extraction module contains a CNN, a BLSTM and attention layer. The glutarylation site prediction module is used for the final prediction. WGAN-GP_Glu performed best with other advanced method results on the same independent test set, indicating that in WGAN-GP_Glu negative samples selected from unlabeled samples were reliable and feasible for glutarylation site prediction. In the future, we consider introducing domain adaptation technology into PU learning, so that the model can better adapt to the data distribution in different domains. Since the reliability of the selected negative samples still cannot be verified by biological methods, in future work, we will introduce an unsupervised algorithm to predict glutarylation sites to avoid the trouble caused by data labeling.

CRediT authorship contribution statement

Qiao Ning: Writing – review & editing, Supervision, Conceptualization, Project administration. **Zedong Qi:** Writing – original draft, Validation, Methodology.

Code availability

All the codes can be downloaded from https://github.com/xbbxhbc/WGAN-GP_Glu.git.

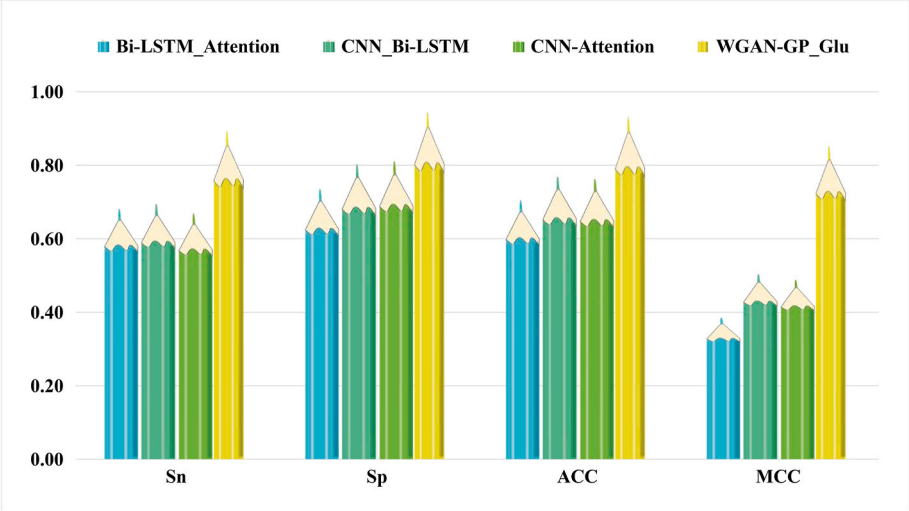


Fig. 12. Performance comparison of different model structures.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by National Natural Science Foundation of China (62302075), Innovation Support Program for Dalian High Level Talents (2023RQ007), the Dalian Excellent Young Project (2022RY35), the Fundamental Research Funds for the Central Universities (3132024251).

References

[1] F. Crick, Central dogma of molecular biology, Nature 227 (5258) (1970) 3–561.
[2] G. Khoury, R. Baliban, C. Floudas, Proteome-wide post translational modifification statistics: frequency analysis and curation of the swiss-prot database, Sci. Rep. 1 (90) (2011).
[3] Z. Liu, Y. Wang, T. Gao, et al., CPLM: a database of protein lysine modifications, Nucleic Acids Res. 42 (2014) D531–D536.
[4] Z. Xie, J. Dai, L. Dai, et al., Lysine succinylation and lysine malonylation in histones, Mol. Cell. Proteomics 11 (2012) 100–107.
[5] E. Kamynina, P. Stover, The roles of SUMO in metabolic regulation, Adv. Exp. Med. Biol. 963 (2017) 143–168.
[6] J. Zhe, J. He, Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC, J. Mol. Graph. Model. 76 (2017) 356–363.
[7] D. Comb, N. Sarkar, C. Pinzino, The Methylation of lysine residues in protein, Biol. Chem. 241 (1966) 1857–1862.
[8] K. Menzies, H. Zhang, E. Katsyuba, et al., Protein acetylation in metabolism—metabolites and cofactors, Nat. Rev. Endocrinol. 12 (2016) 43–60.
[9] M. Tan, C. Peng, K. Anderson, et al., Lysine glutarylation is a protein posttranslational modifification regulated by SIRT5, Cell Metabol. 19 (4) (2014) 605–617.
[10] L. Xie, G. Wang, Z. Yu, et al., Proteome-wide lysine glutarylation profiling of the Mycobacterium tuberculosis H37Rv, J. Proteome Res. 15 (2016) 13791385.
[11] Z. Ju, J. He, Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection, Anal. Biochem. 550 (2018) 1–7.
[12] L. Dou, X. Li, L. Zhang, et al., iGlu AdaBoost: identification of lysine glutarylation using the AdaBoost classifier, J. Proteome Res. 20 (1) (2020).
[13] H. Liu, R. Setiono, Incremental feature selection, Appl. Intell. 9 (3) (1998) 217–230.
[14] H. Albarakati, H. Saigo, R. Newman, et al., RF-GlutarySite: a Random Forest Based Predictor for Glutarylation Sites, Molecular Omics, 2019.
[15] M. Arafat, M. Ahmad, S. Shovan, et al., Accurately predicting glutarylation sites using sequential Bi-Peptide-Based evolutionary features, Genes 11 (9) (2020) 1023.
[16] K.J. Menzies, H. Zhang, E. Katsyuba, et al., Protein acetylation in metabolism—metabolites and cofactors, Nat. Rev. Endocrinol. 12 (1) (2016) 43–60.
[17] Q. Ning, Z. Ma, X. Zhao, et al., SSKM succ: a novel succinylation sites prediction method incorporating K-means clustering with a new semi-supervised learning algorithm, IEEE ACM Trans. Comput. Biol. Bioinf 19 (1) (2020) 643–652.
[18] I. Gulrajani, F. Ahmed, M. Arjovsky, et al., Improved training of wasserstein gans, Adv. Neural Inf. Process. Syst. 30 (2017) 5769–5779.
[19] H. Xu, J. Zhou, S. Lin, et al., PLMD: an updated data resource of protein lysine modifications, Journal of Genetics and Genomics 44 (5) (2017) 243–250.
[20] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.
[21] Y. Huang, B. Niu, Y. Gao, et al., CD-HIT Suite: a web server for clustering and comparing biological sequences, Bioinformatics 26 (2010) 680–682.
[22] G.E. Crooks, WebLogo: a sequence logo generator, Genome Res. 14 (6) (2004) 1188–1190.
[23] M. Bhasin, G. Raghava, Classification of nuclear receptors based on amino acid composition and dipeptide composition, J. Biol. Chem. 279 (2017) 23262–23266.
[24] Y. Chen, Y. Tang, Z. Sheng, et al., Prediction of mucintype Oglycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs, BMC Bioinf. 9 (2008).
[25] Z. Li, J. Fang, S. Wang, et al., Adapt-Kcr: a novel deep learning framework for accurate prediction of lysine crotonylation sites based on learning embedding feature, attention architecture, Briefings Bioinf. (2) (2022) 2.
[26] X. Jing, Q. Dong, D. Hong, et al., Amino acid encoding methods for protein sequences: a comprehensive review and assessment, IEEE ACM Trans. Comput. Biol. Bioinf 17 (6) (2019) 1918–1931.
[27] A. Creswell, T. White, V. Dumoulin, et al., Generative adversarial networks: an overview, IEEE Signal Process. Mag. 35 (1) (2018) 53–65.
[28] G. Huang, Z. Liu, L. Van Der Maaten, et al., Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[29] Y. Li, Y. Pan, A novel ensemble deep learning model for stock prediction based on stock prices and news, International Journal of Data Science and Analytics 13 (2022) 139–149.
[30] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 5998–6008.
[31] M.H. Guo, Z.N. Liu, T.J. Mu, et al., Beyond self-attention: external attention using two linear layers for visual tasks, IEEE Trans. Pattern Anal. Mach. Intell. 45 (5) (2022) 5436–5447.

Qiao Ning, received the B.S. and the PhD degree from School of information science and technology, Northeast Normal University, China, in 2019. She is currently a Lecturer in Department of Information Science and Technology, Dalian Maritime University, Dalian. Her research interests include machine learning and bioinforamtics.

Zedong Qi received his Master degree from Information Science and Technology, Dalian Maritime University. His research interests include protein sites prediction and machine learning in bioinformatics.