ASEAN Artificial Intelligence Journal

Journal homepage:
https://karyailham.com.my/index.php/aaij/index
ISSN: 3083-9971

Prediction of Lysine Malonylation Sites Using Ensemble Learning

Xin Wei[1,2], Muhammad Akmal Remli[1,2,*]

[1] Institute for Artificial Intelligence and Big Data, Universiti Malaysia Kelantan, 16100 Kota Bharu, Kelantan, Malaysia
[2] Faculty of Data Science and Computing, Universiti Malaysia Kelantan, 16100 Kota Bharu, Kelantan, Malaysia

ARTICLE INFO

Article history:
Received 21 January 2026
Received in revised form 19 February 2026
Accepted 8 April 2026
Available online 12 May 2026

Keywords:
Lysine Malonylation; ensemble learning; feature extraction; machine learning

ABSTRACT

Protein post-translational modifications (PTMs) serve as crucial regulators of protein function, referring to chemical modifications of proteins coordinated by PTM enzymes, which play key roles in numerous physiological processes. To date, over 400 distinct types of PTMs have been identified. Malonylation, a newly discovered PTM, involves the chemical modification of positively charged lysine side chains and participates in the regulation of human metabolism, demonstrating significant associations with functional and structural alterations. In this project, we extracted sequence feature information by integrating coupled information from protein sequences with general pseudo-amino acid composition (PseAAC). Multiple ensemble learning methods were employed to train and classify imbalanced datasets. Results from cross-validation models and independent test sets indicate that our approach outperforms existing predictors in terms of Sn (sensitivity) and MCC (Matthews correlation coefficient). Given the importance of Sn and MCC for imbalanced data, the overall improvement achieved remains substantial.

1. Introduction

In the field of bioinformatics, lysine malonylation is a significant, newly discovered protein post-translational modification (PTM) in recent years. It involves the chemical modification of the positively charged side chain of lysine (K), playing a role in regulating human metabolism and demonstrating considerable relevance to functional and structural changes. Among the twenty typical amino acids, lysine is subject to particularly rich and diverse modifications [1-5]. Given the crucial role of protein K-malonylation in both pathological and physiological processes, the detection of protein K-malonylation holds substantial value in medicine and the life sciences. Considerable efforts have been made by numerous experts, scholars, and researchers to enhance the efficiency of experimentally identifying K-malonylation sites. Various proteomic methods have recently been developed for recognizing malonylation sites, including mass spectrometry-based approaches, affinity enrichment, chemical probes, and label-free quantitative methods [6-9].

* *Corresponding author.*
E-mail address: akmal@umk.edu.my

Given the high complexity of the aforementioned experimental methods, we can effectively adopt computational approaches analogous to those established for other types of post-translational modifications (PTMs) to predict lysine malonylation sites.In recent years, a growing number of scholars have dedicated their efforts to the research on lysine malonylation site prediction. In 2016, Xu *et al*., [10] incorporated residue sequence order information, position-specific amino acid propensities, and physicochemical properties, and adopted the minimum redundancy maximum relevance (mRMR) feature selection method to identify optimal features from the entire feature set, with validation performed via the leave-one-out cross-validation (LOOCV) strategy. In 2017, Wang *et al*., [11] developed MaloPred, a novel online predictor for malonylation sites, by integrating diverse informative features across species preferences and implementing an enhanced feature engineering strategy.In 2016, Du *et al*., [12] designed a predictive algorithm applicable to multiple types of lysine modifications, including malonylation.In 2018, Chen *et al*., [13] proposed LEMP, a high-performance predictor that significantly enhanced the accuracy of lysine malonylation site prediction. In 2022, Li *et al*., [14] developed DeepMal, a high-accuracy predictor that utilized deep neural networks integrating both sequence information and evolutionary features to identify protein lysine malonylation sites. In 2024, Chen *et al*., [15] introduced MALSpectraNet, an advanced multimodal deep learning framework designed for cross-species prediction of lysine malonylation sites by comprehensively leveraging sequence attributes and structural profiles. This was achieved by combining long short-term memory (LSTM) networks and random forest classifiers with a novel encoding scheme to optimize amino acid feature representation.

In this study, we developed an effective predictive method for identifying protein malonylation sites. To achieve this, we utilized a learning method by Jia, employing the coupled sequence feature method based on pseudo-amino acid composition (PseAAC) to extract feature information. For the dataset comprising positive and negative samples, we implemented an ensemble classifier and adopted a majority voting principle to ensure more rigorous prediction outcomes. Furthermore, five-fold cross-validation was applied to train the data, taking into account not only training error but also generalization error. Finally, the average results from multiple validation rounds were used to objectively evaluate the accuracy of the predictions.

2. Materials and Methods

2.1 Benchmark Dataset

The original dataset used in the development of this method was retrieved from previous studies, which included experimentally identified malonylation sites from humans and mice [16,17]. Following the same procedure as Chen [13], only non-redundant data were retained. Sequence homology was reduced using CD-Hit to remove sequences with over 30% identity [18]. Briefly, experimentally verified malonylation sites were defined as positive sites, while unknown lysine residues were defined as negative sites. A sliding window comprising a specific number of amino acid residues upstream and downstream of each positive or negative site was applied. Various window sizes were tested, and the optimal size was selected based on performance. Through random sampling, one-fifth of the data samples from both the positive and negative sets were selected to form an independent test set, while the remaining four-fifths were used as the training set. The summary statistics for the independent test set and the training set are presented in Table 1.

Table 1
The details of the benchmark dataset

Original Dataset	Positive	Negative
Training Dataset	4242	71809
Testing Dataset	1046	16827

According to Chou's method, peptide sequences containing K-malonylation modification sites can be represented as follows:

$$\mathrm{P}_{\delta}(\mathbb{K}) = \mathrm{H}_{-\delta}\mathrm{H}_{-(\delta-1)} \cdots \mathrm{H}_{-2}\mathrm{H}_{-1}\mathbb{K}\mathrm{H}_{+1}\mathrm{H}_{+2} \cdots \mathrm{H}_{+(\delta-1)}\mathrm{H}_{+\delta} \quad (1)$$

In this representation, the central position $\mathbb{K}$ denotes the modification site, while the $-\delta$ positions to the left of the central $\mathbb{K}$ represent the upstream amino acid residues, and the $+\delta$ positions to the right represent the downstream amino acid residues. Consequently, the samples $\mathrm{P}_{\delta}(\mathbb{K})$ can be categorized into two classes:

$$\mathrm{P}_{\delta}(\mathbb{K}) \in \begin{cases} \mathrm{P}_{\delta}^{+}(\mathbb{K}), & \text{if } \mathbb{K} \text{ is the malonylation} \\ \mathrm{P}_{\delta}^{-}(\mathbb{K}), & \text{otherwise} \end{cases} \quad (2)$$

Here, $\mathrm{P}_{\delta}^{+}(\mathbb{K})$ denotes peptide fragments with lysine as the central modified residue (true malonylation), while $\mathrm{P}_{\delta}^{-}(\mathbb{K})$ represents peptide fragments with unmodified lysine as the central residue (false malonylation). In the literature, benchmark datasets are typically divided into two types: one for training the model (training dataset) and the other for testing the model (testing dataset). A common approach is to employ subsampling (K-fold) cross-validation for predictive modeling. This method randomly splits the benchmark dataset into multiple subsets, ensuring both diversity and independence of the results. Accordingly, the data in this study are divided into two parts:

$$\mathbb{S}_{\delta} = \mathbb{S}_{\delta}^{+} \cup \mathbb{S}_{\delta}^{-} \quad (3)$$

There are two types of malonylation peptide segments (referring to Eq. 3): true malonylation peptide samples are denoted as $\mathbb{S}_{\delta}^{+}$, while false malonylation peptide segments are denoted as $\cup\, \mathbb{S}_{\delta}^{-}$, where the symbol $\cup$ represents the union of sets.

According to Eq. (1), the peptide can be represented as:

$$L(\delta) = 2\delta + 1 \quad (4)$$

Furthermore, for each δ value, the reference dataset comprises distinct peptide fragments from $\mathbb{S}_{\delta}$, with varying sample sizes due to the differing counts of positive and negative samples. For example:

$$N(\delta) = N^{+}(\delta) + N^{-}(\delta) \quad (5)$$

Here, $N^{+}(\delta)$represents the number of samples in the positive benchmark dataset S_{δ}^{+}, while$N^{-}(\delta)$represents the number in the negative benchmark dataset S_{δ}^{-}. Residues represented by 'X' were filled according to the method described by Jia et al. [19]. As shown in Eq. (1), for a sample $P_{\delta}(K)$, a residue 'X' may appear at the left or right end (but only on one side) of the central site K. This occurs because the protein sequence is truncated during the cleavage process, resulting in insufficient length at the protein's N- or C-terminus. Taking the case where the 'X' residue is on the left end as illustrated in Figure 1(A), the intact sequence from the right end can be mirrored and used to fill the left end, ensuring consistent sample length for all sequences. If the 'X' residue appears on the right end, as in Figure 1(B), the same principle applies as in Figure 1(A).

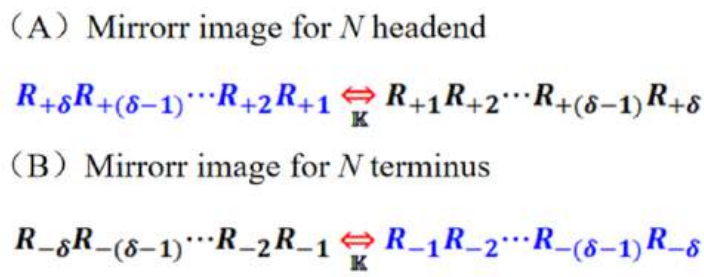

Fig. 1. The residue is "X" processing method The part whose color is marked in blue,the real peptide is black, and the red symbol ⇔ in the middle indicates a mirror.K indicates modification site. (A) Represents the mirror image of the δ residue X at the head of N. (B) Represents the mirror image of the δ residue X at the end of N

Based on preliminary experiments, the optimal results were achieved when δ = 15. As derived from Eq. (1), this yields L(δ)=31. From Table 1, we obtain $N^{+}(\delta) = 5288$ and $N^{-}(\delta) = 88636$. Detailed sample sequence information is available for download at the following link: https://github.com/weixin7112/Chen_data.

2.2 Feature Coding Schemes

2.2.1 One-hot encoding

One-hot encoding is a method well-suited for protein feature extraction. It involves assigning binary labels to the 20 standard amino acids plus an unknown amino acid. For this study, with δ = 15, each amino acid (A, C, …, Y, X) can be encoded as (10000000000000000000), (01000000000000000000), …, (00000000000000000001), respectively, while the unknown amino acid X is represented as (00000000000000000000). Consequently, each sample from Eq. (1 is transformed into a 31 × 21-dimensional feature matrix, which is then used for subsequent model development.

2.2.2 Amino Acid Composition (AAC)

calculating the occurrence frequency of the 20 standard amino acids and the unknown amino acid X within each sequence sample. Assuming a protein sequence is denoted as K with a length L of 31, and letting $f_{(K_i)}$ represent the count of amino acid i in sequence K, the feature value for each amino acid can be expressed as:

$$P_{(K_i)} = \frac{f_{(K_i)}}{31} \tag{6}$$

Where $P_{(K_i)}$ represents the frequency of occurrence of each amino acid, with i denoting the amino acid (A, C, …, Y, X). Finally, the protein sequence K can be represented as a feature vector as follows:

$$P_{(K)} = \left[P_{(K_1)}\ P_{(K_2)}\ \cdots\ P_{(K_{20})}\right] \tag{7}$$

2.1.3 Use general PseAAC to formulate Peptide Samples (PseAAC)

In bioinformatics, transforming biological sequences into mathematical models has long been a challenging issue. Within existing machine learning algorithms, it is difficult to directly process sequence information through mathematical models. However, bioinformatics plays a significant role

in the development of human life and society. With the continuous advancement of human knowledge and science, we can represent PseAAC from Eq. (1) as follows:

$$\mathrm{P}_{\delta}(\mathbb{K}) = \mathrm{P}_{\delta}^{+}(\mathbb{K}) - \mathrm{P}_{\delta}^{-}(\mathbb{K}) \tag{8}$$

In the processing of samples from the Chen dataset, the presence of unknown amino acid "X" in sequences is addressed through the mirror-filling procedure illustrated in Figure 1. After this step, $\mathrm{P}_{\delta}(\mathbb{K})$ can be specifically represented as:

$$\mathrm{P}_{\delta}^{+}(\mathbb{K}) = \begin{cases} \mathrm{P}_{-\delta}^{+}(\mathbb{K}) = \mathrm{P}_{-\delta}^{+}\left(H_{-\delta}|H_{-(\delta-1)}\right), & \text{if } (-\delta) = -1, \mathrm{P}_{-\delta}^{+}(\mathbb{K}) = \mathrm{P}_{-\delta}^{+}(H_{-\delta}) \\ \mathrm{P}_{+\delta}^{+}(\mathbb{K}) = \mathrm{P}_{+\delta}^{+}\left(H_{+\delta}|H_{+(\delta-1)}\right), & \text{if } (+\delta) = +1, \mathrm{P}_{+\delta}^{+}(\mathbb{K}) = \mathrm{P}_{+\delta}^{+}(H_{+\delta}) \end{cases} \tag{9}$$

$$\mathrm{P}_{\delta}^{-}(\mathbb{C}) = \begin{cases} \mathrm{P}_{-\delta}^{-}(\mathbb{K}) = \mathrm{P}_{-\delta}^{-}\left(H_{-\delta}|H_{-(\delta-1)}\right), & \text{if } (-\delta) = -1, \mathrm{P}_{-\delta}^{-}(\mathbb{K}) = \mathrm{P}_{-\delta}^{-}(H_{-\delta}) \\ \mathrm{P}_{+\delta}^{-}(\mathbb{K}) = \mathrm{P}_{+\delta}^{-}\left(H_{+\delta}|H_{+(\delta-1)}\right), & \text{if } (+\delta) = +1, \mathrm{P}_{+\delta}^{-}(\mathbb{K}) = \mathrm{P}_{+\delta}^{-}(H_{+\delta}) \end{cases} \tag{10}$$

It is evident that both Eq. (9) and Eq. (10) are derived based on positional information correlations, and their methods for obtaining conditional probabilities are identical. From Eq. (9), the calculation of $\mathrm{P}_{\delta}^{+}(\mathbb{K})$ yields:

- $\mathrm{P}_{-\delta}^{+}(\mathbb{K})$ and $\mathrm{P}_{+\delta}^{+}(\mathbb{K})$ represent the conditional probabilities of amino acids, respectively.
- $\mathrm{P}_{-\delta}^{+}\left(H_{-\delta}|H_{-(\delta-1)}\right)$ denotes the conditional probability of amino acid $H_{-\delta}$ given its right neighbor $H_{-(\delta-1)}$.
- $\mathrm{P}_{+\delta}^{+}\left(H_{+\delta}|H_{+(\delta-1)}\right)$ denotes the conditional probability of amino acid $H_{+\delta}$ given its left neighbor $H_{+(\delta-1)}$.

From Eq. (1), it can be observed that the immediate left neighbor H_{-1} and right neighbor H_{+1} of the central site K are treated as unconditional probabilities. The rationale behind Eq. (10) follows the same principle as Eq. (9).

2.3 Classifier Algorithms

As shown in Table 1 of this study, it is evident that the number of negative samples significantly exceeds that of positive samples. Consequently, lysine non-malonylation sites represent the majority class. If conventional classifiers are directly applied to such a highly imbalanced dataset, they tend to exhibit a strong bias toward predicting negative samples during the classification process. This leads to a high prediction accuracy for the majority class (negative samples) at the expense of missing or poorly predicting the minority class (positive samples, i.e., malonylation sites). Although the overall prediction accuracy may appear favorable, such an approach is fundamentally flawed. Failing to accurately predict positive samples would defeat the purpose of this research. Therefore, under such circumstances, it is essential to first address the dataset imbalance to create a more balanced sample distribution.

Subsequently, we will enhance conventional classifiers through ensemble techniques to improve the prediction accuracy on the original dataset.

2.3.1 Integrated random forest learning algorithm

Random Forest algorithm is one of the most widely used classification algorithms in current experimental research [20-22], known for its speed and strong classification performance. In this study, to address the imbalanced sample data, we constructed an ensemble classifier based on the Random Forest algorithm.

Here, we established a voting system comprising n integrated predictors:

$$RF^{\varphi} = RF(1) \vee RF(1) \vee \cdots \vee RF(n) = \vee_{i=1}^{n} RF(n) \tag{11}$$

Here, RF^{φ} denotes the ensemble classifier, and V represents the fusion decision operator. Given that the positive-to-negative sample ratio is approximately 1:17, we set $n = 17$. First, the negative samples are randomly divided into 17 subsets, each of which is then combined with the positive samples to form a balanced dataset. After preprocessing the features of these 17 original datasets, bootstrap sampling is performed on each. Subsequently, a decision tree is trained on each sampled dataset. The predictions from all base classifiers (decision trees) are aggregated via the Random Forest method to produce intermediate results through voting. Finally, the outputs of the 17 Random Forests are integrated via another round of voting to obtain the final prediction. The specific computational workflow is illustrated in Figure 2:

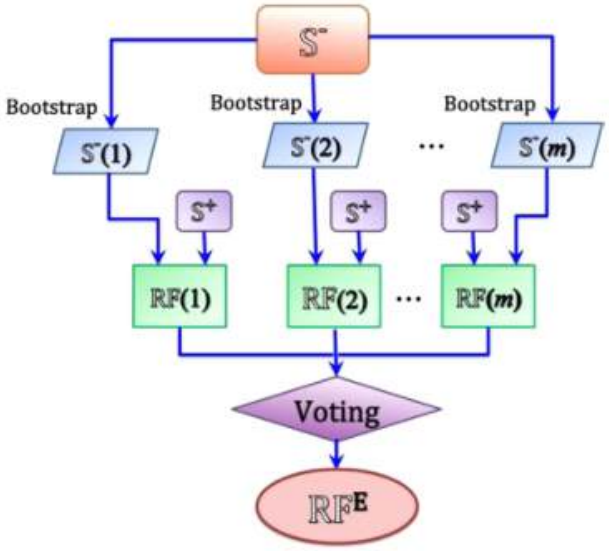

Fig. 2. Flowchart of the ensemble random forest algorithm

2.3.2 Integrated support vector machine learning algorithm

In this study, the original benchmark dataset exhibits $N^{-}(\delta) >> N^{+}(\delta)$, indicating a significant class imbalance where negative samples far outnumber positive samples. Compared to the number of malonylation sites, non-malonylation sites consistently constitute the majority. When dealing with such imbalanced data, conventional machine learning methods often exhibit a "lazy" bias during binary classification. Taking the dataset in this study as an example, a model could achieve a high overall accuracy by simply favoring predictions of the negative class, rendering the classification task meaningless and inevitably introducing substantial predictive error. However, the primary objective of our prediction task is to accurately identify malonylation sites. To achieve this, the integrated Support Vector Machine (SVM) algorithm presents a suitable choice.

The Support Vector Machine (SVM) has been widely applied across numerous domains in bioinformatics [23]. Its principle involves constructing a maximum-margin separating hyperplane to distinguish between positive and negative samples. The algorithm primarily relies on two parameters: c, which controls the tolerance for training errors, and g, which influences the speed of prediction and training. In this work, a grid search algorithm was employed for dataset training, with parameter c ranging from an initial value of -5 to a final value of 15 with a step size of 2, and parameter g ranging from an initial value of 3 to a final value of -15 with a step size of -2. This algorithm is also capable of operating effectively on imbalanced datasets.

This study utilized the libsvm-3.24 software package to conduct the experiments, which is available for download at: https://www.csie.ntu.edu.tw/~cjlin/libsvm/. For such imbalanced datasets, ensemble learning methods can be employed by applying under-sampling techniques to

construct multiple sub-classifiers. These models are then trained via cross-validation, and their performance is evaluated on an independent test set. The use of an ensemble classifier has been shown to significantly enhance the prediction accuracy for protein post-translational modification sites [24]. The formulation of the ensemble classifier is as follows:

$$\mathbb{T}^{\mathcal{E}} = \mathbb{T}(1) \vee \mathbb{T}(2) \vee \cdots \vee \mathbb{T}(i) = \vee_{i=1}^{m} \mathbb{T}(i) \tag{13}$$

Here, $\mathbb{T}^{\mathcal{E}}$ denotes the ensemble classifier, and $\vee$ represents the fusion operator. Each $\mathbb{T}(i)$ functions as a voting component. Given that the positive-to-negative sample ratio in the dataset is 1:17, we set m=17, thereby constructing 17 sub-classifiers. For each working benchmark dataset mentioned above, this study employs an individual SVM classifier. Subsequently, the independent test set is evaluated by all 17 predictors, and the final prediction is determined through majority voting.

2.3.3 Performance evaluation

The performance of the method was evaluated through cross-validation and independent test set validation, yielding four key metrics. Traditional evaluation of predictive performance includes Specificity (Sp), Sensitivity (Sn), Accuracy (Acc), and the Matthews Correlation Coefficient (MCC) [25], defined as follows:

$$\begin{cases} S_n = 1 - \dfrac{N_-^+}{N^+}, \quad 0 \le S_n \le 1 \\ S_p = 1 - \dfrac{N_+^-}{N^-}, \quad 0 \le S_p \le 1 \\ Acc = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-}, \quad 0 \le Acc \le 1 \\ MCC = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}}, \quad -1 \le Mcc \le 1 \end{cases} \tag{14}$$

Here, N^+ represents the total number of true $\mathbb{K}$ (modified sites), N^- denotes the total number of false $\mathbb{K}$ (unmodified sites), N_-^+ indicates the number of misclassified true $\mathbb{K}$ (modified sites), and N_+^- represents the number of misclassified false $\mathbb{K}$ (non-modified sites).

The Receiver Operating Characteristic curve (ROC curve) illustrates the relationship between sensitivity and specificity [4]. The x-coordinate (1-Specificity) approaches 0 as the accuracy increases; the y-coordinate, referred to as sensitivity or the True Positive Rate, indicates better precision when its value is higher.

Based on the position of the curve, the graph is divided into two parts. The area under the curve, known as Area Under the Curve (AUC), serves as an indicator of predictive accuracy. A higher AUC value, corresponding to a larger area under the curve, reflects greater prediction accuracy. The closer the curve is to the upper-left corner (where x is smaller and y is larger), the higher the predictive precision.

3. Results and Discussion

3.1 Comparison of the Four Metrics

Given that the sample size in this study exceeds 90,000 entries, making the dataset substantial, and considering the complexity and variety of the feature extraction methods proposed, extensive

use of classifiers for prediction is required. The first step is to determine whether the methodology is suitable for this dataset. Given that the integrated Random Forest algorithm is efficient and offers relatively high predictive accuracy, we initially applied multiple feature extraction methods to the entire dataset and performed predictions using the integrated Random Forest classifier. If the results show significant variations, switching classifiers is unlikely to yield substantial improvements. While the Support Vector Machine algorithm excels in predictive accuracy, its grid optimization process is time-consuming. If all feature extraction methods were directly tested with this classifier from the outset—especially under ensemble learning, which expands the sample dataset—the research process could become excessively time-consuming, posing practical challenges.

Therefore, under the premise that the Random Forest classifier demonstrates satisfactory performance, it is a reasonable strategy to subsequently employ the Support Vector Machine classifier for comparative prediction, enabling the selection of the superior classifier. The outcomes and comparisons from this study are summarized in Tables 2 and Tables 3.

Table 2
Comparison of cross-validation results from different methods on the Chen Dataset

Predictor	Sn(%)	Sp(%)	Acc(%)	MCC(%)
LEMP[13]	30.53	90.76	87.62	16.00
RF_{AAC}	67.04	59.25	59.69	12.22
$RF_{One-hot}$	57.13	60.61	60.41	8.29
RF_{PseAAC}	76.28	75.18	75.24	26.40
SVM_{PseAAC}	78.67	76.67	76.78	28.78

Table 3
Comparison of independent test set results from different methods on Chen Dataset

Predictor	Sn(%)	Sp(%)	Acc(%)	MCC(%)
LEMP[13]	43.79	90.00	87.30	24.40
RF_{AAC}	65.49	59.37	59.73	11.82
$RF_{One-hot}$	58.03	61.15	60.97	9.19
RF_{PseAAC}	77.36	74.28	75.18	27.75
SVM_{PseAAC}	76.58	77.79	77.72	29.32

From the two tables presented above, we can observe that by employing multiple feature extraction methods combined with classifiers for prediction, the PseAAC feature extraction method demonstrates relatively strong performance, with a notably significant improvement in effectiveness. Subsequently, we performed classification under both the integrated Random Forest and integrated Support Vector Machine classifiers, obtaining the currently optimal method in this research: SVM_{PseAAC}.

When compared with the LEMP method, although the specificity (Sp) for negative sample prediction decreased by 14.09% in cross-validation and 12.21% in the independent test set, and the overall accuracy (Acc) also slightly declined, the sensitivity (Sn) for positive sample prediction increased significantly by 48.14% and 32.79%, respectively. The Matthews correlation coefficient (MCC) also improved by 12.78% and 4.92% correspondingly. Furthermore, the dataset used in this study exhibits a positive-to-negative sample ratio of 1:17, indicating severe data imbalance. In addressing such issues, Sn and MCC serve as the primary evaluation metrics, making improvements in these two indicators particularly crucial. This approach provides substantial assistance for the future prediction of protein K-malonylation sites. Overall, the SVM_{PseAAC} method demonstrates stable and significant improvement over existing methods.

3.2 Robustness and Feasibility Analysis

To obtain a stable and reliable model, this study employs a robust model performance evaluation technique. Cross-validation was used to assess both the training set and the independent test set. The area under the ROC curve, which represents the region enclosed by the curve and the axes, serves as a method to evaluate the predictive effectiveness of the classifier. As shown in Tables 2 and 3, among the classification methods compared for this dataset, the SVM_{PseAAC} method demonstrates the best performance. Figures 4 and 5 present the ROC curve results for the SVM_{PseAAC} method on the independent test set and cross-validation, respectively:

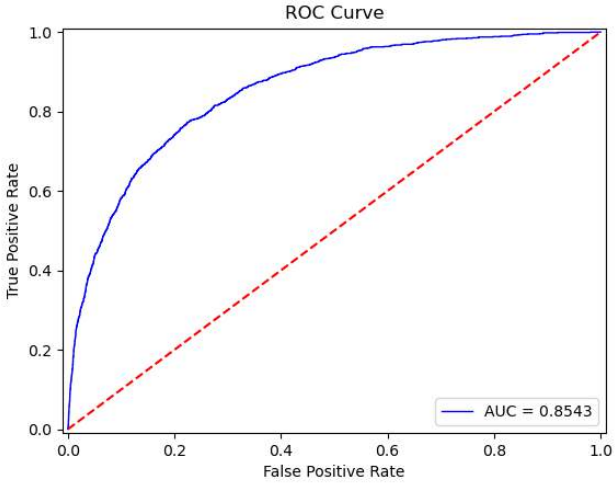

Fig. 2. ROC curve of the independent test set under the SVM_{PseAAC} method

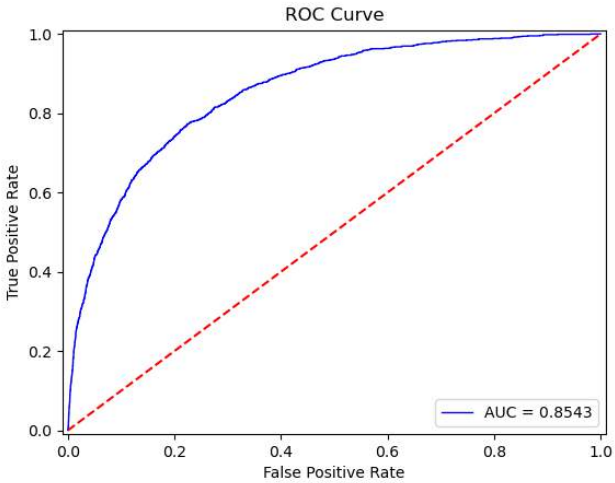

Fig. 3. ROC curve of the cross-validation under the SVM_{PseAAC} method

4. Conclusions

This study demonstrates the effectiveness of a hybrid computational framework integrating diverse feature representations with ensemble learning strategies for tackling the challenging task of predicting malonylation sites. The success of the SVM_PseAAC method highlights the importance of capturing both compositional and sequential information, suggesting that the positional and contextual patterns around lysine residues are crucial for accurate identification.

Future research could explore several promising directions. Firstly, incorporating more advanced feature descriptors, such as evolutionary information from Position-Specific Scoring Matrices

(PSSMs) or physicochemical properties of amino acids, might further enhance predictive power. Secondly, deep learning models, particularly Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), which can automatically learn hierarchical and long-range dependency features from raw sequences, present a compelling alternative to manual feature engineering. Thirdly, extending this methodology to predict other types of lysine modifications (e.g., succinylation, crotonylation) or developing multi-label prediction systems for concurrent modifications would test its generalizability and increase its biological utility. Finally, interpreting the most predictive features or patterns learned by the model could provide valuable biological insights into the mechanisms and preferences of malonylation, potentially guiding experimental validation.

References

[1] Liu, Zexian, Yongbo Wang, Tianshun Gao, Zhicheng Pan, Han Cheng, Qing Yang, Zhongyi Cheng, Anyuan Guo, Jian Ren, and Yu Xue. "CPLM: a database of protein lysine modifications." *Nucleic acids research* 42, no. D1 (2014): D531-D536. https://doi.org/10.1093/nar/gkt1093

[2] Li, Fuyi, Chen Li, Jerico Revote, Yang Zhang, Geoffrey I. Webb, Jian Li, Jiangning Song, and Trevor Lithgow. "GlycoMine struct: a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features." *Scientific reports* 6, no. 1 (2016): 34595. https://doi.org/10.1038/srep34595

[3] Hanley, James A., and Barbara J. McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology* 143, no. 1 (1982): 29-36. https://doi.org/10.1148/radiology.143.1.7063747

[4] Lanouette, Sylvain, Vanessa Mongeon, Daniel Figeys, and Jean-François Couture. "The functional diversity of protein lysine methylation." *Molecular systems biology* 10, no. 4 (2014): MSB134974. https://doi.org/10.1002/msb.134974

[5] Wei, Leyi, Jie Hu, Fuyi Li, Jiangning Song, Ran Su, and Quan Zou. "Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms." *Briefings in Bioinformatics* 21, no. 1 (2020): 106-119. https://doi.org/10.1093/bib/bby107

[6] Bao, Xiucong, Qian Zhao, Tangpo Yang, Yi Man Eva Fung, and Xiang David Li. "A chemical probe for lysine malonylation." *Angewandte chemie international edition* 52, no. 18 (2013): 4883-4886. https://doi.org/10.1002/anie.201300252

[7] Peng, Chao, Zhike Lu, Zhongyu Xie, Zhongyi Cheng, Yue Chen, Minjia Tan, Hao Luo et al. "The first identification of lysine malonylation substrates and its regulatory enzyme." *Molecular & cellular proteomics* 10, no. 12 (2011). https://doi.org/10.1074/mcp.M111.012658

[8] Hirschey, Matthew D., and Yingming Zhao. "Metabolic regulation by lysine malonylation, succinylation, and glutarylation." *Molecular & Cellular Proteomics* 14, no. 9 (2015): 2308-2315. https://doi.org/10.1074/mcp.R114.046664

[9] Du, Yipeng, Tanxi Cai, Tingting Li, Peng Xue, Bo Zhou, Xiaolong He, Peng Wei, Pingsheng Liu, Fuquan Yang, and Taotao Wei. "Lysine malonylation is elevated in type 2 diabetic mouse models and enriched in metabolic associated proteins." *Molecular & Cellular Proteomics* 14, no. 1 (2015): 227-236. https://doi.org/10.1074/mcp.M114.041947

[10] Xu, Yan, Ya-Xin Ding, Jun Ding, Ling-Yun Wu, and Yu Xue. "Mal-Lys: prediction of lysine malonylation sites in proteins integrated sequence-based features with mRMR feature selection." *Scientific reports* 6, no. 1 (2016): 38318. https://doi.org/10.1038/srep38318

[11] Wang, Li-Na, Shao-Ping Shi, Hao-Dong Xu, Ping-Ping Wen, and Jian-Ding Qiu. "Computational prediction of species-specific malonylation sites via enhanced characteristic strategy." *Bioinformatics* 33, no. 10 (2017): 1457-1463. https://doi.org/10.1093/bioinformatics/btw755

[12] Du, Yipeng, Zichao Zhai, Ying Li, Ming Lu, Tanxi Cai, Bo Zhou, Lei Huang, Taotao Wei, and Tingting Li. "Prediction of protein lysine acylation by integrating primary sequence information with multiple functional features." *Journal of proteome research* 15, no. 12 (2016): 4234-4244. https://doi.org/10.1021/acs.jproteome.6b00525

[13] Chen, Zhen, Ningning He, Yu Huang, Wen Tao Qin, Xuhan Liu, and Lei Li. "Integration of a deep learning classifier with a random forest approach for predicting malonylation sites." *Genomics, proteomics & bioinformatics* 16, no. 6 (2018): 451-459. https://doi.org/10.1016/j.gpb.2018.07.003

[14] Li, Y., Wang, M., Xu, H., et al. 2022. DeepMal: Accurate prediction of protein lysine malonylation sites using deep neural networks with integrated sequence and evolutionary features. Briefings in Bioinformatics. 23(4): bbac242. https://doi.org/10.1093/bib/bbac242

[15] Chen, X., Zhang, L., Zhou, Y., et al. 2024. MALSpectraNet: A multimodal deep learning framework for cross-species lysine malonylation site prediction from sequence and structural profiles. Bioinformatics. 40(2): btae036. https://doi.org/10.1093/bioinformatics/btae036

[16] Colak, Gozde, Olga Pougovkina, Lunzhi Dai, Minjia Tan, Heleen Te Brinke, He Huang, Zhongyi Cheng et al. "Proteomic and biochemical studies of lysine malonylation suggest its malonic aciduria-associated regulatory role in mitochondrial function and fatty acid oxidation." *Molecular & Cellular Proteomics* 14, no. 11 (2015): 3056-3071. https://doi.org/10.1074/mcp.M115.048850

[17] Nishida, Yuya, Matthew J. Rardin, Chris Carrico, Wenjuan He, Alexandria K. Sahu, Philipp Gut, Rami Najjar et al. "SIRT5 regulates both cytosolic and mitochondrial protein malonylation with glycolysis as a major target." *Molecular cell* 59, no. 2 (2015): 321-332. https://doi.org/10.1016/j.molcel.2015.05.022

[18] Xiao, Xuan, Xiang Cheng, Genqiang Chen, Qi Mao, and Kuo-Chen Chou. "pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC." *Genomics* 111, no. 4 (2019): 886-892.. https://doi.org/10.1016/j.ygeno.2018.05.017

[19] Jia, Jianhua, Zi Liu, Xuan Xiao, Bingxiang Liu, and Kuo-Chen Chou. "pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach." *Journal of theoretical biology* 394 (2016): 223-230. https://doi.org/10.1016/j.jtbi.2016.01.020

[20] Jia, Jianhua, Zi Liu, Xuan Xiao, Bingxiang Liu, and Kuo-Chen Chou. "iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC." *Journal of theoretical biology* 377 (2015): 47-56. https://doi.org/10.1016/j.jtbi.2015.04.011IF: 2.0

[21] Kandaswamy, Krishna Kumar, Kuo-Chen Chou, Thomas Martinetz, Steffen Möller, P. N. Suganthan, S. Sridharan, and Ganesan Pugalenthi. "AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties." *Journal of theoretical biology* 270, no. 1 (2011): 56-62. https://doi.org/10.1016/j.jtbi.2010.10.037

[22] Lin, Wei-Zhong, Jian-An Fang, Xuan Xiao, and Kuo-Chen Chou. "iDNA-Prot: identification of DNA binding proteins using random forest with grey model." *PloS one* 6, no. 9 (2011): e24756.https://doi.org/10.1371/journal.pone.0024756

[23] Yang, Yang, Huiwen Zheng, Chunhua Wang, Wanyue Xiao, and Taigang Liu. "Predicting apoptosis protein subcellular locations based on the protein overlapping property matrix and tri-gram encoding." *International Journal of Molecular Sciences* 20, no. 9 (2019): 2344. https://doi.org/10.3390/ijms20092344

[24] Chou, Kuo-Chen, and Hong-Bin Shen. "Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers." *Journal of proteome research* 5, no. 8 (2006): 1888-1897. https://doi.org/10.1021/pr060167c

[25] Chen, Wei, Hao Lv, Fulei Nie, and Hao Lin. "i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome." *Bioinformatics* 35, no. 16 (2019): 2796-2800. https://doi.org/10.1093/bioinformatics/btz015