


METHODOLOGY ARTICLE

Open Access



# Mal-Prec: computational prediction of protein Malonylation sites via machine learning based feature integration

## Malonylation site prediction

Xin Liu<sup>1\*†</sup>, Liang Wang<sup>1,2†</sup> , Jian Li<sup>3</sup>, Junfeng Hu<sup>1</sup> and Xiao Zhang<sup>1\*</sup>

### Abstract

**Background:** Malonylation is a recently discovered post-translational modification that is associated with a variety of diseases such as Type 2 Diabetes Mellitus and different types of cancers. Compared with experimental identification of malonylation sites, computational method is a time-effective process with comparatively low costs.

**Results:** In this study, we proposed a novel computational model called Mal-Prec (Malonylation Prediction) for malonylation site prediction through the combination of Principal Component Analysis and Support Vector Machine. One-hot encoding, physio-chemical properties, and composition of k-spaced acid pairs were initially performed to extract sequence features. PCA was then applied to select optimal feature subsets while SVM was adopted to predict malonylation sites. Five-fold cross-validation results showed that Mal-Prec can achieve better prediction performance compared with other approaches. AUC (area under the receiver operating characteristic curves) analysis achieved 96.47 and 90.72% on 5-fold cross-validation of independent data sets, respectively.

**Conclusion:** Mal-Prec is a computationally reliable method for identifying malonylation sites in protein sequences. It outperforms existing prediction tools and can serve as a useful tool for identifying and discovering novel malonylation sites in human proteins. Mal-Prec is coded in MATLAB and is publicly available at <https://github.com/flyinsky6/Mal-Prec>, together with the data sets used in this study.

**Keywords:** Post-translational modification, Malonylation, Machine learning, Principal component analysis, Support vector machine

### Background

Post-translational modification (PTM) participates in many biological processes through protein function regulations. It has been well recognized that PTM identification is critical in the prevention and medical treatment of certain diseases. Lysine malonylation (Kmal) is a novel

type of PTMs that was initially detected by mass spectrometry and is widely present in both eukaryotic and prokaryotic organisms [1]. For instance, Kmal has been enriched in key signaling molecules in mouse liver [2], plant cells [3] and the gram-positive bacterium *Saccharopolyspora spinosa*, etc. [4, 5]. Although many efforts have been devoted to investigating the cellular mechanisms of Kmal, its biological significance remains poorly understood [2, 6]. Recognition of malonylation sites in substrates represents an initial but crucial step in elucidating the molecular mechanisms underlying protein

\* Correspondence: [flyinsky6@gmail.com](mailto:flyinsky6@gmail.com); [changshui@hotmail.com](mailto:changshui@hotmail.com)

†Xin Liu and Liang Wang contributed equally to this work.

<sup>1</sup>Department of Bioinformatics, School of Medical Informatics and Engineering, Xuzhou Medical University, Xuzhou 221004, Jiangsu, China  
Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

malonylation. With the development of high-throughput mass spectrometry techniques, many Kmal-containing peptides have been identified [7, 8]. However, considering the dynamic properties and low abundance of malonylation and the limitation of experiment methods, identification of the exact substrates or sites on a large scale remains challenging.

To date, various computational tools have been developed to predict malonylation sites in protein sequences [9–14]. For instance, Xu et al. [9] used minimum Redundancy Maximum Relevance (mMRM) model to construct a prediction tool named Mal-Lys by incorporating residue sequence order information, position-specific amino acid propensity, and physicochemical properties for each peptide. Wang et al. [10] built a predictor called MaloPred, which took into accounts of five features including amino acid compositions (AAC), amino acids binary encoding (BINA), encoding based on grouped weight (EBGW), K nearest neighbors feature (KNN), and position specific scoring matrix (PSSM). Their information gains (IG) were then evaluated to select most meaningful and significant features. Hasan and Kurata [11] proposed a prediction tool called identification of Lysine-Malonylation Sites (iLMS), which used the composition of profile-based k-Spaced Amino Acid Pairs (pkSAAP), dipeptide amino acid compositions (DC) and amino acid index properties (AAindex) to encode the segment. Chen et al. [12] constructed a LSTM-based ensemble malonylation predictor (LEMP), which combined the long short-term memory (LSTM) algorithm with word embedding and the random forest algorithm with novel encoding of enhanced amino acid content. In addition, Taherzadeh et al. [13] developed the SPRINT-Mal tool and found that evolutionary information and physicochemical properties are the two most discriminative features. A structural feature called half-sphere exposure provides additional improvement to the prediction performance. Bao et al. [14] proposed the IMKPse model that utilized general PseAAC as the classification features and employed flexible neural tree as classification model. Although many achievements have been made in the prediction of malonyl acylation modification sites, there is still much room for improvement in the prediction performance.

In this study, we investigated whether dimensionality reduction algorithm PCA is useful for predicting malonylation sites. Another issue that we attempted to address here is whether the integration of sequence features could generate better prediction accuracy. On the basis of our results, Mal-Prec significantly outperformed existing predictors and indicated that PCA, together with three sequence features, one-hot encoding, physicochemical properties (AAindex), and composition of k-spaced amino acid pairs (CKSAAP), is able to

improve the accuracy of prediction. Thus, Mal-Prec could serve as a powerful tool for identifying malonylation sites in proteins.

## Results and discussion

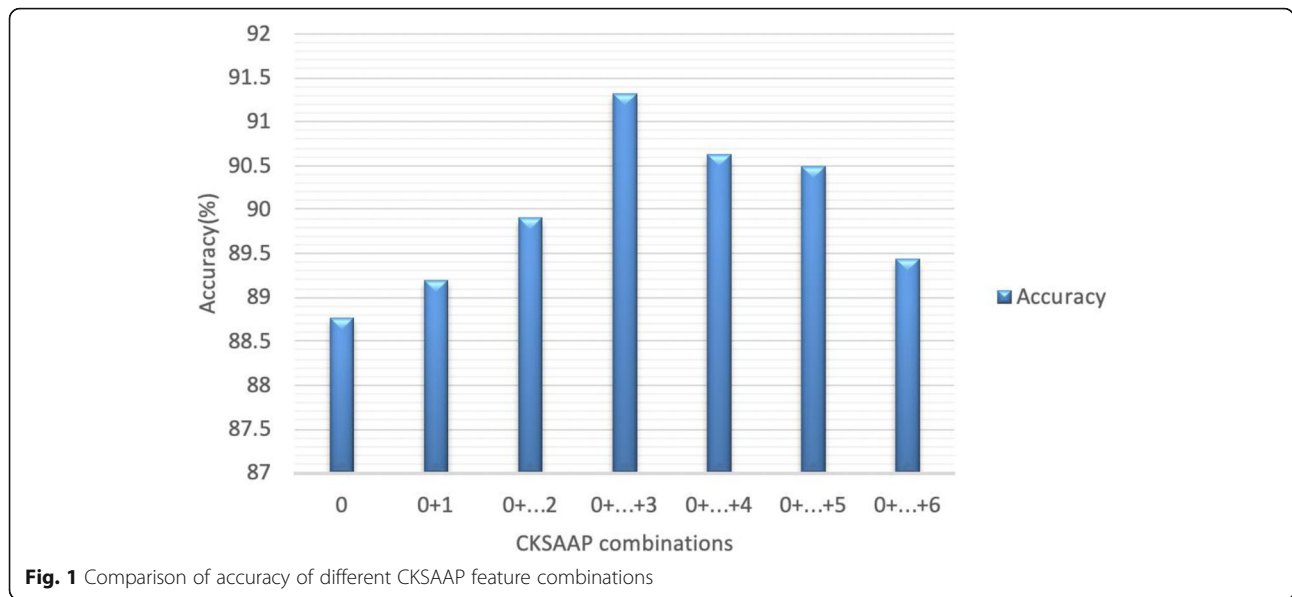
### Determination of CKSAAP features

Though many approaches have adopted CKSAAP features to predict PTM sites, most of them only used the CKSAAP features generated by single K value and did not identify optimal K for constructing the CKSAAP feature. In order to obtain valid CKSAAP features, we analyzed the performance of different combination of CKSAAP features. In particular, we not only analyzed the CKSAAP features obtained by single K value ranging from 0 to 6, but also analyzed their combined effects. All data sets were dimensioned to 100 using PCA. We used LIBSVM tool which is available on <https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/>. By using the grid search method, we optimized the two important parameters of SVM,  $c$  and  $g$ , which are the penalty parameters and kernel parameters respectively in the SVM algorithm. Finally, we set  $c = 10$  and  $g = 2$  in the SVM model and the radial basis function was adopted as the kernel function. 5-fold cross-validation was executed for 50 times to optimize the parameters in the training model. The results are shown in Supplementary Table 1, according to which, Acc, F1, and MCC do not change much under different K value. For example, the Sen value changes from 81.46 to 98.63%, the Spec value changes from 65.72 to 82.24%. Thus, it is difficult to figure out which is more suitable.

Thereafter, we made comparisons by combining all features together (CKSAAP, one-hot encoding, AAindex). The parameters  $c$  and  $g$  in SVM were set to 1.9 and 0.07 by grid search, respectively. The performance is shown in Supplementary Table 2, according to which, we can see that when K was set to 0 to 6, the performance of the proposed method did not change too much. Acc, Sen, Spec, F1, and MCC changed from 88.58 to 89.87%, 89.01 to 90.38%, 87.53 to 89.77%, 88.65 to 89.87%, and 79.79 to 81.81%, separately. When combining feature vectors computed by different K value, the result has a certain law, which is shown in Fig. 1, from which we could see that when we combined the first 4 CKSAAP features together, the accuracy achieves the best score, so do in the other four metrics. Thus, in this paper, we set K as 0, 1, 2, and 3, and got the CKSAAP feature vectors were  $441 \times 4 = 1764$ -dimensions.

### Effectiveness of PCA

In order to determine the suitable dimensions of PCA for our prediction, we run the training model when the dimensions equal to 50, 100, 150, 200, 250, and 300, separately. 5-fold cross-validation was executed for 50 times



**Fig. 1** Comparison of accuracy of different CKSAAP feature combinations

to optimize the parameters. The results are shown in Table 1.

In Table 1, when the dimensions equal to 100, the proposed method performed best, and average ACC, Sen, Spec, F1, and MCC can reach to 91.24, 91.71, 90.83, 91.18, and 84.03%, separately. Supplementary Figure 1 shows the accuracy curve in different dimensions. It is apparent that accuracy curve is a convex function. When the dimensions are equal to 100, the accuracy reaches the maximum of 91.24%. When the dimension value is greater than 100, larger the dimension gets, lower the accuracy is.

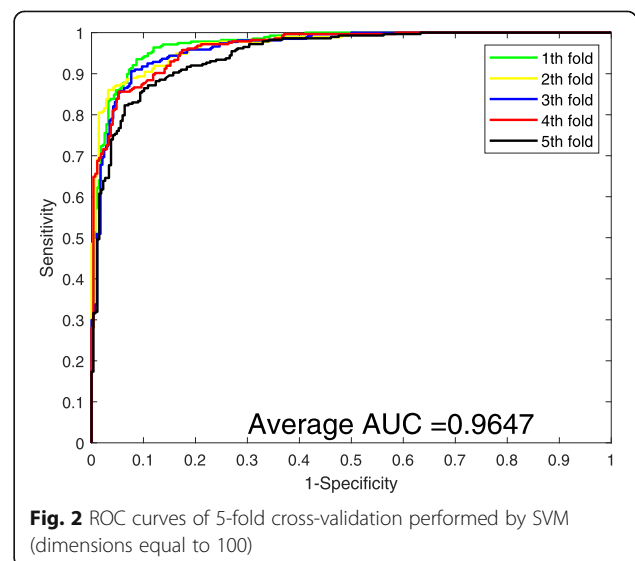
Supplementary Table 3 shows the performance of 5-fold cross-validation when implementing the proposed method on human data set. It can be seen that the average Acc, Sen, Spec, F1, and MCC can reach 91.24, 91.71, 90.83, 91.18, and 84.03%, separately. The standard deviations of these criteria values are 1.24, 2.50, 2.10, 1.43, and 2.09%, respectively. The ROC curves of the 5-fold cross-validation are listed in Fig. 2. The average AUC value is 96.47%.

For the purpose of analyzing the role of PCA in our proposed method, we applied the same procession of our proposed approach without PCA. The parameter c

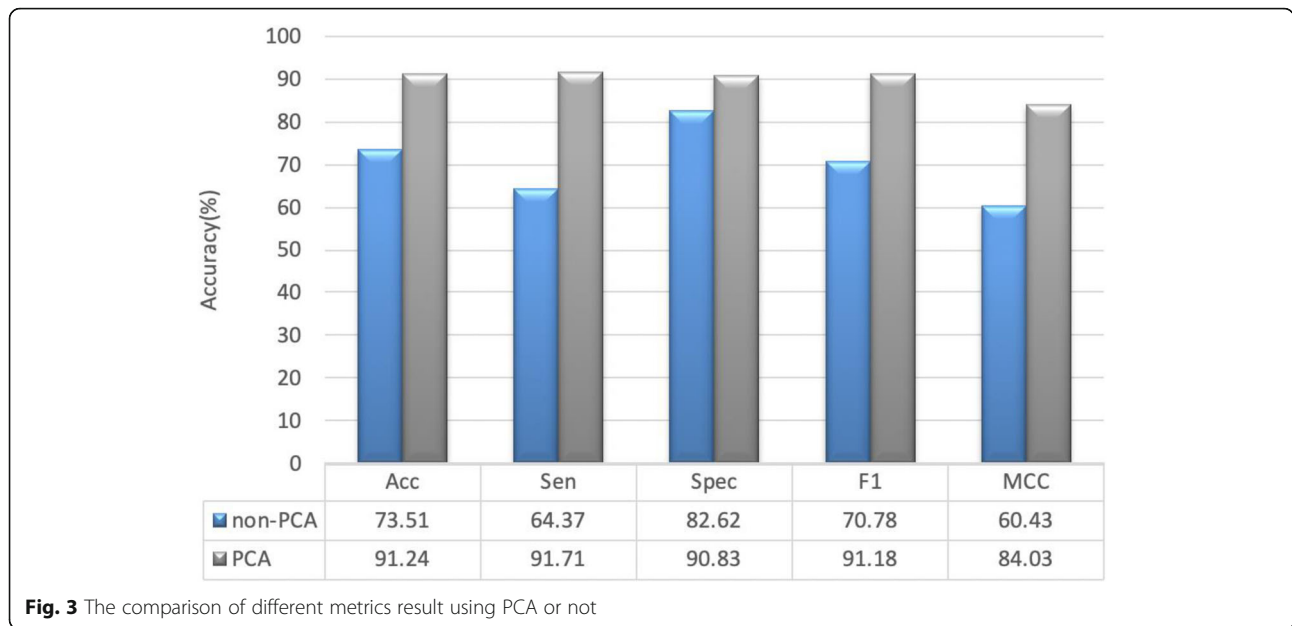
and g were set to 2 and 0.1 by grid search. The performance of the 5-fold cross-validation is shown in Supplementary Table 3, in which, the average Acc, Sen, Spec, F1, and MCC reach to 73.51, 64.37, 82.62, 70.78, and 60.43%. And the standard deviations of these criteria values are 1.99, 2.32, 2.63, 2.64, and 1.89%, respectively. For a more intuitive analysis, we adopted Fig. 3 to show the comparison of different metrics result using PCA or not. Non-PCA represents PCA was not used, PCA represents the dimensions are reduced to 100 using PCA. From Fig. 3 we could see that, by comparing to the proposed method without PCA, the average Acc, Sen, Spec, F1, and MCC of the proposed method with PCA could increase 17.73, 27.34, 8.21, 20.4, and 19.6%,

**Table 1** 5-fold cross-validation results of different dimensions

dimensions	Acc (%)	Sen (%)	Spec (%)	F1 (%)	MCC (%)
50	85.91	86.09	85.68	85.94	75.77
100	<b>91.24</b>	<b>91.71</b>	<b>90.83</b>	<b>91.18</b>	<b>84.03</b>
150	90.20	91.00	89.48	90.30	82.31
200	88.11	89.53	86.65	88.39	79.02
250	84.61	86.15	83.10	84.73	73.91
300	82.31	84.34	80.36	82.27	70.89



**Fig. 2** ROC curves of 5-fold cross-validation performed by SVM (dimensions equal to 100)



**Fig. 3** The comparison of different metrics result using PCA or not

respectively. That means PCA can effectively improve the performance of the algorithm.

**Performance comparison of different feature combination**

For the purpose of further identifying the role of various features, we analyzed the performance of each feature and multiple feature combinations. The performance comparison of each single feature was shown in Supplementary Table 4, from which we could see that the CKSAAP outperforms the other two features, especially in terms of ACC, Sen, Spec, and F1, which are almost 20% ~ 30% higher than the other two features. Meanwhile, while the performance comparison of multiple features was shown in Supplementary Table 5, which shows the performance of different features combination. The CKSAAP (exclude) means exclude the CKSAAP from the three features, so it represents the combination of AAindex and One-hot. The AAindex (exclude) and One-hot (exclude) also has the same meaning. All represents the combination of three features. From Supplementary Table 6 we could see that the combination of AAindex and One-hot performs best in all of those two features combined. It is interesting because we know CKSAAP performs best in the comparison of a single feature. Thus, we can use a Chinese saying to summarize this phenomenon, three cobblers combined makes a genius mind. This is to say, the combination of the three features works best. For a more intuitive analysis, we applied the column chart to show the performance comparison of the seven kinds of feature combination. In Supplementary Figure 2, the ECKSAAP means exclude the CKSAAP from the three features. The AAindex and One-hot also have similar

meaning. It can be seen that the proposed method which combined all features achieves best in all metrics. The ECKSAAP ranks second in terms of Acc, Spec, F1, and MCC.

According to the above analysis, after combining the four attributes of CKSAAP, one-hot encoding and nine attributes of AAindex, and then using PCA to reduce the dimension to 100, Mal-Prec can achieve better performance.

**Comparison of classical algorithms**

We also compared Mal-Prec with other four classical classifiers on the training data sets, including Random Forest (RF), K-nearest neighbors (KNN), Ensemble of decision tree and Naive Bayes (NB) [15–17]. The Euclidean distance was used in KNN algorithm, and the number of its neighbor is 2. The number of decision trees in RF and Ensemble was 20 and 50, separately. 5-fold cross-validation was conducted 50 times to each of them. The performance comparisons are shown in Table 2.

Even though it is well known that the ensemble classifier is more accurate and robust than individual

**Table 2** The performance comparisons of different classical classifiers

classifier	Acc (%)	Sen (%)	Spec (%)	F1 (%)	MCC (%)
KNN	59.68	26.73	92.18	34.34	34.98
NB	83.24	84.17	82.36	83.39	72.11
RF	68.25	62.41	74.04	66.27	56.36
Ensemble	64.11	60.11	68.20	62.50	53.69
Mal-Prec (SVM)	<b>91.24</b>	<b>91.71</b>	<b>90.83</b>	<b>91.18</b>	<b>84.03</b>

classifiers, it can be seen from Table 2 that, compared with other classical classifiers, Mal-Prec model performs best in all metrics. That means different data set requires different models.

### Performance on independent data set

For objective performance comparison, the independent data set which is truly blind to the training data set was adopted to evaluate the performance of the proposed method. As seen in Table 3, the proposed method performs best, including Acc, Sen, Spec, F1, and MCC values of 90.65, 89.71, 91.59, 90.62, and 83.04%, respectively.

Figure 4 shows the ROC curves from combinations of different features on the independent data set. It can be seen that, on the independent data set, the proposed method (all features) has a AUC value of 90.72%, the ECKSAAP ranks second, and the rest are ECKSAAP, EOne-hot, EAAindex, AAindex, One-hot, which are the same as the result on the testing data set. This further confirms that Mal-Prec constructed by incorporating those three features and PCA has a good effect.

### Comparison of the state-of-the-art approaches

We compared the proposed method with some state-of-the-art approaches for predicting malonylation sites, including Mal-Lys, MaloPred, iLMS, LEMP, SPRINT-Mal. Table 4 shows the comparison of the proposed method and some state-of-the-art approaches.

The reasons for the good performance of our proposed method can be summarized as two points. Firstly, PCA is utilized to extract features. PCA is a dimensionality reduction method, which extracts more effective characteristic information. Secondly, the support vector machine classifier is used for classification. All the above proves that the SVM classifier combined with principal component analysis and three features (PseAAC, One-hot, CKSAAP) is more suitable for predicting the malonylation sites than the state-of-the-art approaches.

**Table 3** Performance of different feature combinations on the independent data set

Features	Acc (%)	Sen (%)	Spec (%)	F1 (%)	MCC (%)
CKSAAP	77.55	77.14	77.97	77.59	65.18
AAindex	61.73	65.43	57.97	63.26	52.61
One-hot	58.71	61.43	55.94	59.97	51.44
CKSAAP (exclude)	86.19	86.86	85.51	86.36	76.19
AAindex (exclude)	79.42	80.57	78.26	79.77	67.30
One-hot (exclude)	71.08	76.29	65.80	72.65	58.65
ALL	90.65	89.71	91.59	90.62	83.04

### Feature analysis

We also analyzed sequence occurrence frequency on every position using Two Sample Logo with t-test ( $P$ -value  $< 0.05$ ). Figure 5 shows that the malonylation and non-malonylation peptides have considerably different sequence preferences. Glycine (G), Leucine (L), Alanine (A), and Valine (V) were significantly richer than those in non-malonylation ones. However, Lysine (K) and Glutamic acid (E) were much abundant in non-malonylation peptides. Thus, we believe that the difference between the two peptides could be a new method to distinguish them.

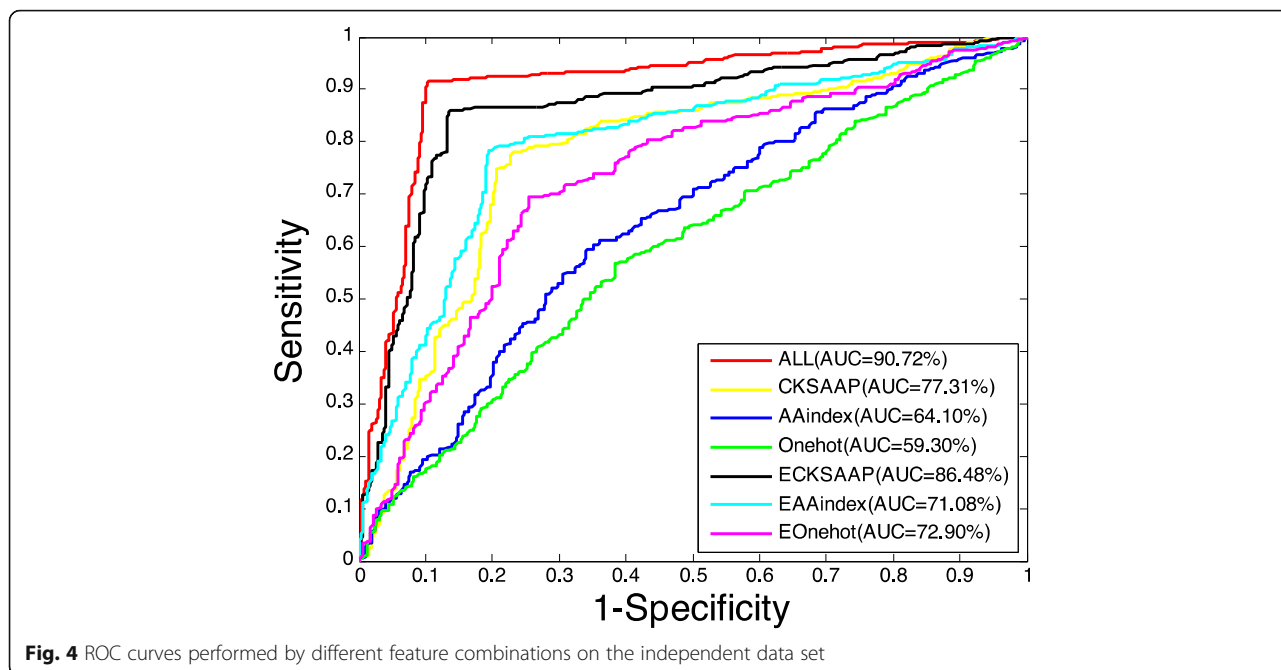
### Conclusions

In this study, a novel method entitled Mal-Prec was developed to predict human malonylation sites. The best prediction performance was achieved when using PCA to reduce the dimensionality of feature combination (CKSAAP, AAindex, and one-hot) to 100, rather than combined those features all together. By individual comparison of three features (CKSAAP, AAindex, and One-hot), we found that CKSAAP with the incorporation of the first four features, performed best. While the AAindex and one-hot combination performed best in two features combination. This indicated that simply incorporating more features may not achieve the best results. Based on the results obtained by 5-fold cross-validation, Mal-Prec remarkably outperforms existing predictors and could serve as a useful tool for identifying and discovering novel malonylation sites in human proteins. In addition, although good performance has been obtained by using Mal-Prec, there is still space for the method to be refined. First of all, more peptide features, such as structure properties, evolutionary information, and so on, could be incorporated for the prediction. In future, we will take more feature constructions into account to achieve better prediction performance. Secondly, we have not solved the data set imbalance problem. Down-sampling method is popular but not good enough for data set imbalance. We will introduce other approaches to solve the imbalance problem, such as one-side selection (OSS) and sampling based on clustering (SBC), etc. Finally, we are planning to develop a webserver for the method, by doing which other researchers could try this novel method for malonylation site predictions.

### Methods

#### Data collection and preprocessing

In this study, the data sets were retrieved from literature [10, 13]. A total of 1768 sequence fragments from 934 human proteins were collected. To reduce the redundancy and avoid artificial bias, CD-HIT was employed to remove redundant sequences with equal to or more than



40% similarities [18]. Then the processed sequences were truncated into 17-residue long sequence segments with lysine (K) located at the center. Each of peptide fragment was defined as follows:

$$P = R_{-n}R_{-n+1} \dots R_{-1}KR_1R_2 \dots R_\epsilon \tag{1}$$

Where  $R_\epsilon$  represents the  $\epsilon$ -th downstream peptide from the center K while  $R_{-n}$  represents the  $n$ -th upstream sequence fragment, and so forth. The length of the sequence fragment is  $n + \epsilon + 1$ . Since there might be fewer amino acids around the center K, as shown in Supplementary Figure 3, the downstream peptide from the center K is less than  $\epsilon$ , so we can use X to fill in those residues. Thus, the dataset was made up of 20 native amino acids and the dummy code X. Different studies may select varied length of malonylation peptide

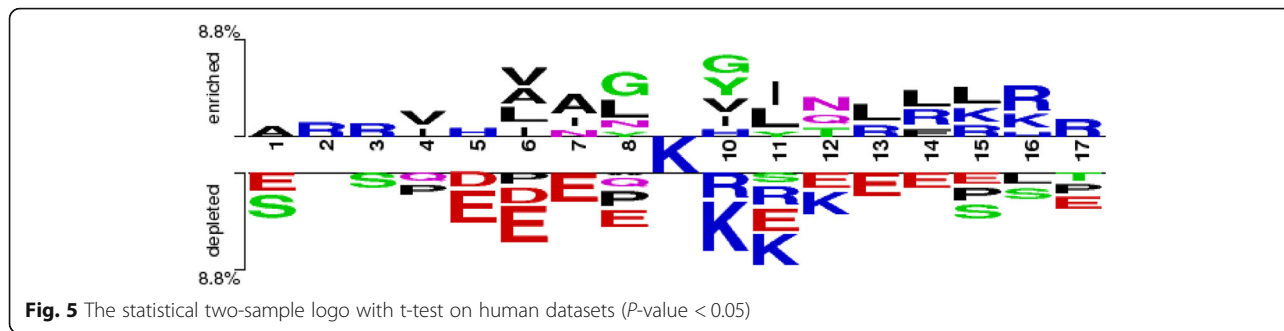
segments for analysis. In this project, we set  $\epsilon = 8$  and  $n = 8$  and the length of the peptide segment is 17. Thus, the complete sequence segment P describing a lysine belongs to either of two classes ( $\delta_1, \delta_2$ ). If the represented lysine is a malonylation site, then  $\delta_1 = 0$ , otherwise  $\delta_2 = 1$ .

$$P \in (\delta_1, \delta_2)^T \quad \delta_1, \delta_2 \in (0, 1)^T \tag{2}$$

Accordingly, 1735 sequence fragments from 931 human proteins were selected as positive dataset. Sequence fragments around lysine (abbreviated as Lys or K) that are not included in the positive data set were constituted as negative dataset. After doing all of this, we obtained 45,607 negative samples. Unbalanced dataset may lead to false prediction, hence we used the down-sampling method to construct a balanced dataset [19]. Therefore, our data set is balanced which contains 3470 sets of

**Table 4** Comparison of state-of-the-art approaches in terms of Acc and AUC in different organisms

Approach	Feature	Species	Acc (%)	AUC (%)
mRMR+SVM [9]	K-gram+AAindex	N/A	N/A	79.35
IG + SVM [10]	AAC + BINA (sequence-based)	<i>E. coli</i>	72.30	75.50
	EBGW (physicochemical)	Mouse	74.65	82.70
	KNN + PSSM (evolutionary)	<i>Homo sapiens</i>	73.72	87.10
IG + SVM [11]	PKsaap+AAindex+DC	Mouse	N/A	73.90
		<i>Homo sapiens</i>	N/A	74.30
LSTM+RF [12]	EAAC+word embedding	Mouse	88.00	82.40
		<i>Homo sapiens</i>		
SVM [13]	Binary+PSSM+AAindex+Structured (ASA + SS + HSE + IDR)	Mouse	N/A	76.00
Proposed method	AAindex+One-hot+CKSAAP	<i>Homo sapiens</i>	<b>90.65</b>	<b>90.91</b>



data, half of the positive and negative sets. In order to validate the performance of the predictor, we split 20% of the dataset (695) as independent dataset, the remaining are training dataset (2775).

**Flowchart of the proposed method**

Flowchart of the malonylation site prediction method Mal-Prec proposed in this paper is shown in Fig. 6. The prediction steps of the Mal-Prec are described as follows:

1. Data collection and preprocessing. Dataset was collected through literature and NCBI websites. Sliding window was then used to select a peptide having a length of 17 with lysine at the center point. Positive data set and negative data set were constructed with equal quantity by down-sampling method.
2. Feature representation. CKSAAP, AAindex, and One-hot coding method were chosen as features to represent each peptide segment in this study.
3. Dimensionality reduction. High dimensional data set may lead to the curse of dimensionality [19]. To

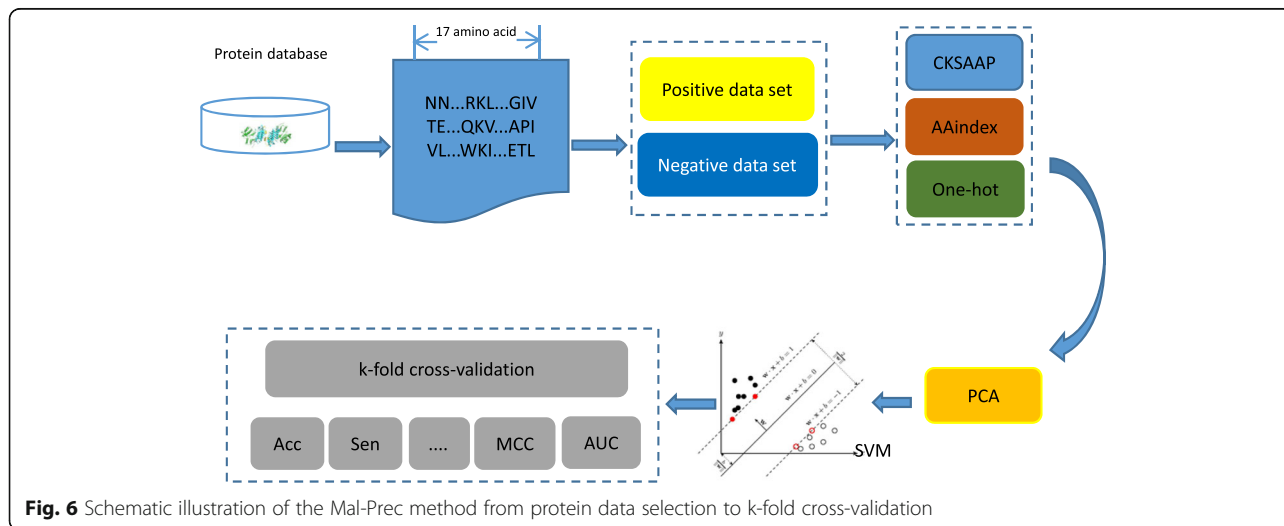
solve this problem, we used PCA for dimensionality reduction, and also analyzed the suitable dimension of the data set.

4. Classification. Different data requires corresponding algorithms [20]. Comparing to other classical classifier algorithms, we chose SVM as classification algorithm in Mal-Prec.
5. Model performance evaluation. To find the suitable parameters and avoid potential over-fitting issue, we adopted the 5-fold cross-validation algorithm and employed classical metrics, such as Acc and Sen, etc., to assess the performance of the algorithm.

**Feature construction**

**Binary encoding (one-hot encoding)**

Binary encoding is also called one-hot encoding, which could transform amino acids into orthogonal numeric vectors, and has been applied in many protein sequence analyses. Since there are 21 types of amino acids (20 conventional amino acids and 1 pseudo amino acid X), each peptide sequence can be represented as a 21-dimensional vector. For example, the protein sequence is ‘ACDEFGHIKLMNPQRSTVWYX’. Thus, alanine (A) is







Matthews correlation coefficient (MCC). The selected performances have been demonstrated in eqs. (5)–(9).

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (5)$$

$$\text{Sen} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{Spec} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (7)$$

$$\text{F1} = 2 \times \frac{\text{SN} \times \text{PPV}}{\text{SN} + \text{PPV}} \quad (8)$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FP}) \times (\text{TN} + \text{FN})}} \quad (9)$$

Where TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives and false negatives, respectively [44]. In addition, the receiver operating characteristic (ROC) curves are plotted based on Sen and Spec by taking different thresholds [45] and their area under the ROC (AUC) values were also calculated based on the trapezoidal approximation [46].

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-020-07166-w>.

**Additional file 1: Table S1.** The performance of the proposed method using different CKSAAP features. **Table S2.** Performance of the proposed method using different CKSAAP combinations. **Table S3.** The performance of 5-fold cross-validation (dimensions equal to 100). **Table S4.** The performance of 5-fold cross-validation without PCA. **Table S5.** The performance comparison of different single feature. **Table S6.** The performance comparison of different feature combination.

**Additional file 2: Figure S1.** Comparison of accuracy in different dimensions got by using PCA. **Figure S2.** The comparison of different feature combinations. **Figure S3.** Schematic diagram of malonylation sequence fragment. X represents filled-up residues in the fragment. n and ε represent the n-th upstream peptide and ε-th downstream peptide from the center K, respectively. **Figure S4.** Transformation of the 17-amino-acid peptide VAERAALEKLDANQEYK into a 17\*21 dimensional vector after one-hot encoding process.

## Abbreviations

Mal-Prec: Malonylation Prediction; PTM: Post-Translational Modification; T2DM: Type 2 Diabetes Mellitus; PCA: Principal Component Analysis; SVM: Support Vector Machine; AAindex: Amino Acid Index Properties; CKSAAP: Composition of K-Spaced Amino Acid Pairs

## Acknowledgements

Not applicable.

## Authors' contributions

XL and XZ proposed the core ideas of the project. XL and LW collected and processed the data, performed the experiments, and contribute to the writing of the manuscript. JL and JH critically reviewed and revised the manuscript. All authors read and approved the final manuscript and consent the publication of this study.

## Funding

Prof. Xin Liu greatly appreciated the funding by Xuzhou Science and Technology Project (KC17123), Jiangsu Postdoctoral Science Foundation

(1601080C, 1701062B), Jiangsu University Natural Science Foundation (17KJB310015), and Research Foundation for Talented Scholars in Xuzhou Medical University (D2015001). Prof. Liang Wang acknowledged the financial support of National Natural Science Foundation of China (31900022), Natural Science Foundation of Jiangsu Province (BK20180997), Jiangsu Qinglan Project (2020), and the Funding of Innovative Science and Technology Team of Young Scientists at Xuzhou Medical University (TD202001).

## Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

Not applicable.

## Author details

<sup>1</sup>Department of Bioinformatics, School of Medical Informatics and Engineering, Xuzhou Medical University, Xuzhou 221004, Jiangsu, China. <sup>2</sup>Jiangsu Key Laboratory of New Drug Research and Clinical Pharmacy, School of Pharmacy, Xuzhou Medical University, Xuzhou 221000, Jiangsu, China. <sup>3</sup>School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA 70118, USA.

Received: 23 April 2020 Accepted: 20 October 2020

Published online: 23 November 2020

## References

- Peng C, Lu Z, Xie Z, Cheng Z, Chen Y, Tan M, Luo H, Zhang Y, He W, Yang K, et al. The First Identification of Lysine Malonylation Substrates and Its Regulatory Enzyme. *Mol Cell Proteomics*. 2011;10:12.
- Xie Z, Dai J, Dai L, Tan M, Cheng Z, Wu Y, Boeke JD, Zhao Y. Lysine Succinylation and lysine Malonylation in histones. *Mol Cell Proteomics*. 2012;11(5):100–7.
- Colak G, Pougovkina O, Dai L, Tan M, te Brinke H, Huang H, Cheng Z, Park J, Wan X, Liu X, et al. Proteomic and biochemical studies of lysine Malonylation suggest its Malonic Aciduria-associated regulatory role in mitochondrial function and fatty acid oxidation. *Mol Cell Proteomics*. 2015; 14(11):3056–71.
- Foster DW. Malonyl-CoA: the regulator of fatty acid synthesis and oxidation. *J Clin Invest*. 2012;122(6):1958–9.
- Liu J, Wang G, Lin Q, Liang W, Gao Z, Mu P, Li G, Song L. Systematic analysis of the lysine malonylation in common wheat. *BMC Genomics*. 2018;19:1.
- Nishida Y, Rardin Matthew J, Carrico C, He W, Sahu Alexandria K, Gut P, Najjar R, Fitch M, Hellerstein M, Gibson Bradford W, et al. SIRT5 regulates both cytosolic and mitochondrial protein Malonylation with glycolysis as a major target. *Mol Cell*. 2015;59(2):321–32.
- Hirschey MD, Zhao Y. Metabolic regulation by lysine Malonylation, Succinylation, and Glutarylation. *Mol Cell Proteomics*. 2015;14(9):2308–15.
- Bao X, Zhao Q, Yang T, Fung YME, Li XD. A chemical probe for lysine Malonylation. *Angew Chem Int Ed*. 2013;52(18):4883–6.
- Xu Y, Ding Y-X, Ding J, Wu L-Y, Xue Y. Mal-Lys: prediction of lysine malonylation sites in proteins integrated sequence-based features with mRMR feature selection. *Sci Rep*. 2016;6:1.
- Wang L-N, Shi S-P, Xu H-D, Wen P-P, Qiu J-D. Computational prediction of species-specific malonylation sites via enhanced characteristic strategy. *Bioinformatics*. 2016.
- Hasan MM, Kurata H. iLMS, Computational Identification of Lysine-Malonylation Sites by Combining Multiple Sequence Features. In: 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE); 2018. p. 356–9.
- Chen Z, He N, Huang Y, Qin WT, Liu X, Li L. Integration of a deep learning classifier with a random Forest approach for predicting Malonylation sites. *Genom Proteomics Bioinformatics*. 2018;16(6):451–9.

13. Taherzadeh G, Yang Y, Xu H, Xue Y, Liew AW-C, Zhou Y. Predicting lysine-malonylation sites of proteins using sequence and predicted structural features. *J Comput Chem*. 2018;39(22):1757–63.
14. Bao W, Yang B, Huang D-S, Wang D, Liu Q, Chen Y-H, Bao R. IMKPse: identification of protein Malonylation sites by the key features into general PseAAC. *IEEE Access*. 2019;7:54073–83.
15. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
16. Patrick EA, Fischer FP. A generalized k-nearest neighbor rule. *Inf Control*. 1970;16(2):128–52.
17. Webb GI, Boughton JR, Wang Z. Not so naive Bayes: aggregating one-dependence estimators. *Mach Learn*. 2005;58(1):5–24.
18. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics*. 2010;26(5):680–2.
19. Allen Elena A, Erhardt Erik B, Calhoun Vince D. Data visualization in the neurosciences: overcoming the curse of dimensionality. *Neuron*. 2012;74(4):603–8.
20. Ali S, Smith KA. On learning algorithm selection for classification. *Appl Soft Comput*. 2006;6(2):119–38.
21. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci*. 1981;78(6):3824–8.
22. Radzicka A, Wolfenden R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*. 2002;27(5):1664–70.
23. Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol*. 1968;21(2):170–201.
24. Treece JM, Sheinson RS, McMeekin TL. The solubilities of  $\beta$ -lactoglobulins a, B, and AB. *Arch Biochem Biophys*. 1964;108(1):99–108.
25. Bhaskaran R, Ponnuswamy PK. Positional flexibilities of amino acid residues in globular proteins. *Int J Pept Protein Res*. 2009;32(4):241–55.
26. Chothia C. Structural invariants in protein folding. *Nature*. 1975;254(5498):304–8.
27. Cosic I. Macromolecular bioactivity: is it resonant interaction between macromolecules?—theory and applications. *IEEE Trans Biomed Eng*. 1994;41(12):1101–14.
28. Bull HB, Breese K. Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues. *Arch Biochem Biophys*. 1974;161(2):665–70.
29. Eisenberg D, Weiss RM, Terwilliger TC. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Natl Acad Sci*. 1984;81(1):140–4.
30. Huang K-Y, Kao H-J, Hsu JB-K, Weng S-L, Lee T-Y. Characterization and identification of lysine glutarylation based on intrinsic interdependence between positions in the substrate sites. *BMC Bioinformatics*. 2019;19:S13.
31. Wang X, Yan R, Song J. DephosSite: a machine learning approach for discovering phosphatase-specific dephosphorylation sites. *Sci Rep*. 2016;6:1.
32. Chen Z, Zhou Y, Song J, Zhang Z. hCKSAAP\_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*. 2013;1834(8):1461–7.
33. Ju Z, Wang S-Y. Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition. *Gene*. 2018;664:78–83.
34. Ahmed MS, Shahjaman M, Kabir E, Kamruzzaman M. Prediction of protein acetylation sites using kernel naive Bayes classifier based on protein sequences profiling. *Bioinformation*. 2018;14(05):213–8.
35. Cui X, Yu Z, Yu B, Wang M, Tian B, Ma Q. UbiSitePred: a novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components. *Chemom Intell Lab Syst*. 2019;184:28–43.
36. Noble WS. What is a support vector machine? *Nat Biotechnol*. 2006;24(12):1565–7.
37. Cui G, Fang C, Han K. Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics*. 2012;13(Suppl 7):S5.
38. Huang S, Cai N, Pacheco P, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics Proteomics*. 2018;15:1.
39. Huang Y-F, Chen S-Y. Protein secondary structure prediction based on physicochemical features and PSSM by SVM. In: 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB); 2013. p. 9–15.
40. Jolliffe IT. Principal component analysis; 2002.
41. Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. *Bioinformatics*. 2001;17(9):763–74.
42. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904–9.
43. Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans Pattern Anal Mach Intell*. 2010;32(3):569–75.
44. Forbes AD. Classification-algorithm evaluation: five performance measures based on confusion matrices. *J Clin Monit*. 1995;11(3):189–206.
45. Landgrebe TCW, Duin RPW. Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis. *IEEE Trans Pattern Anal Mach Intell*. 2008;30(5):810–22.
46. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr*. 2008;17(2):145–51.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

