Genomics & Informatics

RESEARCH

Open Access

Rore: robust and efficient antioxidant protein classification via a novel dimensionality reduction strategy based on learning of fewer features

Chaolu Meng[1,2], Yongqi Hou[3], Quan Zou[4], Lei Shi[5], Xi Su[6] and Ying Ju[7*]

Abstract

In protein identification, researchers increasingly aim to achieve efficient classification using fewer features. While many feature selection methods effectively reduce the number of model features, they often cause information loss caused by merely selecting or discarding features, which limits classifier performance. To address this issue, we present Rore, an algorithm based on a feature-dimensionality reduction strategy. By mapping the original features to a latent space, Rore retains all relevant feature information while using fewer representations of the latent features. This approach significantly preserves the original information and overcomes the information loss problem associated with previous feature selection. Through extensive experimental validation and analysis, Rore demonstrated excellent performance on an antioxidant protein dataset, achieving an accuracy of 95.88% and MCC of 91.78%, using vectors including only 15 features. The Rore algorithm is available online at http://112.124.26.17:8021/Rore.

Keywords Dimensionality reduction, Robust feature, Protein classifier, Protein sequence

1 Introduction

Antioxidant proteins produced by the human body can resist free radical damage. Identifying which human proteins are antioxidant can help prevent diseases such as cancer and cardiovascular disease [1–4]. For this purpose, machine learning models can be used. Model's fewer features can improve the interpretability of the model and help researchers understand the underlying biological mechanisms [5–13]. Antioxidant protein identification based on machine learning has been performed in the past [14–17]. In 2016, Feng et al. developed the AodPred model based on optimal 3-gap dipeptides for feature selection to obtain 158-dimensional features for classification [18]. In 2020, Chun et al. identified 9808-dimensional hybrid features using 188D, N-gram, ACC-PSSM, and g-gap. The authors performed feature selection using MRMD, t-SNE, and optimal feature set selection methods and classified the resulting feature vectors using the random forest algorithm [19]. In 2023, Chao et al. used MRMD and dynamic programming to select of 473D feature vectors and to obtain 17-dimensional feature vectors for classification [20]. In general, the methods used in these studies use feature selection to reduce the

*Correspondence:
Ying Ju
yju@xmu.edu.cn
[1] College of Computer and Information Engineering, Inner Mongolia Agricultural University, Hohhot, China
[2] Inner Mongolia Autonomous Region Key Laboratory of Big Data Research and Application of Agriculture and Animal Husbandry, Hohhot, China
[3] School of Computer Science, Inner Mongolia University, Hohhot, China
[4] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China
[5] Department of Spine Surgery, Changzheng Hospital, Naval Medical University, Huangpu District, No. 415, Fengyang Road, Shanghai, China
[6] Foshan Women and Children Hospital, Foshan, China
[7] School of Informatics, Xiamen University, Xiamen, China

antioxidant protein dimensions, which inevitably leads to information loss. Most importantly, potential relationships between the features are ignored, limiting the performance of the classifiers. In addition, unbalanced datasets for antioxidant protein identification lead to low MCC values, which indicates poor model performance in predicting a small number of classes [21–24]. Notably, in clinical applications, misdiagnoses caused by targeting only few classes may have fatal consequences [25–29]. To address this problem, we propose the Rore model, which achieves superior MCC values based on a feature dimensionality reduction algorithm and also preserves the original information and relationships between the features. The usage of a SMOTE algorithm that rebalances the dataset. Specifically, the Rore model was found to achieve 95.88% accuracy and a 91.78% MCC value using only 15-dimensional features. Our findings demonstrated Rore shows significant performance with fewer features.

2 Methods

To address the problem of information loss, we developed a classification method based on a new feature-dimensionality reduction strategy called Rore. For this purpose, we first constructed an antioxidant protein dataset containing 3104 samples by screening and rebalancing data in the UniProt database using the SMOTE algorithm. Subsequently, a 473-dimensional feature vector was extracted (Fig. 1). To reduce feature dimensionality and maximize the retention of the original feature information, we propose a feature dimensionality reduction method, the variational feature compressor (VFC). This method is based on the idea of an information bottleneck. Through VFC processing, we obtain a final 15-dimensional feature vector. Finally, this 15-dimensional feature vector was fed to the XGBoost algorithm for classification.

2.1 Benchmark dataset

We used the dataset collected in previous studies to allow a fair and comprehensive performance comparison with existing methods [18, 30, 31]. This dataset is unbalanced. The positive sample consists of sequences labeled as "antioxidant" from the UniProt database [32]. Samples containing "B," "X," "Z," "O," "U," and "J" were eliminated due to uncertainty regarding their meanings [33], and only the sequences labeled "review" were retained for experimental validation. Negative samples were obtained from the PDB database using the PISCES procedure with an identification rate of no more than 20% (values less than 20%). The resulting dataset contained 1805 protein sequences, with 253 positive and 1552 negative samples.

$$\text{Data} = \text{positive}_{+} + \text{negative}_{-}$$

Unbalanced datasets can lead to overfitting when training models. Thus, we rebalanced the dataset using the positive sample oversampling method, SMOTE [34], to

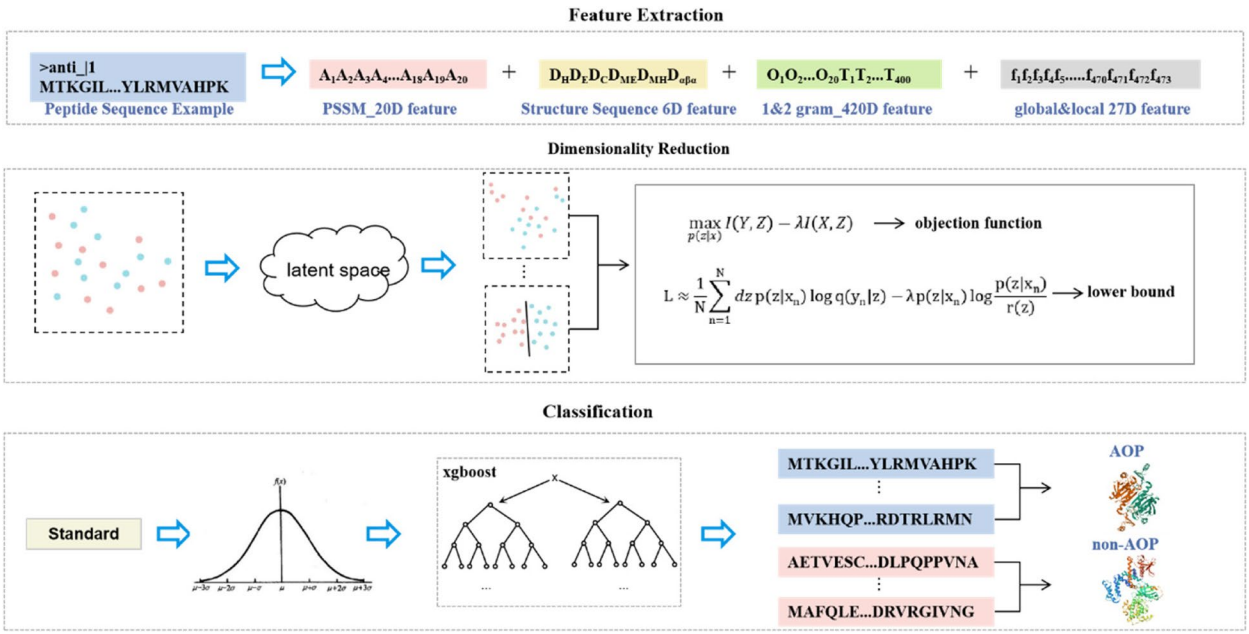

Fig. 1 Overall structure of the Rore classifier. The Rore model is built in three steps: extracting 473-dimensional feature vectors from protein sequences, using the feature extraction method proposed by Wei, and selecting the most informative vectors among these according to the VFC feature dimensionality reduction method. Finally, the model is trained using the XGBoost algorithm

obtain a 1:1 balanced dataset. The algorithm randomly selects a similarity point from the M similarity points closest to the sample point [35], and a new data point is generated through a linear link between the sample point and the randomly selected similarity point. After processing using the SMOTE algorithm, a dataset containing 3104 samples was obtained.

2.2 Feature extraction

We used the feature extraction method proposed by Wei et al. [36]. This method uses PSI-BLAST [37] and PSI-PRED [38] algorithms to extract features. These two algorithms obtain feature information from the sequence and structural points of view [39–41]. We reasoned that the complementarity of these two types of features can improve the prediction accuracy. The specific steps were as follows:

The PSI-BLAST algorithm was first used to obtain a position-specific score matrix. Twenty feature vectors are obtained from the job-specific score matrix. The position-specific score matrix can be expressed as follows:

$$S_{pssm} = \begin{bmatrix} pse_{1,1} & \cdots & pse_{1,20} \\ pse_{2,1} & \cdots & pse_{2,20} \\ \vdots & \ddots & \vdots \\ pse_{20,1} & \cdots & pse_{20,20} \end{bmatrix}$$

where the score for mutation of residues at position i in the protein sequence S to j-type residues is denoted as $pse_{i,j}$. The 20 feature vectors to be extracted were the mean values D_{pssm} of the mutated residues based on the 20 different types of mutated residues obtained during the evolutionary process [42], and D_{pssm} can be denoted as follows:

$$D_{pssm} = \left\{ \overline{A_j} = \frac{1}{L} \sum_{i=1}^{L} A_{ij}; 1 \le j \le 20 \right\}$$

where A_j denotes the mean value of the mutations occurring in the jth residue type during evolution.

In order to extract features from sequences containing richer evolutionary information, it is first necessary to transform each $A_{i,j}$ into a consensus sequence $A_{i,j}{}'$ with the following transformation formula:

$$A_{ij}{}' = 2^{A_{ij} \times BF_j}$$

where BF_j is obtained by dividing the number of amino acids by the number of sequences for all sequences in the PDB25 database [43]. Then, only a maximum value remains each row, and the new sequence A_{con} obtained is the consensus sequence containing evolutionary information.

Next, 20 and 400 features were extracted from the consensus sequence using 1-g and 2-g algorithms, respectively [44]. The feature extraction using the 1-g and 2-g algorithms can be formulated as follows:

$$D_{1-gram} = \left\{ \frac{O(P_i)}{L}; 1 \le i \le 20 \right\}$$

$$D_{2-gram} = \left\{ \frac{O(P_i P_j)}{L}; 1 \le i \le 20, 1 \le j \le 20 \right\}$$

where P_i denotes the residue i, $O(P_i)$ is the frequency of occurrence of the residue, and $O(P_i P_j)$ denotes the frequency of occurrence of residue pairs. Using a weighted combination of the 1-g and 2-g algorithms [45, 46], 420 features D_{con} were obtained, which can be denoted as follows:

$$D_{con} = \left\{ \frac{20 D_{1-gram}}{420}, \frac{400 D_{2-gram}}{420} \right\}$$

PSI-PRED algorithm allows obtaining sequences and matrices of information about the structure from which features can be extracted, with 6 features for the former and 27 features for the latter. Where the secondary structure sequence is noted as $A_{str} = S_1 S_2 S_3 \cdots S_L (S \in \{H, E, C\})$, H, E, and C denote the three states. The six features extracted from the sequence of relevant structural information obtained using the PSI-PRED algorithm were as follows:

$$D_H = \frac{\sum_{i=1}^{total_H} I_{H_i}}{L(L-1)}$$

$$D_E = \frac{\sum_{i=1}^{total_E} I_{E_i}}{L(L-1)}$$

$$D_C = \frac{\sum_{i=1}^{total_C} I_{C_i}}{L(L-1)}$$

$$D_{ME} = \frac{maxE}{L}$$

$$D_{MH} = \frac{maxH}{L}$$

where $total_H$, $total_E$, and $total_C$ are the sums of A_{str} in three states and I_H, I_E, and I_C are the location indices of the three states. maxE and maxH represent the maximum continuous lengths of the two states in space. D_{ME} and D_{MH} are the normalized maximum lengths. Replacing consecutive helices in the secondary structure sequence

with α and consecutive strands with β, ignoring coiled coils, results in a set of fragment sequences consisting of α and β. Based on the difference in the α and β arrangement in α/β proteins and α + β proteins, the two proteins can be distinguished by the following features:

$$D_{\alpha\beta\alpha} = \frac{total_{\alpha\beta\alpha}}{L-2}$$

where $total_{\beta\alpha\beta}$ is the total number of occurrences of βαβ segments in the segmentation sequence A_{str}.

From the structural correlation matrix, we have extracted 27 global structural features. Local structural features were extracted from the structural correlation matrix. Matrix consists of L rows and three columns, with each column representing the three states. The structural probability matrix SPM_{pro} can be expressed as follows:

$$SPM_{pro} = \begin{bmatrix} p_{1,C} & p_{1,H} & p_{1,E} \\ p_{2,C} & p_{2,H} & p_{2,E} \\ \vdots & \vdots & \vdots \\ p_{L,C} & p_{L,H} & p_{L,E} \end{bmatrix}$$

where $p_{i,C}\, p_{i,H}\, p_{i,E}$ are the probabilities of states C, H, and E, respectively. L is the size of the protein sequence. Based on the structure probability matrix, 3 and 24 features can be obtained from the global and local perspectives, respectively, where the global structural features can be denoted as follows:

$$D_{global\_C} = \frac{1}{L}\sum_{i=1}^{L} p_{i,C}$$

$$D_{global\_H} = \frac{1}{L}\sum_{i=1}^{L} p_{i,H}$$

$$D_{global\_E} = \frac{1}{L}\sum_{i=1}^{L} p_{i,E}$$

The local structural features were obtained by dividing the structural correlation matrix obtained according to the PSI-PRED algorithm into eight submatrices by row, and each submatrix consisted of three columns. The computation of the features for each submatrix is consistent with the computation of the global structural features. We obtained the following eight local structural features:

$$D_{local} = \{D_{local\_1C}, D_{local\_1H}, D_{local\_1E}, \cdots, D_{local\_8C} D_{local\_8H} D_{local\_8E}\}$$

where $D_{local\_iC}, D_{local\_iH}, D_{local\_iE}$ denote the probabilities of the submatrix in the three states; that is, 24 local structural features are obtained from the structural probability matrix.

In conclusion, the PSI-BLAST algorithm obtained 440 features, including 20 features, using a location-specific score matrix, and 420 features based on a weighted combination of 1-g and 2-g algorithms. In addition, PSI-PRED extracted 33 features, including 6 features based on the secondary structure sequence and 3 and 24 features based on the structural probability matrix from global and local perspectives, respectively. The total number of features was set to 473.

In conclusion, the PSI-BLAST algorithm obtained altogether 440 features, including 20 features by means of a location-specific score matrix and 420 features based on a weighted combination of 1-g and 2-g algorithms. In addition, PSI-PRED extracted 33 features, including 6 features based on the secondary structure sequence, and 3 features and 24 features based on the structural probability matrix from global and local perspectives, respectively. The total number of features is 473.

2.3 Variational feature compressor

To make the variational information bottleneck [47] (VIB) available for improved feature dimensionality reduction, we propose a variational feature compressor (VFC). In order to extract from the 473D feature vectors all possible representations of nonlinear interaction effects between features, it is necessary to convert the 473D feature vectors into more compact nonlinear latent variables [48]. The conversion process can be represented as follows:

$$F_{potential} = RELU(Dense(F_{473D}))$$

where F_{473D} denotes the 473D feature vector obtained from 473D feature extraction [36], dense denotes the fully connected layer, and RELU is the activation function that captures the key to nonlinear relationships.

We then adopted the idea of an information bottleneck to reduce feature dimensionality. Specifically, the goal of the idea is to compress the input-redundant features R to obtain the potential variable P that can maximally represent its category C, that is, minimize the mutual information M(R,P) and maximize the mutual information M(C,P) [49]. Based on the information bottleneck theory, the above process can be regarded as a maximization problem with the following formula:

$$\max_{p(p|r)} M(C,P) - \lambda M(R,P)$$

where λ is the Lagrangian quantity and M denotes the mutual information operation between two variables. The mutual information is calculated as follows:

$$\mathrm{M(R,P)} = \int \mathrm{q(r,p)\big(\log q(r,p) - \log q(r) - \log q(p)\big)}\, dr\, dp$$

Because q(c|p) and q(p) are not computable and the Kullback–Leibler scatter is positive, the above equation can be approximated using a variational approximation as follows:

$$\mathrm{KL}[\mathrm{q(c|p)}, \mathrm{t(c|p)}] \geq 0$$

$$\int q(c|p) \log q(c|p)\, dc \geq \int q(c|p)\, \log t(c|p)\, dc$$

Thus, M(C,P) can be approximated as follows:

$$\begin{aligned} \mathrm{M(C,P)} &\geq \int \mathrm{q(c,\ p)}\big(\log \mathrm{(c|p)} - \log\ \mathrm{q(c)}\big)\, dc\, dp \\ &= \int \mathrm{q(c,p)}\ \log\ \mathrm{t(c|p)}\, dc\, dp - \int \mathrm{q(c)}\ \log\ \mathrm{q(c)}\, dc \\ &= \int \mathrm{q(c,\ p)}\ \log\ \mathrm{t(c|p)}\, dc\, dp + \mathrm{G(C)} \end{aligned}$$

where G(C) is irrelevant to the optimization objective and can be ignored. Furthermore, M(C,P) can be expressed using the Markov chain and edge probability density formula as follows:

$$\mathrm{M(C,P)} \geq \int \mathrm{q(r)q(c|r)q(p|r)} \log \mathrm{t(c|p)} dr\, dc\, dp$$

Similarly, M(R,P) can be expressed as follows:

$$\mathrm{M(R,P)} \leq \int \mathrm{q(r)q(p|r)}\big(\log \mathrm{q(p|r)} - \mathrm{s(p)}\big) dr\, dp$$

Combining the boundaries of M(C,P) and M(R,P) can be expressed as follows:

$$M(C,P) - \lambda M(R,P) \geq L$$

L is the lower bound, and the optimization objective can be approximated as follows:

$$\mathrm{L} \approx \frac{1}{\mathrm{K}} \sum_{\mathrm{i}=1}^{\mathrm{K}} \left[dz\ \mathrm{q}(p|r_i) \log\ \mathrm{t(c_i|p)} - \lambda q(\mathrm{p|r_i}) \log \frac{\mathrm{q(p|r_i)}}{\mathrm{s(p)}} \right]$$

We estimated the gradient of the lower bound L using the reparameterization technique and calculated it as follows:

$$z = E(r) + \varepsilon D(r)$$

where E(r) is the mean of r, D(r) is the variance of r, and is an auxiliary noise variable and follows a standard normal distribution.

Because the VFC is random in terms of weight initialization and sampling, the feature set processed by the algorithm exhibits a certain degree of randomness. Therefore, we simultaneously generated multiple features with the same dimensions. After several rounds of validation, the feature set with the highest MCC value and the smallest possible number of feature dimensions was the final feature set. Finally, we obtained a final feature set containing 15 features.

2.4 Classification method

Since Chen first proposed the XGBoost algorithm in 2017 [50], it has received considerable attention as an integrated learning algorithm based on gradient enhancement [51–54]. XGBoost accumulates the predictions of the k-tree by weighting at each iteration step and uses the information of the first-order derivative from the loss function to adjust the model and optimize the final prediction of the model using the following formula:

$$\phi(\mathrm{x_i}) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{f_k(x_i)} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \omega_{\mathrm{q_k(x_i)}}$$

where $q_k(x_i)$ is the prediction model for the kth tree, w is the leaf weight, x_i is the characteristics of the sample, and y is the algorithm's prediction result. Given an input, it will be determined the leaf node to which it belongs is based on the branching condition (denoted as q) of the tree structure species. The prediction is obtained by accumulating the scores (denoted as w) of all the leaf nodes through which this input passes. For the algorithm to learn the sample information better, the objective function can be expressed as follows:

$$\mathrm{obj}(\theta) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \ell\big(\mathrm{y_i}, \widehat{\mathrm{y}}_1\big) + \sum_{\mathrm{k}=1}^{\mathrm{K}} \Omega(\mathrm{t_k}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \ell\big(\mathrm{y_i}, \widehat{\mathrm{y}}_1\big) + \gamma T + \frac{1}{2}\lambda \parallel \omega \parallel^2$$

where ℓ is used to measure the difference between the algorithm's prediction and the actual labels. Ω is the regularization term, consisting of the leaf node count and leaf weights. The parameter in the regularization term controls leaf number, and the parameter controls leaf weights.

2.5 Evaluation metrics

The k-fold cross-validation assessment model was previously found to yield more objective assessment results [55–59] and was used in this study as well. This method splits the dataset equally into k-folds, selecting onefold at a time as the validation set and the rest of the folds as the training set, and repeating this k times. The averages of

the evaluations were used as the cross-validation results. To demonstrate that the model performs better on both positive and negative samples, we used the following relevant formulas as evaluation metrics:

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

$$SN = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

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

$$F1_{score} = \frac{2TP}{2TP + FP + FN}$$

where TP represents the amount of data categorizing antioxidant proteins as antioxidant proteins, FP represents the amount of data categorizing nonantioxidant proteins as antioxidant proteins, TN represents the amount of data categorizing nonantioxidant proteins as nonantioxidant proteins, and FN represents the amount of data categorizing antioxidant proteins as nonantioxidant proteins. The Matthews correlation coefficient (MCC) represents the ability of the model to balance the predictive accuracy of positive and negative samples [48].

3 Results and discussion

3.1 Algorithm overview

To assess the performance of the dimensionality reduction algorithm, we first plotted sample distributions before and after dimensionality reduction. Owing to the clear description of feature contributions in DP-AOPs, we refer to the table provided and visualize these features for comparison. For a fair assessment, we chose the same number of features as the number of feature dimensions we adopted, i.e., the first 15 features, based on the contribution of the features in the DP-AOPs table from highest to lowest. Data points before dimensionality reduction were found to be highly interconnected (Fig. 2a), with the red and blue data points appearing particularly overlapping each other in some areas. In addition, the decision boundary was found to be complex and irregular, hindering the capture of useful information. In Fig. 2b, the classification boundaries between the red and blue data points are relatively more clear, with several regions densely populated with a single category, yet some regions appeared to contain both types. Upon dimensionality reduction, the separation between the red and blue data points in some regions was improved compared to the original features and DP-AOPs. In addition, the decision boundary line separated the two types of data points. In summary, we conclude that the low-dimensional features obtained by the VFC dimensionality reduction method enable a more generalized and efficient classification.

Although the main idea behind VFC feature dimensionality reduction is to transform and combine the original features nonlinearly in the latent space, the method still has a preference for certain features. Thus, to some extent, certain features and their potential relationships are more important to identify antioxidant proteins. In addition, the source of the features after dimensionality reduction may reveal their biological significance. We determined the importance of the original features and the relationship between the features before and after dimensionality reduction using the SHAP method. First, the SHAP values between the original feature vector and 15-dimensional feature vector were calculated for all samples. The SHAP values of all samples were then

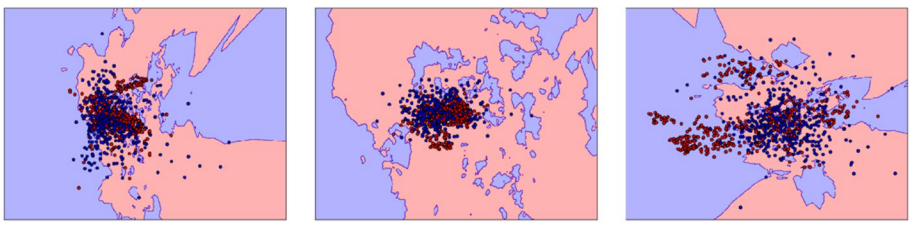

Fig. 2 Sample distribution visualization of the feature set obtained based on different algorithms. **a** Sample distribution visualization of the feature set using the original 473D feature set, **b** sample distribution visualization of the 17-dimensional enlistment used in DP-AOP, and **c** sample distribution visualization of the 15-dimensional enlistment obtained based on the dimensionality reduction of the VFC features. Red dots represent positive samples, and blue dots represent negative samples in the figure

averaged to obtain absolute values. To obtain the main components of the dimensionality reduction feature, we sorted the SHAP values of the 473-dimensional features corresponding to each dimensionality reduction feature in descending order, and thus obtained the features in descending order of importance. Considering the large number of features, we chose features with SHAP values greater than 0.012 as references.

The relationship between the original feature vector obtained based on the SHAP value and the dimensionality-reduction feature vector is shown in Fig. 3. The original features on the left side of the graph are listed in decreasing order of SHAP values, and the features on the right side are listed in order of feature name. Ten features were based on the structure probability matrix, and 26 features were based on the consensus sequence, as shown in the figure. Additionally, the five features with the largest SHAP values were based on a consensus sequence. Here, consensus sequences were found to play an essential role in predicting antioxidant proteins.

3.2 Performance evaluation based on k-fold cross-validation

To estimate the model performance more reliably, the k-fold cross-validation method was used to estimate model performance. Specifically, the antioxidant dataset was split into five equally sized subsets, and a different subset was selected as test data, whereas the remaining four were used for model training each time model performance was evaluated. This process was repeated five times. This cross-validation process reduces accidental errors by repeating it many times for different

Table 1 Performance results and average performance results of the fivefold cross-validation method based on the benchmark dataset

Times	ACC	SN	SP	MCC	F1
1	96.62	96.44	96.80	93.24	96.60
2	96.78	96.30	97.22	93.55	96.62
3	95.49	95.25	95.74	90.98	95.56
4	95.65	93.31	98.29	91.44	95.79
5	94.84	95.68	94.04	89.69	94.74
Average	95.88	95.40	96.42	91.78	95.86

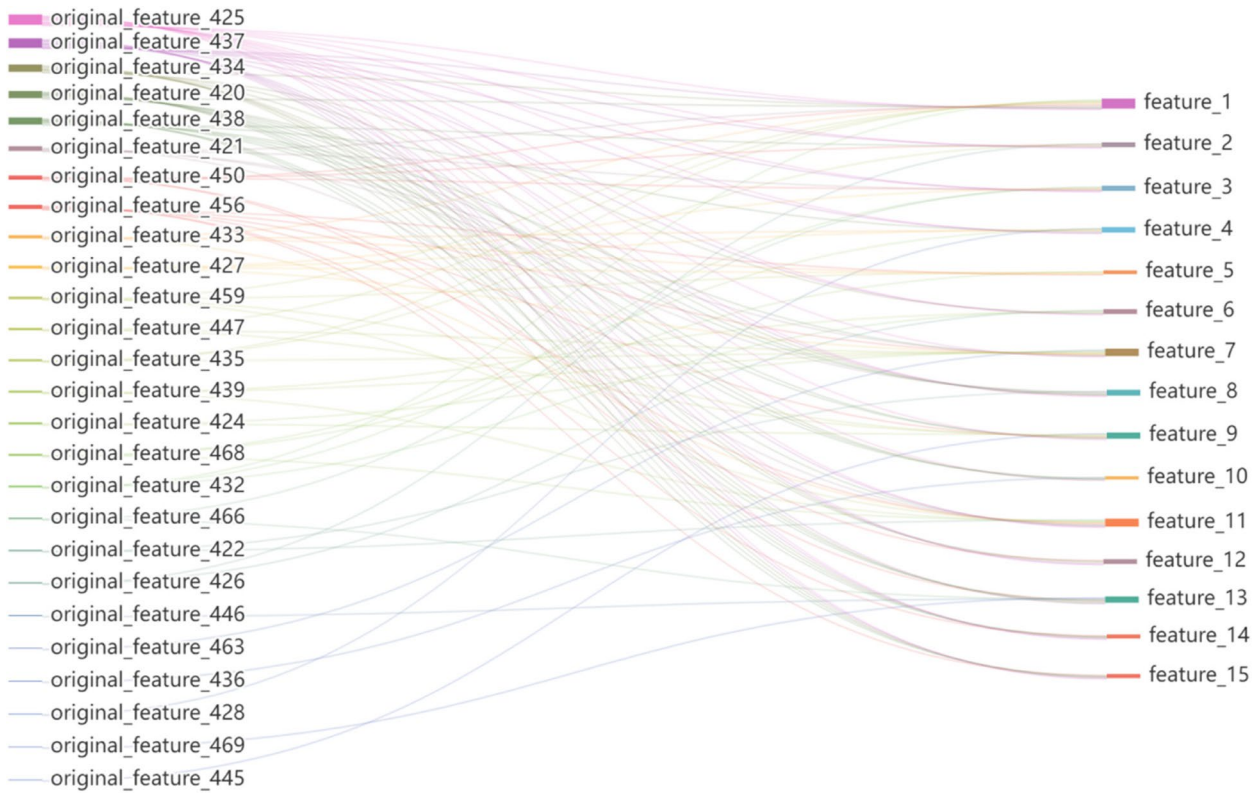

Fig. 3 Visualization of the relationship between the original and dimensionality reduction feature vectors obtained based on SHAP values. The original features on the left side of the figure are arranged in descending order based on SHAP values, and the features on the right side are arranged in the order of feature names

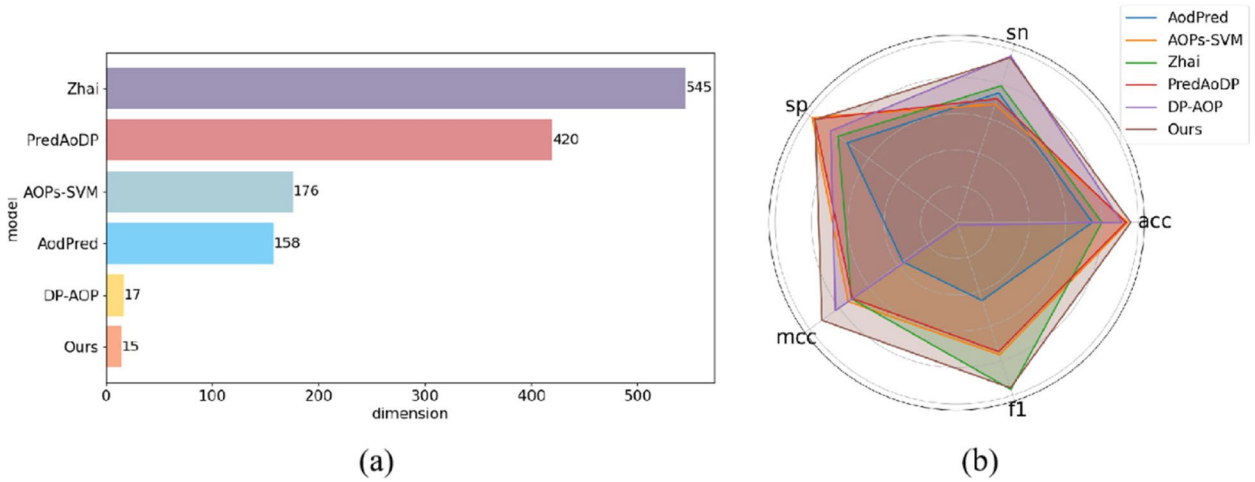

Fig. 4 Comparison of Rore's performance with other models. **a** Comparison based on the dimension of the feature vectors used. **b** Comparison based on SN, SP, MCC, ACC, and F1 evaluation metrics

Table 2 Performance comparison of Rore with other classifiers

Model	ACC	SN	SP	MCC	F1	Dimension
AodPred	74.79	75.09	74.48	36.8	45.2	158
Zhai	80	79.2	80.8	71.65	**96.77**	545
DP-AOP	91.08	**96.4**	85.8	82.6	91.5	17
PredAoDP	93.18	71.65	96.77	71.2	74.9	420
AOPs-SVM	94.2	68.5	**98.5**	74.1	76.7	176
Rore	**95.88**	95.4	96.42	**91.78**	95.86	**15**

combinations of subsets, ultimately yielding more objective model evaluation results. For performance evaluation, a benchmark dataset was constructed using the dataset developed by Feng et al. (2016). Table 1 presents the performance results of the developed classifier on this dataset for five instances after fivefold cross-validation, as well as the average performance of the five instances. The results of the experiments objectively demonstrated the accuracy and generalization ability of the model, as well as its ability to classify antioxidant proteins.

3.3 Comparison of Rore classifier performance with those of other classifiers

We adopted the benchmark dataset used in the k-fold cross-validation method and tested the performances of the Rore model and the other five methods. As shown in Fig. 2, Fig. 4, and Table 2, Rore model outperforms all the other methods owing largely to retaining of the original information and discarding of noise. In clinical applications, the advantage in terms of the MCC values is significant, indicating that the classifier performs well on unbalanced datasets. In practice, more importance is given to MCC and ACC values. Therefore, we sacrificed SN, SP, and F1 scores within acceptable limits.

4 Conclusion

To overcome the limitations of feature selection in existing protein identification methods, a classifier based on feature dimensionality reduction was proposed in this study. By mapping the original features into the potential space to obtain a compressed representation with all feature information, the problem of information loss caused by the previous use of feature selection is solved. Experimental results show that on the benchmark dataset, our classifier model outperforms other existing classifiers in three metrics, MCC, F1, and ACC, using only 15-dimensional features. A high MCC value indicates superiority when dealing with unbalanced datasets. It shows great potential in clinical applications. Therefore, we can conclude that the Rore classifier can obtain more robust features to achieve superior recognition ability. Given the excellent robustness of the Rore classifier, its application potential in other recognition areas may be investigated in future studies.

Acknowledgements
Not applicable

Authors' contributions
C.M. and Y.H wrote the main manuscript text, Q.Z. and L.S. prepared Figs. 1, 2, 3 and 4. All authors reviewed the manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (62262051, 62373203, 62271489, 62072385), the Natural Science Foundation of Inner Mongolia (2022MS06030), Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT23056), and young innovative talents of the "Grassland Talents" project of Inner Mongolia Autonomous Region (to Chaolu Meng).

Data availability
Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 27 July 2024 Accepted: 3 October 2024
Published online: 04 December 2024

References

1. Cheeseman KH, Slater TF. An introduction to free radical biochemistry. Br Med Bull. 1993;49(3):481–93.
2. Phaniendra A, Jestadi DB, Periyasamy L. Free radicals: properties, sources, targets, and their implication in various diseases. Indian J Clin Biochem. 2015;30:11–26.
3. DiMartini ET, Lowe CJ, Shreiber DI. Alternative chemistries for free radical-initiated targeting and immobilization. J Functional Biomater. 2023;14(3):153.
4. Li H, Liu B. BioSeq-Diabolo: biological sequence similarity analysis using Diabolo. PLoS Comput Biol. 2023;19(6):e1011214.
5. Massonis G, Villaverde AF, Banga JR. Distilling identifiable and interpretable dynamic models from biological data. PLoS Comput Biol. 2023;19(10):e1011014.
6. Tonner PD, Pressman A, Ross D. Interpretable modeling of genotype–phenotype landscapes with state-of-the-art predictive power. Proc Natl Acad Sci. 2022;119(26): e2114021119.
7. Poursabzi-Sangdeh, F., et al. Manipulating and measuring model interpretability. in Proceedings of the 2021 CHI conference on human factors in computing systems. 2021.
8. Wang, Y., Zhai, Y., Ding, Y., Zou, Q, *SBSM-Pro: support bio-sequence machine for proteins.* arXiv preprint, 2023: p. arXiv:2308.10275.
9. Guo X, et al. Highly accurate estimation of cell type abundance in bulk tissues based on single-cell reference and domain adaptive matching. Adv Sci. 2024;11(7):2306329.
10. Ai C, et al. MTMol-GPT: De novo multi-target molecular generation with transformer-based generative adversarial imitation learning. Plos Comput Biol. 2024;20(6):e1012229.
11. Jiang Y, et al. Explainable deep hypergraph learning modeling the peptide secondary structure prediction. Adv Sci. 2023;10(11):2206151.
12. Wei L, et al. Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. Briefings in Bioinformatics. 2020.
13. Li H, Pang Y, Liu B. BioSeq-BLM: a platform for analyzing DNA, RNA, and protein sequences based on biological language models. Nucleic Acids Res. 2021;49(22): e129.
14. Cao C, et al. webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. Nucleic Acids Res. 2022;50(D1):D1123–30.
15. Dao F-Y, et al. AcrPred: a hybrid optimization with enumerated machine learning algorithm to predict anti-CRISPR proteins. Int J Biol Macromol. 2023;228:706–14.
16. Yang S, et al. MASQC: next generation sequencing assists third generation sequencing for quality control in N6-methyladenine DNA identification. Front Genet. 2020;11: 507302.
17. Jin J, et al. Rapid screening of multi-point mutations for enzyme thermostability modification tools. Future Gen Comput Syst Int J Esc. 2024;160:160.
18. Feng P, Chen W, Lin H. Identifying antioxidant proteins by using optimal dipeptide compositions. Interdiscipl Sci. 2016;8:186–91.
19. Ao C, et al. Prediction of antioxidant proteins using hybrid feature representation method and random forest. Genomics. 2020;112(6):4666–74.
20. Meng C, et al. DP-AOP: A novel SVM-based antioxidant proteins identifier. Int J Biol Macromol. 2023;247: 125499.
21. Basit MS, Khan A, Farooq O, et al. Handling imbalanced and overlapped medical datasets: a comparative study[C]//2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT). IEEE; 2022. p. 1–7.
22. Moore RC, Ellis DP, Fonseca E, Hershey S, Jansen A, Plakal M. Dataset balancing can hurt model performance. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE; 2023. p. 1–5.
23. Su R, et al. Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. Brief Bioinform. 2020;21(2):408–20.
24. Tang Y, Pang Y, Liu B. IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning. Bioinformatics. 2021;36(21):5177–86.
25. Newman-Toker DE, et al. Burden of serious harms from diagnostic error in the USA. BMJ Qual Saf. 2024;33(2):109–20.
26. Newman-Toker DE, et al. Serious misdiagnosis-related harms in malpractice claims: the "Big Three"–vascular events, infections, and cancers. Diagnosis. 2019;6(3):227–40.
27. Newman-Toker DE, et al. Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the "Big Three." Diagnosis. 2021;8(1):67–84.
28. Ma K, et al. PPRTGI: a Personalized PageRank Graph Neural Network for TF-Target Gene Interaction Detection. IEEE/ACM Trans Comput Biol Bioinf. 2024;21(3):480–91.
29. Su R, et al. Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response. Methods. 2019;166:91–102.
30. Meng C, et al. AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. Frontiers in Bioengineering and Biotechnology. 2019;7: 224.
31. Ahmed S, et al. PredAoDP: accurate identification of antioxidant proteins by fusing different descriptors based on evolutionary information with support vector machine. Chemom Intell Lab Syst. 2022;228:104623.
32. Consortium, U. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2012;40(D1):D71–5.
33. Wu S, et al. Machine learning aided construction of the quorum sensing communication network for human gut microbiota. Nat Commun. 2022;13(1):3079.
34. Chawla NV, et al. SMOTE: synthetic minority over-sampling technique. J Artificial Intellig Res. 2002;16:321–57.
35. Yang Y, et al. DeepIDC: a prediction framework of injectable drug combination based on heterogeneous information and deep learning. Clin Pharmacokinet. 2022;61(12):1749–59.
36. Wei L, et al. Enhanced protein fold prediction method through a novel feature extraction technique. IEEE Trans Nanobiosci. 2015;14(6):649–59.
37. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
38. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999;292(2):195–202.
39. Kong R, et al. 2SigFinder: the combined use of small-scale and large-scale statistical testing for genomic island detection from a single genome. BMC Bioinformatics. 2020;21:1–15.

40. Dai Q, et al. A segmentation based model for subcellular location prediction of apoptosis protein. Chemom Intell Lab Syst. 2016;158:146–54.
41. Qian Y, et al. Identification of DNA-binding proteins via hypergraph based Laplacian support vector machine. Curr Bioinform. 2022;17(1):108–17.
42. Wei L, et al. ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. Brief Bioinform. 2021;22(5):bbab041.
43. Sussman JL, et al. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr. 1998;54(6):1078–84.
44. Zhang H, et al. Distance-based support vector machine to predict DNA N6-methyladenine modification. Curr Bioinform. 2022;17(5):473–82.
45. Jin J, et al. iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. Genome Biol. 2022;23(1):219.
46. Wei L, et al. ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning. Bioinformatics. 2022;38(6):1514–24.
47. Alemi, A.A., et al., *Deep variational information bottleneck.* arXiv preprint arXiv:1612.00410, 2016.
48. Zhu, W., et al., A first computational frame for recognizing heparin-binding protein. Diagnostics (Basel). 2023;13(14).
49. Gu T, Xu G, Luo J. Sentiment analysis via deep multichannel neural networks with variational information bottleneck. IEEE Access. 2020;8:121014–21.
50. Chen T, C Guestrin. Xgboost: A scalable tree boosting system. in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.
51. Yu B, et al. SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. Bioinformatics. 2020;36(4):1074–81.
52. Zhang D, et al. iBLP: An XGBoost-based predictor for identifying bioluminescent proteins. Comput Math Methods Med. 2021;2021(1):6664362.
53. Abbas Z, et al. XGBoost framework with feature selection for the prediction of RNA N5-methylcytosine sites. Mol Ther. 2023;31(8):2543–51.
54. Zhu H, Hao H, Yu L. Identifying disease-related microbes based on multi-scale variational graph autoencoder embedding Wasserstein distance. BMC Biol. 2023;21(1):294.
55. Ogundunmade T, Adepoju A, Allam A. Stock price forecasting: machine learning models with K-fold and repeated cross validation approaches. Mod Econ Manag. 2022;1(1):2.
56. Oyedele O. Determining the optimal number of folds to use in a K-fold cross-validation: a neural network classification experiment. Res Mathematics. 2023;10(1): 2201015.
57. Phinzi K, Abriha D, Szabó S. Classification efficacy using k-fold cross-validation and bootstrapping resampling techniques on the example of mapping complex gully systems. Remote Sensing. 2021;13(15):2980.
58. Zulfiqar H, et al. Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings. Front Med. 2024;10:10.
59. Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Res. 2019;47(20):e127.

Publisher's Note