Research on protein solubility prediction based on ensemble learning and feature fusion

Hongqi Feng[1,a], Tao Wu[1,b,*]

[1]School of Computer Science and Artificial Intelligence, Aliyun School of Big Data, School of Software, Changzhou University, Changzhou 213164, China
[a]hqfeng@cczu.edu.cn, [b]952060862@qq.com
**Corresponding author*

Abstract: *Protein solubility is one of the momentous properties of a protein that can effectively participate in and inhibit the physiological and biochemical processes of cancer cells in the human body. Therefore, understanding the solubility of proteins may be significant to find the mechanism of diseases caused by the solubility of proteins. In this paper, to improve the protein solubility prediction performance and address the inadequacy of existing protein solubility prediction methods that more feature information about protein sequences is difficult to be obtained. A protein solubility prediction model named EL-FFsol is proposed, which is based on the CatBoost ensemble learning framework and multiple feature fusion of protein sequences. First of all, protein sequence features were introduced to build fusion representation, including the Physicochemical Properties, One-hot Feature Encoding, Amino Acid Composition and Statistical Features. Additionally, the CatBoost was employed to construct an ensemble learning model to predict protein solubility. Finally, EL-FFsol was tested on the benchmark dataset to predict the solubility of proteins. In terms of accuracy, matthews correlation coefficient, sensitivity, specificity, area under ROC curve and area under P-R curve, EL-FFsol achieved 0.7679, 0.5480, 0.6630, 0.8729, 0.8540 and 0.8440 performances. Compared with the DeepSOL and DDcCNN, the matthews correlation coefficient was increased by 1.68% and 0.79%, the area under ROC curve was increased by 1.60% and 2.20% and the area under P-R curve was increased by 1.70% and 2.40%, respectively.*

Keywords: *protein solubility; sequence information; multiple feature fusion; ensemble learning*

1. Introduction

As the material basis of life, proteins play important roles in cell activities[1–3]. The protein function is decided by molecular structures and inherent features. Protein solubility is one of the vital properties of a protein, which is significant to human health. For example, the decrease of protein solubility may lead to the formation of insoluble aggregates. These aggregates may give rise to varieties of neurodegenerative diseases[4], such as Alzheimer's Disease, Amyotrophic Lateral Sclerosis and Parkinson's Disease, which have a serious effect on human cognitive functions and behavior[5–7]. In addition, the sickness rate of cataract may be increased markedly along with the added number of insoluble proteins in the lens[8]. The disease not only has a long-term influence on the normal life of the elderly but also brings about a series of complications[9]. Additionally, several soluble immune checkpoints[10] are based on soluble proteins such as soluble CTLA-4 and soluble PD-1. They may diffuse in serums and regulate the immune system in a positive or negative direction. Thus, they are crucial to the examination, prognosis and treatment of cancers[11]. Therefore, predicting the solubility of proteins is essential and necessary[12].

At present, for the detection of protein solubility, there are mainly laboratory wet experimental methods and computer methods.

Protein solubility refers to the mass of proteins dissolved in a certain amount of potassium hydroxide solution[13]. For example, to detect the rawness of soybeans by the solubility of proteins, the following procedures can be carried out. Firstly, soybeans are heated to different degrees. Secondly, these soybeans are soaked in the 0.2% potassium hydroxide solution. Finally, the ratio is used to compute the rawness of soybeans by using the Kjeldahl method[14]. The ratio means the protein content in soybeans after potassium hydroxide dissolving to the protein content in the original sample. From this, it can be seen that the laboratory wet experimental method is a waste of time, costly, highly repeatable and cannot satisfy the demands of high-throughput. While protein solubility is associated with amino acid residues that make up these proteins[15]. By studying repetitive regulars between soluble and insoluble proteins in the Escherichia coli expression system[16], protein solubility can be calculated. Thus, computer methods may be adopted to assist laboratory wet experimental methods.

In the environment of high-throughput[17] and big data[18], the methods of machine learning or deep learning may be used to predict protein solubility[19]. Several machine learning methods are designed to predict protein solubility. For example, SOLpro proposed by Magnan *et al*.[20] in 2009 obtains the k-mer features of protein sequences and utilizes the sequential minimal optimization (SMO) to train a SVM model to predict protein solubility. PROSO II proposed by Smialowski *et al*.[21] in 2012 extracts the k-mer features from protein sequences and makes use of a new classifier to predict protein solubility. CCSOL proposed by Agostini *et al*.[22] in 2014 uses the hydrophobicity and helicity of protein sequences as the main features and combines a SVM model to predict protein solubility. PaRSnIP proposed by Rawi *et al.*[23] in 2018 reveals that a high proportion of exposed amino acid residues are positively correlated with protein solubility and tripeptide combinations composed of multiple amino acid residues are negatively correlated with protein solubility. And then protein solubility may be predicted by using the k-mer features and Statistical Features of protein sequences through the gradient boosting machine (GBM) model. Most of these methods mentioned above are depended on the SVM, which are unsuitable for big data, while deep learning can acquire more nonlinear relationships[24] and its processing ability is more powerful. For instance, DeepSOL proposed by Khurana *et al.*[25] in 2018 adopts the convolutional neural network (CNN) for the first time, together with the One-hot Feature Encoding and Statistical Features to generate a 25,257-dimensional feature vector to predict protein solubility. While the model with the large input dimension may result in the consumption of a great deal of time and numerous computing resources. SCNN, DCNN, DTcCNN and DDcCNN proposed by Wang *et al.*[26] in 2021 apply the convolutional neural network (CNN) and combine the G-Gap Structural Features and Statistical Features to predict protein solubility. While certain deficiencies exist, first of all, numerous convolutional layers lead to an increased number of parameters and some shortages in performance and efficiency. When the amount of data is insufficient, overfitting may easily come to pass. Besides that, few convolutional layers create inadequate training and performance limitations. Furthermore, several operations in pooling layers cause the lack of certain valuable information, and then the relevance between the whole and the parts may be attenuated.

Under given experimental conditions, the EL-FFsol model was designed to correctly predict protein solubility, which was founded on multiple feature information and an ensemble learning framework. Firstly, the Cluster Database at High Identity with Tolerance (CD-HIT) was utilized to denoise protein sequences[27]. And then various features were extracted from protein sequences, including Hydropathy Index, Electron-Ion Interaction Potential, Molecular Weight, Residue Molecular Weight, Charge and Polarity, Acid-base Features, One-hot Feature Encoding, Amino Acid Composition, Sequence Features[20], Structural Features[25,26,28] and Relative Solvent Accessibility[25,26,28]. Additionally, feature fusion was constructed on these features and a brand-new feature vector was formed. Finally, protein solubility was predicted on the benchmark dataset through the model based on the CatBoost ensemble learning framework.

Briefly, to attain more protein sequence information and enhance the performance and generalization, EL-FFsol extracted the Physicochemical Properties, One-hot Feature Encoding, Amino Acid Composition and Statistical Features from protein sequences. And then feature fusion was built on these features to generate an all-new feature vector. To cope with the huge amount of data and mitigate the inadequacy of existing protein solubility prediction methods, EL-FFsol employed the CatBoost algorithm to substitute for the convolutional neural network to process the vector after feature fusion.

2. Method for protein solubility prediction

This chapter introduces the method used in the protein solubility prediction model. Figure 1 sketches the development flow chart of the EL-FFsol model, which contains three parts: (a) dataset preprocessing of protein sequences, (b) feature extraction of protein sequences and (c) construction and training of the protein solubility prediction model.

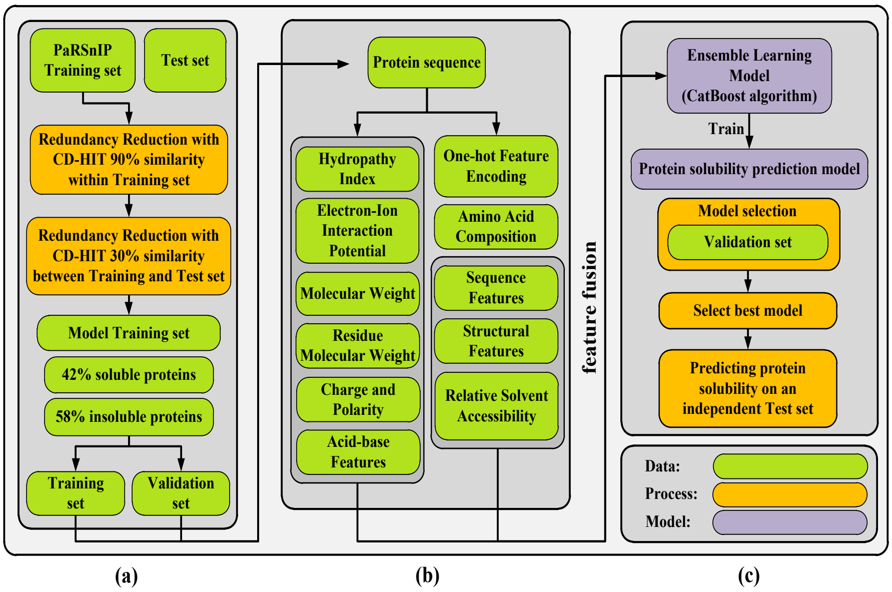


Figure 1: EL-FFsol development flow chart: (a) dataset preprocessing of protein sequences; (b) feature extraction of protein sequences; (c) construction and training of the protein solubility prediction model

2.1. Dataset preprocessing

The protein sequence dataset applied comes from the research[23]. The training set consists of 70,954 soluble proteins and 58,689 insoluble proteins; while the independent test set contains 1,000 soluble proteins and 999 insoluble proteins[28]. To ensure the mutual independence of data in the training set, the CD-HIT was utilized to denoise protein sequences[27]. The dataset preprocessing process is shown in Figure 1(a).

Sequence identity indicates the percentage that the number of identical residues in two sequences to the total length of the sequence[29]. It measures the degree of similarity between protein sequences. Firstly, in the training set, the sequence identity was set to 90%. Secondly, all sequences were deleted, which have over 30% sequence identity between the training set and independent test set. Eventually, the processed dataset was divided into two mutually exclusive subsets of different sizes by using the hold-out method, which was considered as the training set and validation set. The final training set includes 26,075 soluble proteins and 36,403 insoluble proteins, a total of 62,478 protein sequences; while the validation set comprises 2,921 soluble proteins and 4,020 insoluble proteins, a total of 6,941 protein sequences.

2.2. Feature extraction

Based on the preprocessed dataset of protein sequences, the Physicochemical Properties, One-hot Feature Encoding, Amino Acid Composition and Statistical Features were extracted. And then feature fusion was constructed to form a new feature vector. The feature extraction process is depicted in Figure 1(b).

2.2.1. Physicochemical Properties of protein sequences

The Physicochemical Properties used in EL-FFsol contained Hydropathy Index (HI), Electron-Ion Interaction Potential (EIIP), Molecular Weight (MW), Residue Molecular Weight (RMW), Charge and Polarity (CP) and Acid-base Features (AF). These features describe the physical properties and chemical properties of amino acid residues that constitute these proteins. Thus, more protein sequence information is provided and the problem that feature information is difficult to be extracted due to the complex structure of proteins may be effectively addressed. The above features are expressed as $\boldsymbol{h}_x^{(k)} = (a_1, a_2, ..., a_i, ..., a_L)$, where $x(1 \le x \le 6)$ denotes the above six different features, k means the k-th protein sequence in the dataset, L shows the protein sequence length and $a_i (1 \le i \le L)$ represents the feature value or binary vector of the Physicochemical Properties corresponding to each amino acid residue. The values of the Physicochemical Properties are shown in Table 1.

Table 1: Values of the Physicochemical Properties of each amino acid residue

Amino acid residue	HI	EIIP	MW	RMW	CP	AF
A	1.8	0.0373	89.09	71.07	[0,0,0,1]	[0,0,1]
C	2.5	0.0829	121.16	103.10	[0,0,1,0]	[0,0,1]
D	-3.5	0.1263	133.10	115.08	[0,1,0,0]	[0,1,0]
E	-3.5	0.0058	147.13	129.11	[0,1,0,0]	[0,1,0]
F	2.8	0.0946	165.19	147.17	[0,0,0,1]	[0,0,1]
G	-0.4	0.005	75.07	57.05	[0,0,1,0]	[0,0,1]
H	-3.2	0.0242	155.16	137.14	[0,1,0,0]	[1,0,0]
I	4.5	0	131.17	113.15	[0,0,0,1]	[0,0,1]
K	-3.9	0.0371	146.19	128.17	[0,1,0,0]	[1,0,0]
L	3.8	0	131.17	113.15	[0,0,0,1]	[0,0,1]
M	1.9	0.0823	149.21	131.19	[0,0,0,1]	[0,0,1]
N	-3.5	0.0036	132.12	114.10	[0,0,1,0]	[0,0,1]
P	-1.6	0.0198	115.13	97.11	[0,0,0,1]	[0,0,1]
Q	-3.5	0.0761	146.15	128.13	[0,0,1,0]	[0,0,1]
R	-4.5	0.0959	174.20	156.18	[0,1,0,0]	[1,0,0]
S	-0.8	0.0829	105.09	87.07	[0,0,1,0]	[0,0,1]
T	-0.7	0.0941	119.16	101.14	[0,0,1,0]	[0,0,1]
V	4.2	0.0057	117.15	99.13	[0,0,0,1]	[0,0,1]
W	-0.9	0.0548	204.22	186.20	[0,0,0,1]	[0,0,1]
Y	-1.3	0.0516	181.19	163.17	[0,0,1,0]	[0,0,1]

2.2.2. One-hot Feature Encoding of protein sequences

Amino acid residues of protein sequences were encoded by One-Hot Encoding, and then a feature vector $\boldsymbol{h}_7$ of length $L*20$ was obtained, where each amino acid residue is shown by a binary vector of length 20 and L denotes the protein sequence length. Since each protein is composed of twenty classes of amino acid residues and they have hydrophilic or hydrophobic because of the characteristic difference of side chains, amino acid residues have a certain influence on protein solubility. Thus, data information about amino acid residues can be detailed by using the One-hot Feature Encoding.

2.2.3. Amino Acid Composition of protein sequences

By arranging each protein sequence sequentially according to the initial sequence, the combination sequence of every two amino acid residues and the combination sequence of every three amino acid residues, the unary combination sequence, binary combination sequence and ternary combination sequence were acquired.

The Amino Acid Composition of the unary combination sequence may be expressed as $\boldsymbol{A}_1^{(k)}=(a_1,a_2,...,a_i,...,a_{19},a_{20})$, where k means the k-th protein sequence in the dataset and $a_i(1\le i\le 20)$ represents the frequency of amino acid residues.

The Amino Acid Composition of the binary combination sequence and ternary combination sequence may be expressed as $\boldsymbol{A}_x^{(k)}=(a_1,a_2,...,a_i,...,a_L)$, where $x=2$ denotes the binary combination sequence, $x=3$ represents the ternary combination sequence, k shows the k-th protein sequence in the dataset, L means the protein sequence length and $a_i(1\le i\le L)$ indicates the feature value about the histogram information of each amino acid residue, which is calculated through the use of the histogram function in the NumPy library.

By combining the Amino Acid Composition of the above three combination sequences, the feature vector $\boldsymbol{h}_8$ was gained. The Amino Acid Composition describes the combination information of amino acid residues and these amino acid residues have an impact on protein solubility. Thus, the Amino Acid Composition helps predict the solubility of proteins.

2.2.4. Statistical Features of protein sequences

The Statistical Features were used as a feature vector $\boldsymbol{h}_9$, which contained three categories: Sequence Features, Structural Features and Relative Solvent Accessibility. The Sequence Features were calculated by formulas referred to in the Biopython library and Propy3 library, including Sequence Length, Molecular Weight (MW), Fraction Turn-forming Residues (FTR), Aliphatic Indices (AI), Average

Hydropathicity (AH) and Absolute Charge (AC), which have a total of 6 dimensions. All these features narrate the protein sequence in the aspect of physical properties and chemical properties.

The Structural Features and Relative Solvent Accessibility were calculated by the SCRATCH[30], a bioinformatics tool, which have a total of 51 dimensions, including 3-dimensional Secondary Structural Features (the number of categories of amino acid residue combinations was 3), 8-dimensional Secondary Structural Features (the quantity of categories of amino acid residue combinations was 8), 20-dimensional RSA Features (obtained by using the cutoff value of the Relative Solvent Accessibility that ranged from 0% to 95% with an interval of 5%) and 20-dimensional RSA-AH Features (acquired by the RSA Features multiplying by the Average Hydrophobicity of exposed residues). Although compositions and structures of proteins are complex and diverse, the Structure Features not only sketch out the arrangement of polypeptide chains in proteins but also provide more feature information about protein formations and frameworks; the Relative Solvent Accessibility refers to the surface area of the solvent in contact with biomolecules, which can intuitively describe the protein solubility in an aqueous solution.

After extracting the feature information from protein sequences, feature fusion was implemented to form a fresh feature vector. Figure 2 shows the process of feature extraction and fusion.

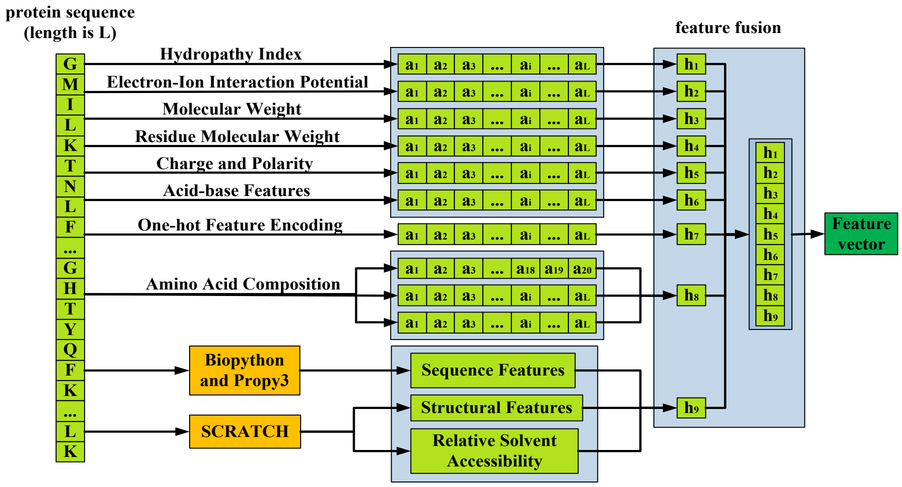


Figure 2: Extraction and fusion of protein sequence features

2.3. Construction and training of the model

The CatBoost algorithm was employed to build an ensemble learning model to predict protein solubility. Firstly, the best model was trained and selected on the training set and validation set. Additionally, the best model was tested on the benchmark to verify the prediction performance. The construction and training process of the protein solubility prediction model is pictured in Figure 1(c).

2.3.1. CatBoost algorithm

In this paper, as the basic classifier for protein solubility prediction, the CatBoost algorithm was used to calculate the vector after feature fusion. In particular, the CatBoost is a novel algorithm that combines gradient boosting and categorical knowledge. Same as all other gradient-based[31] methods, it has two processes: the first process is to select the structure of the tree; the second process is to assign values to the leaf nodes of the fixed tree. Additionally, the CatBoost algorithm has two advantages: firstly, the categorical knowledge is directly trained in the model without manual processing; secondly, the CatBoost algorithm utilizes the Ordered Boosting method[32] to change the gradient estimate from biased to unbiased, which not only slows down the gradient bias but also effectively controls overfitting and improves the model generalization.

2.3.2. Training of the model

The logarithmic loss function was utilized in EL-FFsol, which can be expressed in formula (1).

$$\text{Loss}=-\log\text{P}(Y|X)=-\frac{1}{N}\sum_{i-1}^{N}\sum_{j=1}^{M}y_{ij}\log(p_{ij}) \quad (1)$$

Where Y denotes the outputs, X denotes the inputs, N denotes the count of input samples, M denotes the count of possible categories, y_{ij} denotes a binary index whether the category *j* is the true description of the input x_i, and p_{ij} denotes the probabilistic value that the input instance x_i belongs to the category *j* through the classifier.

To ensure the EL-FFsol model was trained normally, several hyperparameters were applied. The hyperparameter values are exhibited in Table 2.

Table 2: Hyperparameter values

Hyperparameter	Value
eval_metric	accuracy
learning_rate	0.01
iterations	70,000
l2_leaf_reg	49
depth	9
early_stopping_rounds	15,000
border_count	64

3. Experiment and Analysis

3.1. Evaluation metrics

Several evaluation metrics utilized in EL-FFsol included accuracy (ACC), matthews correlation coefficient (MCC), sensitivity (SE), specificity (SP) and F1-score. These metrics are expressed in formula (2) to formula (6).

$$\text{ACC}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (2)$$

$$\text{MCC}=\frac{\text{TP}\times\text{TN-FP}\times\text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \quad (3)$$

$$\text{SE}=\frac{\text{TP}}{\text{TP+FN}} \quad (4)$$

$$\text{SP}=\frac{\text{TN}}{\text{TN+FP}} \quad (5)$$

$$\text{F1-score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \quad (6)$$

Where false positive (FP), false negative (FN), true positive (TP) and true negative (TN) can be combined to form a confusion matrix. Figure 3 shows the heat map of the confusion matrix.

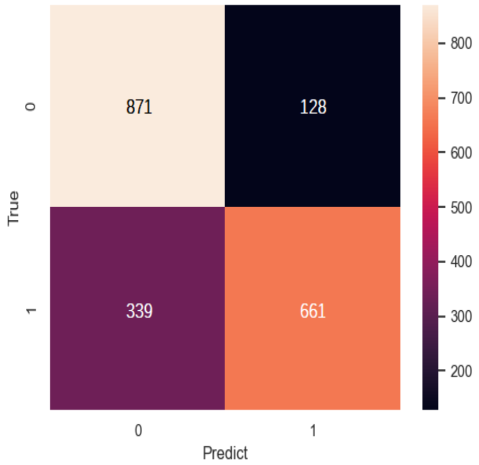


Figure 3: Heat map of the confusion matrix

3.2. Result in analysis

3.2.1. Feature comparison of protein sequences

In order to compare the effects of multiple features of protein sequences on the model performance, the Physicochemical Properties, One-hot Feature Encoding, Amino Acid Composition and Statistical Features of protein sequences were used for ablation studies. The ablation studies of multiple features are shown in Table 3.

Table 3: Ablation studies of multiple features

Features	ACC	MCC	SE	SP	F1-score
Statistical Features	0.6718	0.3679	0.4940	0.8498	0.6010
Physicochemical Properties	0.6893	0.3909	0.5660	0.8128	0.6458
One-hot Feature Encoding	0.6698	0.3741	0.4610	0.8789	0.5828
Amino Acid Composition	0.5718	0.1667	0.3190	0.8248	0.4270
Statistical Features + Physicochemical Properties	0.7494	0.5134	0.6310	0.8679	0.7158
Statistical Features + One-hot Feature Encoding	0.7479	0.5124	0.6220	0.8739	0.7117
Statistical Features + Physicochemical Properties + One-hot Feature Encoding	0.7658	0.5470	0.6490	**0.8829**	0.7350
Statistical Features + Physicochemical Properties + One-hot Feature Encoding + Amino Acid Composition	**0.7679**	**0.5480**	**0.6630**	0.8729	**0.7408**

As can be seen, directly using the Physicochemical Properties on the EL-FFsol model, the accuracy was 0.6893, matthews correlation coefficient was 0.3909, sensitivity was 0.5660 and f1-score was 0.6458. Compared with only making use of the Statistical Features, One-hot Feature Encoding and Amino Acid Composition, the accuracy was increased by 1.75%, 1.95% and 11.75%, the matthews correlation coefficient was increased by 2.30%, 1.68% and 22.42%, the sensitivity was increased by 7.20%, 10.50% and 24.70% and the f1-score was increased by 4.48%, 6.30% and 21.88%, respectively. Adopting the Statistical Features individually, EL-FFsol achieved the accuracy value of 0.6718, the sensitivity value of 0.4940 and the f1-score value of 0.6010. In addition, 0.20% and 10.00% improvements in accuracy, 3.30% and 17.50% improvements in sensitivity and 1.82% and 17.40% improvements in f1-score were acquired while contrasting with solely utilizing the One-hot Feature Encoding and Amino Acid Composition. It demonstrates that the Physicochemical Properties supply more protein characteristic information from the physical and chemical aspects through different attribute values of each amino acid residue, which may significantly increase the feature information in the data. In the Statistical Features, the Sequence Features describe the sequence information through the feature values of each protein sequence to a small degree. The Structural Features detail the local spatial structure of polypeptide chains in proteins. The Relative Solvent Accessibility represents whether the protein is exposed or hidden. Although some evaluation metrics obtained by using the Statistical Features were not as good as those acquired by applying the Physicochemical Properties, the Statistical Features can help sketch out the structural information of protein sequences. Employing the One-hot Feature Encoding or Amino Acid Composition merely narrates the basic information of amino acid residues that make up these proteins and the offered information is limited to physical structures and relationships between protein sequences. Therefore, certain evaluation metrics gained by making use of the One-hot Feature Encoding or Amino Acid Composition were worse than those achieved by utilizing the Physicochemical Properties or Statistical Features. Nevertheless, adopting the One-hot Feature Encoding can map discrete features to Euclidean space. Each amino acid residue is represented by a binary vector of length 20 and stored in a vertical space. Thus, relative to using the Amino Acid Composition, several evaluation metrics obtained by applying the One-hot Feature Encoding were better. The accuracy, matthews correlation coefficient, sensitivity, specificity and f1-score were improved by 9.80%, 20.74%, 14.20%, 5.41% and 15.58%.

Besides that, adopting the Statistical Features together with the Physicochemical Properties and One-hot Feature Encoding successively to train and test the EL-FFsol model. The results show that employing hybrid features can gain better evaluation metrics than utilizing a single feature. In particular, when the mixed features were composed of the Statistical Features and Physicochemical Properties, El-FFsol acquired 0.7494, 0.5134, 0.6310 and 0.7158 performances in the accuracy, matthews correlation coefficient, sensitivity and f1-score. Contrasted with making use of the Statistical Features and One-hot Feature Encoding, the accuracy was increased by 0.15%, the matthews correlation coefficient was increased by 0.10%, the sensitivity was increased by 0.90% and the f1-score was increased by 0.41%. It indicates that features including the Statistical Features and One-hot Feature Encoding cannot describe protein sequences in more detail to a certain degree. The main reason is that the feature information obtained by using the Statistical Features and One-hot Feature Encoding is not detailed in more attribute aspects contrasted with utilizing the Statistical Features and Physicochemical Properties.

Ultimately, based on features including the Statistical Features and Physicochemical Properties, the One-hot Feature Encoding and Amino Acid Composition were gradually added to the EL-FFsol model. It is noted that when all features are adopted in the model, EL-FFsol outperformed better than making use of other feature combinations and achieved 0.7679 in accuracy, 0.5480 in matthews correlation coefficient, 0.6630 in sensitivity and 0.7408 in f1-score. At this time, the protein sequence is described in the aspect of physical properties, chemical properties, amino acid residue information and structural features. The provided information is more abundant and comprehensive.

3.2.2. Performance comparison of CatBoost and Deep Learning

For the sake of mitigating several inadequacies of deep learning and enhancing the model performance, the CatBoost algorithm was employed to take the place of the convolutional neural network. Therefore, it is important to assess the performances between the CatBoost algorithm and deep learning. The multiple features of protein sequences used in this paper were first fused to form a brand-new feature vector. And then DeepSOL, DDcCNN and the model based on the CatBoost algorithm were utilized to predict protein solubility on the independent test set. The performance comparison of CatBoost and Deep Learning is expressed in Table 4.

Table 4: Performance comparison of CatBoost and Deep Learning

Models and Methods	ACC	MCC	SE	SP
DeepSOL	0.7630	0.5363	0.6540	0.8658
DDcCNN	0.7634	0.5419	0.6560	0.8488
CatBoost	**0.7679**	**0.5480**	**0.6630**	**0.8729**

It can be seen that the model based on the CatBoost algorithm acquired better metrics and attained an accuracy of 0.7679, a matthews correlation coefficient of 0.5480, a sensitivity of 0.6630 and a specificity of 0.8729. Relative to the DeepSOL and DDcCNN, the accuracy, matthews correlation coefficient, sensitivity and specificity were increased by 0.49% and 0.45%, 1.17% and 0.61%, 0.90% and 0.70% and 0.71% and 2.41%, respectively. It indicates that the CatBoost algorithm is more suitable than deep learning in hybrid feature fusion problems and improves the model performance and generalization. The foremost reason is that in the training process when using deep learning, firstly, the model with numerous convolutional layers leads to an increased number of parameters and several deficiencies in performance and efficiency. And then the model with few convolutional layers brings about inadequate training and performance limitations. Additionally, a number of operations in pooling layers produce the loss of certain vital information and the connection between the whole and the parts may be weakened. The CatBoost algorithm not only avoids these problems but also effectively controls overfitting through the Ordered Boosting method. To realize this purpose, in each iteration of gradient boosting, a certain sample is deleted from the training set of the current ensemble model to ensure the authenticity of the gradient estimate of each sample. Furthermore, the CatBoost algorithm considers the combination with the greedy algorithm to improve the accuracy of the current tree when creating split nodes.

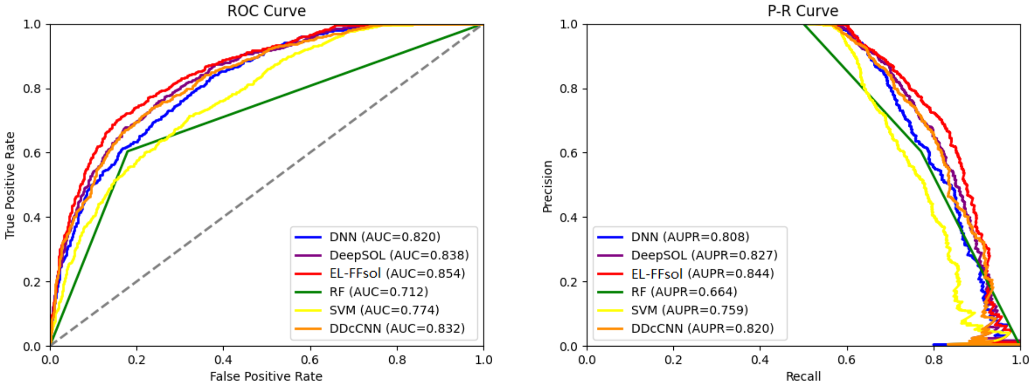
3.2.3. Overall performance comparison of different models and methods

For the aim of further verifying the overall model performance, the random forest (RF), support vector machine (SVM), deep neural network (DNN) and several existing prediction methods were contrasted with the EL-FFsol model on the independent test set. The overall performance comparison of different models and methods is shown in Table 5.

Table 5: Overall performance comparison of different models and methods

Models and Methods	ACC	MCC	SE	SP
RF	0.7019	0.4153	0.5850	0.8188
SVM	0.7273	0.4709	0.5980	0.8569
PaRSnIP	0.7411	0.4811	-	-
DNN	0.7464	0.5065	0.6310	0.8619
SCNN	0.7556	0.5197	0.6410	0.8318
DCNN	0.7568	0.5211	0.6410	0.8324
DTcCNN	0.7582	0.5284	0.6430	0.8341
DeepSOL	0.7625	0.5312	0.6550	0.8658
DDcCNN	0.7631	0.5401	0.6530	0.8417
My model	**0.7679**	**0.5480**	**0.6630**	**0.8729**

From Table 5, it follows that EL-FFsol acquired accuracy in 0.7679, matthews correlation coefficient in 0.5480, sensitivity in 0.6630 and specificity in 0.8729. Compared with methods founded on the RF and SVM, the accuracy was increased by 6.60% and 4.06%, the matthews correlation coefficient was increased by 13.27% and 7.71%, the sensitivity was increased by 7.80% and 6.50% and the specificity was increased by 5.41% and 1.60%, respectively. And contrasted with the DeepSOL and DDcCNN, EL-FFsol achieved 0.54% and 0.48% improvements in accuracy, 1.68% and 0.79% improvements in matthews correlation coefficient, 0.80% and 1.00% improvements in sensitivity and 0.71% and 3.12% improvements in specificity. Additionally, Figure 8 plots the ROC and P-R curves of different models and methods. It can be seen that the EL-FFsol model wrapped other models in a wider space, where the AUC and AUPR of the EL-FFsol model were 0.8540 and 0.8440, respectively. Meanwhile, the AUC and AUPR of EL-FFsol were 1.60% and 1.70% higher than DeepSOL. In addition, the EL-FFsol improved the AUC by 2.20% and AUPR by 2.40% compared with the DDcCNN. The experimental results express that the overall performance of the model based on the CatBoost ensemble learning framework and multiple feature fusion of protein sequences is better than other models. Thus, EL-FFsol can predict the solubility of proteins more accurately and reliably.



(a) (b)

Figure 4: ROC and P-R curves: (a)ROC curves of different models and methods; (b)P-R curves of different models and methods

4. Conclusions

EL-FFsol implements feature fusion on multiple features, including the Physicochemical Properties, One-hot Feature Encoding, Amino Acid Composition and Statistical Features of protein sequences. The Physicochemical Properties represent the physical properties and chemical properties of each amino acid residue, which are consisted of Hydropathy Index, Electron-Ion Interaction Potential, Molecular Weight, Residue Molecular Weight, Charge and Polarity and Acid-base Features. The Statistical Features detail the protein sequence in the aspect of physicochemical and structural attributes, which are composed of Sequence Features, Structural Features and Relative Solvent Accessibility. The One-hot Feature Encoding and Amino Acid Composition narrate the basic combination information of amino acid residues. Therefore, the constructed feature information is more abundant and comprehensive. Furthermore, the CatBoost is an algorithm for gradient enhancement of decision trees together with gradient boosting and categorical knowledge, which can slow down the gradient bias, effectively control overfitting and improve the model generalization by adopting the Ordered Boosting method when building the tree for an unbiased gradient estimation in each iteration. Therefore, the CatBoost ensemble learning framework is more suitable than deep learning in hybrid feature fusion problems.

Therefore, in this paper, the EL-FFsol model based on ensemble learning and feature fusion was designed for the prediction of protein solubility. Through the CatBoost ensemble learning framework, EL-FFsol can better control overfitting and enhance the model performance and generalization. Furthermore, the fusion of protein sequence features can help the model acquire more data information in the sequence data. The experimental results indicate that several evaluation metrics achieved by EL-FFsol are better than those gained by the existing protein solubility prediction method. The accuracy, matthews correlation coefficient, sensitivity, specificity, area under ROC curve and area under P-R curve were 0.7679, 0.5480, 0.6630, 0.8729, 0.8540 and 0.8440, respectively. Contrasted with the DeepSOL and DDcCNN, EL-FFsol improved the matthews correlation coefficient by 1.68% and 0.79%, the area under

ROC curve by 1.60% and 2.20% and the area under P-R curve by 1.70% and 2.40%, respectively. Thus, EL-FFsol can predict protein solubility with high confidence.

In future research, the model compression technology and knowledge distillation technology may be used to reduce the model reasoning time to further improve the efficiency and performance of EL-FFsol. Furthermore, the structural information of protein sequences is significant. Thus, exploring the solubility of proteins from structural information through deep learning is the next important work.

References

[1] Tanaka S, Takizawa K, Nakamura F. One-step visualization of natural cell activities in non-labeled living spheroids. Sci Rep 2022; 12:1–11.
[2] Cho H, Li Y, Archacki S, et al. Splice variants of lncRNA RNA ANRIL exert opposing effects on endothelial cell activities associated with coronary artery disease. RNA Biology 2020; 17:1391–1401
[3] Monteiro L, Da Silva L, Lipinski B, et al. Assessing Cell Activities rather than Identities to Interpret Intra-Tumor Phenotypic Diversity and Its Dynamics. iScience 2020; 23:101061.
[4] Havugimana PC, Hart GT, Nepusz T, et al. A Census of Human Soluble Protein Complexes. Cell 2012; 150:1068–1081.
[5] Aqeel A, Hassan A, Khan MA, et al. A Long Short-Term Memory Biomarker-Based Prediction Framework for Alzheimer's Disease. Sensors 2022; 22:1475.
[6] Meng L, Li X, Li C, et al. Effects of Exercise in Patients With Amyotrophic Lateral Sclerosis: A Systematic Review and Meta-Analysis. American Journal of Physical Medicine & Rehabilitation 2020; 99:801–810.
[7] Goh GS, Zeng GJ, Tay DK, et al. Patients With Parkinson's Disease Have Poorer Function and More Flexion Contractures After Total Knee Arthroplasty. The Journal of Arthroplasty 2021; 36:2325–2330.
[8] Vihinen M. Solubility of proteins. ADMET and DMPK 2020; 8:391–399.
[9] Marmamula S, Barrenakala NR, Challa R, et al. Visual outcomes after cataract surgery among the elderly residents in the 'homes for the aged' in South India: the Hyderabad Ocular Morbidity in Elderly Study. British Journal of Ophthalmology 2021; 105:1087–1093.
[10] Peng Y, Zhang C, Rui Z, et al. A comprehensive profiling of soluble immune checkpoints from the sera of patients with non-small cell lung cancer. Journal of Clinical Laboratory Analysis 2022; 36:e24224.
[11] Gu D, Ao X, Yang Y, et al. Soluble immune checkpoints in cancer: production, function and biological significance. j. immunotherapy cancer 2018; 6:132.
[12] Wang X-F, Gao P, Liu Y-F, et al. Predicting Thermophilic Proteins by Machine Learning. Current Bioinformatics 2020; 15:493–502.
[13] Parsons CM, Hashimoto K, Wedekind KJ, et al. Soybean protein solubility in potassium hydroxide: an in vitro test of in vivo protein quality. Journal of Animal Science 1991; 69:2918–2924.
[14] Rizvi NB, Aleem S, Khan MR, et al. Quantitative Estimation of Protein in Sprouts of Vigna radiate (Mung Beans), Lens culinaris (Lentils), and Cicer arietinum (Chickpeas) by Kjeldahl and Lowry Methods. Molecules 2022; 27:814.
[15] Hou Q, Bourgeas R, Pucci F, et al. Computational analysis of the amino acid interactions that promote or decrease protein solubility. Sci Rep 2018; 8:1–13.
[16] Guzman-Chavez F, Arce A, Adhikari A, et al. Constructing Cell-Free Expression Systems for Low-Cost Access. ACS Synth. Biol. 2022; 11:1114–1128.
[17] Yan M, Zhang X, Hu L, et al. Bacterial Community Dynamics During Nursery Rearing of Pacific White Shrimp (Litopenaeus vannamei) Revealed via High-Throughput Sequencing. Indian J Microbiol 2020; 60:214–221.
[18] Grant-Kels JM, Sloan B, Kantor J, et al. Big data and cutaneous manifestations of COVID-19. Journal of the American Academy of Dermatology 2020; 83:365–366.
[19] Han G-S, Yu Z-G, Anh V. A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC. Journal of Theoretical Biology 2014; 344:31–39.
[20] Magnan CN, Randall A, Baldi P. SOLpro: accurate sequence-based prediction of protein solubility. Bioinformatics 2009; 25:2200–2207.
[21] Smialowski P, Doose G, Torkler P, et al. PROSO II – a new method for protein solubility prediction. The FEBS Journal 2012; 279:2192–2200.
[22] Agostini F, Cirillo D, Livi CM, et al. cc SOL omics : a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli. Bioinformatics 2014; 30:2975–2977.
[23] Rawi R, Mall R, Kunji K, et al. PaRSnIP: sequence-based protein solubility prediction using

gradient boosting machine. Bioinformatics 2018; 34:1092–1098.

[24] Savojardo C, Bruciaferri N, Tartari G, et al. DeepMito: accurate prediction of protein sub-mitochondrial localization using convolutional neural networks. Bioinformatics 2020; 36:56–64.

[25] Khurana S, Rawi R, Kunji K, et al. DeepSol: a deep learning framework for sequence-based protein solubility prediction. Bioinformatics 2018; 34:2605–2613.

[26] Wang X-F, Liu Y-F, Du Z-Y, et al. Design of protein solubility prediction model based on deep neural network. Journal of Henan Normal University(Natural Science Edition) 2021; 49:31–39.

[27] Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012; 28:3150–3152.

[28] Chang CCH, Song J, Tey BT, et al. Bioinformatics approaches for improved recombinant protein production in Escherichia coli: protein solubility prediction. Briefings in Bioinformatics 2014; 15:953–962.

[29] Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003; 19:1589–1591.

[30] Cheng J, Randall AZ, Sweredoski MJ, et al. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Research 2005; 33:W72–W76.

[31] Seibert P, Raßloff A, Ambati M, et al. Descriptor-based reconstruction of three-dimensional microstructures through gradient-based optimization. Acta Materialia 2022; 227:117667.

[32] Zhang F, Fleyeh H, Bales C. A hybrid model based on bidirectional long short-term memory neural network and Catboost for short-term electricity spot price forecasting. Journal of the Operational Research Society 2022; 73:301–325.