Proceedings of ICMMGH 2026 Symposium: Biomedical Imaging and AI Applications in Neurorehabilitation
DOI: 10.54254/2753-8818/2026.AU28733

Protein Secondary Structure Prediction Based on SSA-DNN Regression Algorithm

Feiyang Yang

Shanghai Pinghe School, Shanghai, China
yangfeiyang1221@gmail.com

Abstract. Protein secondary structure prediction is a fundamental task in bioinformatics, crucial for understanding protein function and guiding drug discovery. Traditional regression and ensemble models show limited performance due to their inability to capture nonlinear dependencies and sequential features of protein sequences. To address these challenges, this study proposes a hybrid model that integrates the Sparrow Search Algorithm (SSA) with Deep Neural Networks (DNN). SSA optimizes the initialization and hyperparameters of DNN, improving convergence and generalization. Furthermore, Online Low-rank Subspace Tracking by Tensor Decomposition (OLSTEC) is incorporated to exploit multi-dimensional correlations among sequence, evolutionary, and physicochemical features. Experimental results demonstrate that the SSA-DNN framework achieves superior accuracy over regression baselines, and the addition of OLSTEC further improves test accuracy to 36.82% with a Macro-F1 score of 0.1555. These findings highlight the advantages of combining metaheuristic optimization with tensor decomposition for large-scale protein structure prediction.

Keywords: Protein secondary structure prediction, deep neural networks (DNN), Sparrow Search Algorithm (SSA), OLSTEC tensor decomposition, time-series modeling

1. Introduction

Protein secondary structure prediction is a crucial step in understanding protein function and 3D conformation, with broad applications in molecular biology and drug development. Enhancing the accuracy and robustness of prediction models has become a key research focus.

The formation of protein structures is influenced by amino acid sequences, physicochemical properties, and evolutionary context. These factors interact in highly nonlinear and complex ways, making it difficult for traditional mathematical models to capture the full structural rules.

Artificial neural networks, known for their adaptability and strong learning capabilities, have been widely used to address such nonlinear problems. Among them, deep neural networks (DNNs) have shown strong performance in feature extraction due to their multi-layered architecture and are commonly applied to protein structure prediction. However, the training of DNNs is often sensitive to parameter initialization and prone to local minima, which can hinder overall performance.

To address these challenges, this study integrates the Sparrow Search Algorithm (SSA) with DNNs to build an SSA-DNN hybrid model. SSA serves to optimize the initial parameters of the

DOI: 10.54254/2753-8818/2026.AU28733

network, improving convergence speed and generalization. By combining evolutionary profiles, sequence features, and physicochemical descriptors, the proposed approach offers an efficient solution to the problem of protein secondary structure prediction.

2. Principle of the SSA-DNN regression algorithm

2.1. Principle of Sparrow Search Algorithm (SSA)

The SSA is a nature-inspired optimization technique that simulates the foraging and anti-predation behaviors of sparrows. The algorithm divides the population into three types of roles: producers, scroungers, and scouts, each with specific responsibilities and update mechanisms. By randomly initializing individual positions and updating them based on fitness values, the algorithm aims to find the optimal solution to the problem.

Producers account for 10% to 20% of the population and are responsible for global exploration of food sources. Their position update is governed by a vigilance threshold, distinguishing between safe and risky states [1,2]:

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^{t} \cdot \exp\left(\frac{-i}{\alpha \cdot m}\right) & \text{if } R_2 < ST \\ X_{i,j}^{t} + \mu \cdot L & \text{if } R_2 \geq ST \end{cases} \quad (1)$$

Where, $X_{i,j}^{t+1}$: updated position of the i -th individual at iteration $t+1$; $X_{i,j}^{t}$: global historical best position;

μ : random variables obeying normal distribution;

L: a matrix of 1×d for which each element inside is 1;

R_2 : ($R_2 \in [0,1]$) alert value;

ST : ($ST \in [0.5,1.0]$) safety threshold;

m : maximum number of iterations;

α : ($\alpha \in (0,1]$) a random number.

Scroungers follow producers to perform local search. The population is split into two halves, each with distinct update rules:

$$X_{i,j}^{t+1} = \begin{cases} \mu \cdot \exp\left(\frac{X_{worst}^{t} - X_{i,j}^{t}}{i^2}\right) & \text{if } i > n/2 \\ X_{P}^{t+1} + \left| X_{i,j}^{t} - X_{P}^{t+1} \right| \cdot A^{+} \cdot L & \text{otherwise} \end{cases} \quad (2)$$

Where, X_{worst}^{t} : worst position in the current population;

X_{P}^{t+1} : best position among producers at iteration $t+1$;

A : a matrix of 1×d for which each element inside is randomly assigned 1 or -1;

A^{+} : $A^{m}(AA^{m})^{-1}$.

10% to 20% of sparrows are designated as scouts to detect external threats. To reduce their predation risk, their position is updated based on global best information:

DOI: 10.54254/2753-8818/2026.AU28733

$$X_{i,j}^{t+1} = \begin{cases} X_{best}^{t} + \beta \cdot \left| X_{i,j}^{t} - X_{best}^{t} \right| & \text{if } f_i > f_g \\ X_{i,j}^{t} + k \cdot \left(\frac{\left| X_{i,j}^{t} - X_{worst}^{t} \right|}{(f_i - f_w) + \gamma} \right) & \text{if } f_i = f_g \end{cases} \tag{3}$$

Where: X_{best}^{t} : current globally best position; f_i : fitness of the i -th individual; f_g : current best fitness in the population;

f_w : current worst fitness in the population; k , β : step control parameters; γ : a small constant to avoid division by zero.

2.2. Deep Neural Network (DNN)

Deep neural networks (DNNs) are a type of machine learning model rooted in the broader field of artificial intelligence (AI), which aims to enable machines to perform tasks that typically require human intelligence. Within AI, machine learning refers to techniques that allow systems to improve automatically through experience. DNNs represent a specialized category of machine learning that takes inspiration from the structure of the human brain [3].

In biological terms, neurons communicate via electrical signals. Each neuron receives inputs, processes them, and transmits outputs to other neurons through connections known as synapses. This signal processing is roughly analogous to operations in an artificial neural network, where each connection has an associated weight that determines its influence. When such artificial models consist of many layers, they are referred to as deep neural networks. By modeling this behavior computationally, artificial neural networks were developed, and when such networks involve many layers of neurons, they are referred to as deep neural networks.

Deep Neural Networks (DNNs) are composed of multiple layers of interconnected neurons, including an input layer, several hidden layers, and an output layer. Each neuron computes a weighted sum of its inputs, adds a bias, and applies a nonlinear activation function:

$$y_j = f\left(\sum_i \varphi_{ij} x_i + k\right) \tag{4}$$

where:

x_i : input from the i -th neuron in the previous layer; φ_{ij} : weight between input x_i and output neuron j ;

k : bias term added to the weighted sum; $f(\cdot)$: nonlinear activation function (e.g., ReLU, sigmoid);

y_j : output of the j -th neuron after activation.

The layered structure enables DNNs to learn hierarchical features from data. For instance, in image processing, lower layers may detect edges, middle layers may capture shapes, and deeper layers may identify high-level objects.

DNN training involves two key phases: forward propagation and backward propagation. During forward propagation, the input is passed through the network to generate an output. In training, a loss function quantifies the prediction error, and the gradient of this loss with respect to each weight is computed via backpropagation. The weights are then updated using gradient descent:

$$\varphi_{ij}^{t+1} = \varphi_{ij}^{t} - \theta \frac{\partial L}{\partial \varphi_{ij}} \tag{5}$$

DOI: 10.54254/2753-8818/2026.AU28733

In this formula:

φ_{ij}^{t} : the weight between input i and neuron j at iteration t;

$\frac{\partial L}{\partial \varphi_{ij}}$: the partial derivative of the loss with respect to the weight, indicating the gradient direction;

θ : the learning rate that controls the step size.

The subtraction updates the weight in the direction that reduces the loss, which is the essence of the gradient descent algorithm.

This iterative process continues until the network achieves acceptable performance. DNNs have been successfully applied in fields such as image classification, speech recognition, and time-series prediction, demonstrating strong capabilities in modeling complex, nonlinear data relationships.

2.3. Optimization of DNN using SSA

SSA-DNN refers to a hybrid framework that integrates the global optimization ability of the SSA with the nonlinear modeling capability of DNNs. While DNNs excel at learning complex input–output relationships, their performance can be affected by factors such as poor parameter initialization and slow convergence. SSA is introduced as a metaheuristic search strategy to improve the training process and enhance model generalization.

SSA-DNN neural network flowchart is shown in Figure 1.

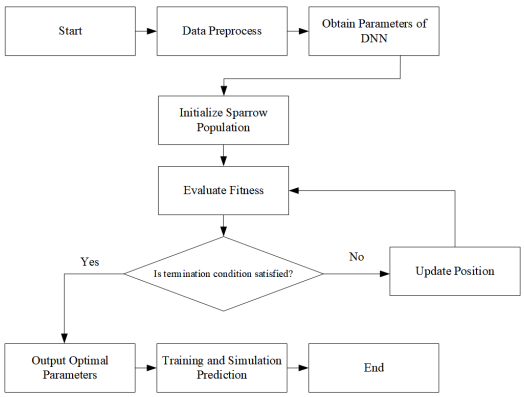

Figure 1. The process begins with data preprocessing and parameter acquisition for the deep neural network

The sparrow search algorithm (SSA) is then initialized to iteratively update candidate solutions based on fitness evaluations. The loop continues until the termination condition is met, at which point the optimal parameters are output and used for final training and simulation prediction.

3. Protein secondary structure prediction

3.1. Hierarchical classification of protein structures

Proteins are composed of amino acids. There are over 500 different amino acids and 20 standard essential amino acids in nature, all sharing a common amino group and carboxyl group, but differing in their side chains (R groups) [4]. These variations in side chain composition lead to diverse combinations of amino acids, which give rise to a wide variety of proteins, each with its own distinct conformations (also known as structure decoys).

DOI: 10.54254/2753-8818/2026.AU28733

Through peptide bonds formed between amino acids, linear sequences of amino acid residues called peptide chains are generated, which constitute the primary structure of a protein. This level has no three-dimensional configuration and can be represented by a simple string of letters. Under the influence of hydrogen bonds, the peptide chain folds into various local regular sub-structures, such as alpha helices, beta-pleated sheets, and beta turns, which together define the secondary structure.

Building upon these local secondary structures, additional interactions—including ionic bonds, hydrogen bonds, and disulfide bridges—induce further folding of the protein into a stable three-dimensional configuration, known as the tertiary structure. This conformation corresponds to the native-stable state, which represents the minimum energy configuration of the protein.

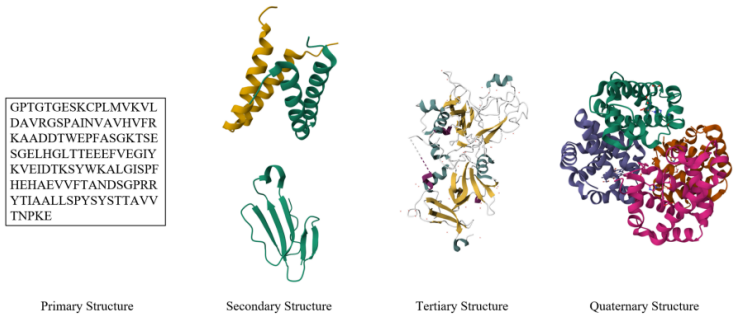

Figure 2. The four levels of protein structure: primary (amino acid sequence), secondary (local structures like α-helices and β-sheets), tertiary (3D folding of a single chain), and quaternary (assembly of multiple chains)

The quaternary structure arises when multiple peptide chains assemble into a complex protein structure. This may involve only polypeptide chains (non-conjugated proteins), or polypeptide chains in combination with non-polypeptide units (conjugated proteins).

Collectively, the primary, secondary, tertiary, and quaternary structures (as shown in Figure 2) determine the overall conformation of a protein, and this conformation in turn dictates the protein's biological function.

3.2. Common input features

The effective secondary structure prediction relies heavily on the design of informative input features that capture both evolutionary and physicochemical contexts. A widely used class of features is derived from multiple sequence alignment (MSA), which aggregates homologous sequences to reveal conservation patterns. From these alignments, position-specific scoring matrices (PSSMs) are computed to encode the log-odds of observing each amino acid at a given sequence position, offering residue-level insights into evolutionary constraints.

$$PSSM\,[i]\,[j] = \log\left(\frac{P(\alpha_j|i)}{P(\alpha_j)}\right) \tag{6}$$

Where:

$P(\alpha_j|i)$ is the probability of observing amino acid α_j at position i in the multiple sequence alignment;

DOI: 10.54254/2753-8818/2026.AU28733

$P(\alpha_j)$ is the background frequency of amino acid α_j across the general protein database.

Alternatively, hidden Markov models (HMMs) generated via tools such as HHblits can provide complementary probabilistic profiles, including emission and transition probabilities across alignment states [5]. In addition to evolutionary information, the raw primary sequence is typically represented using one-hot encoding, where each amino acid is mapped to a unique 20-dimensional binary vector, ensuring that the sequence identity is explicitly preserved. Furthermore, physicochemical descriptors—such as hydrophobicity, polarity, molecular volume, and solvent accessibility—can be incorporated to reflect the chemical environment of each residue [6]. By concatenating these heterogeneous features, a comprehensive input representation is formed for each position in the sequence, enabling deep neural networks to learn complex mappings from sequence to structural space.

3.3. Deep learning-based PSSP workflow

As shown in Figure 3, the process begins by collecting homologous sequences for a target protein through MSA techniques, such as DeepMSA, which searches protein sequence databases to identify evolutionarily related proteins. The resulting alignments are used to compute PSSMs and HMMs, which encode the conservation and variation patterns at each residue position. These features, along with primary sequence identity and physicochemical descriptors, form the input to deep learning models.

In parallel, structural annotations are derived from known protein structures using Define Secondary Structure of Proteins (DSSP), which labels each residue in experimentally solved structures with its secondary structure class. These labels serve as training targets for supervised learning. The combined features are then processed through a feature extraction module—such as Meiler's encoding scheme or T5 protein language model—and passed into deep neural networks. Some models also incorporate pre-trained protein language models to capture contextual relationships within protein sequences. The trained model outputs predicted secondary structures, which can aid in downstream tasks such as tertiary structure prediction, function annotation, or molecular docking analysis.

4. Case study

4.1. Regression model

Regression and ensemble models, including Logistic Regression, SVM, KNN, Random Forest, XGBoost, LightGBM, and CatBoost, were tested as baselines. Their accuracies remained at 14.70%–23.3%. This result confirmed that conventional regression models fail to capture the nonlinear and sequential patterns of protein secondary structures. These results provide only weak references for subsequent model improvements.

4.2. Time series model

To address this limitation, a Sparrow Search Algorithm optimized Deep Neural Network (SSA-DNN) was implemented. SSA improves initialization and hyperparameter settings, enhancing convergence and robustness. Without tensor decomposition, the SSA-DNN achieved around 26.00% test accuracy. The result outperformed conventional regression models. This demonstrates the

DOI: 10.54254/2753-8818/2026.AU28733

advantage of sequence-based deep learning combined with evolutionary optimization, though further improvements were still necessary.

4.3. Advantages of tensor decomposition

By integrating OLSTEC tensor decomposition, multi-dimensional correlations among sequence, evolutionary, and physicochemical features were more effectively preserved. The combined SSA-DNN + OLSTEC achieved 37.00% training accuracy, 36.82% test accuracy, and a Macro-F1 of 0.1555, representing a ~10% improvement over plain SSA-DNN.

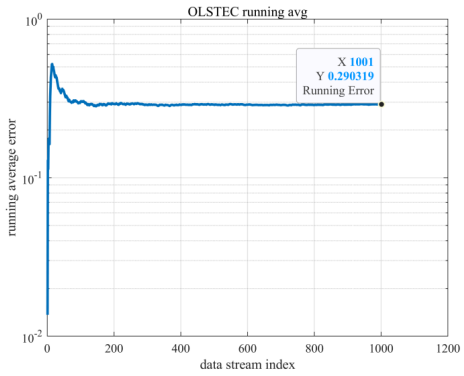

Figure 4. Running average error of OLSTEC

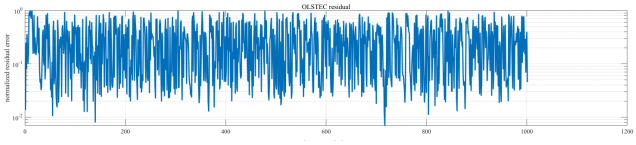

Figure 5. Residual error distribution of OLSTEC

Figure 6. Confusion matrix of the test set

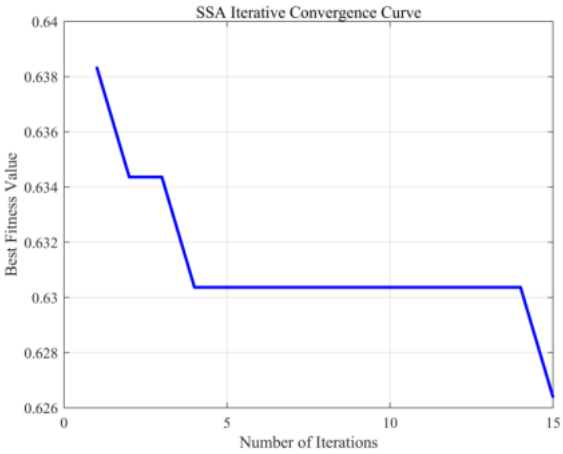

Figure 7. Iterative convergence curve of SSA

In Figure 6, Classes 1 and 2 are recognized with moderate accuracy, while minority classes remain more difficult to classify, reflecting the effect of class imbalance. In Figure 7, the best fitness value steadily decreases over iterations, confirming the effectiveness of the Sparrow Search Algorithm in optimizing DNN parameters.

DOI: 10.54254/2753-8818/2026.AU28733

The benefits are also reflected in the visual analysis. Figure 5 shows the residual error distribution, indicating fluctuations but overall stability around 10^{-1}. Figure 4 presents the running average error, which converges near 0.29. This confirmed OLSTEC's stability. Figure 6 gives the confusion matrix of the test set, where classes 1 and 2 are recognized with moderate success, but minority classes remain challenging, highlighting class imbalance issues. Figure 7 shows the SSA convergence curve, where fitness decreases steadily across iterations, demonstrating the effectiveness of SSA in optimizing the DNN parameters. Together, these analyses confirm that OLSTEC not only improves predictive performance but also enhances model stability and interpretability.

Table 1. Overall performance of all the models

Model	Test Accuracy (%)
Logistic Regression	14.70
SVM	14.70
KNN	23.30
Decision Tree	15.30
Random Forest	13.98
AdaBoost	15.30
ExtraTrees	14.70
XGBoost	14.70
GBDT	14.70
LightGBM	15.30
CatBoost	14.70
BP Neural Network	14.70
SSA-DNN	26.00
SSA-DNN + OLSTEC	36.82

As shown in Table 1, traditional regression and ensemble models such as Logistic Regression, SVM, and Random Forest achieved relatively low accuracies, mostly between 13% and 15%, indicating their limited ability to capture the nonlinear and sequential dependencies of protein structures. Among these, KNN reached the highest accuracy of 23.30%, but the overall performance remained unsatisfactory. The BP neural network slightly improved feature learning but still failed to achieve stable convergence.

In contrast, the proposed SSA-DNN model substantially improved predictive performance, achieving a test accuracy of 26.00%, which demonstrates the advantage of combining time-series deep learning with evolutionary optimization. When the OLSTEC tensor decomposition was integrated, the model achieved a further increase in accuracy to 36.82%, highlighting the effectiveness of multi-dimensional tensor representation in enhancing both precision and model robustness.

5. Conclusion

This study presented a progressive pipeline for protein secondary structure prediction, moving from conventional regressors to a time-series deep model and finally to a tensorized framework. Classical regression and ensemble baselines—Logistic Regression, SVM, KNN, Decision Tree, Random

DOI: 10.54254/2753-8818/2026.AU28733

Forest, AdaBoost, ExtraTrees, XGBoost, GBDT, LightGBM, CatBoost, and a shallow BP network—delivered low test accuracies (≈14.70-23.30%), confirming that hand-crafted, vectorized features are insufficient to capture the nonlinear dependencies and contextual signals embedded in protein sequences. Building on this observation, a Sparrow Search Algorithm-optimized deep neural network (SSA-DNN) was introduced. By globally tuning initialization and hyperparameters, SSA improved convergence stability and helped the network escape poor local minima, raising test accuracy to 26.00% and establishing the value of time-series modeling combined with metaheuristic optimization.

To further exploit multi-way structure in the data, this study integrated OLSTEC tensor decomposition with SSA-DNN. Modeling sequence, evolutionary, and physicochemical descriptors as a high-order tensor allowed online low-rank factorization to denoise inputs and preserve cross-mode correlations. The combined model achieved 37.00% training accuracy, 36.82% test accuracy, and a Macro-F1 of 0.1555, outperforming plain SSA-DNN by ≈10.8 percentage points. Convergence analyses supported these gains: OLSTEC's running average error rapidly stabilized (~0.29), residuals fluctuated within a narrow band (~10-1), and SSA's fitness curve descended steadily across iterations, indicating effective parameter search and stable online updates. The confusion matrix revealed that minority classes remain harder to recognize, explaining the gap between accuracy and Macro-F1 and pointing to class imbalance as a principal bottleneck.

References

[1] J. Xue and B. Shen. (2020) A novel swarm intelligence optimization approach: sparrow search algorithm, Systems Science & Control Engineering, vol. 8, no. 1, pp. 22–34. doi: 10.1080/21642583.2019.1708830.

[2] Y. Zhou, S. Ye, and L. Yang. (2024) Research on industrial robot positioning error compensation based on improved SSA–DNN, Autom. Appl., vol. 65, no. 3. DOI: 10.19769/j.zdhy.2024.03.004.

[3] V. Sze, Y.H. Chen, T.J. Yang, and J. S. Emer. (2017) Efficient Processing of Deep Neural Networks: A Tutorial and Survey, Proc. IEEE, vol. 105, no. 12, pp. 2295–2329. doi: 10.1109/JPROC.2017.2761740.

[4] W. Wardah, M. G. M. Khan, A. Sharma, and M. A. Rashid. (2019) Protein secondary structure prediction using neural networks and deep learning: A review. Computational Biology and Chemistry, vol. 81, pp. 1–8. doi: 10.1016/j.compbiolchem.2019.107093.

[5] J. Zhao. (2024) Research on protein secondary structure prediction based on deep learning. M.S. thesis, Control Theory and Control Engineering, Xi'an Technological University, Xi'an, Shaanxi, China. DOI: 10.27391/d.cnki.gxagu.2024.000040.

[6] J. Liang, J. Liu, X. Guan, and Y. Chen. (2024) Research on prediction and optimization of cereal protein secondary structure based on multi-kernel LSSVM. J. Food Sci. Biotechnol., vol. 43, no. 7, July. DOI: 10.12441/jxsjwsx.20211221003.