Construction and Validation of a Supply Chain Demand Forecasting Model Based on Embedded AI

Yonghui Ding[1*], Cancan Li[2]
[1]Business School, Nantong Institute of Technology, Nantong 226001, China
[2]The Logistics and Capital Construction Department, Nanjing University of Finance and Economics, Nanjing 210023, China
E-mail: dingyh201606@163.com
[*]Corresponding author

Keywords: embedded, demand, forecasting, BPNN, CNN-LSTM, parameter optimization

Received: July 4, 2025

Supply chain demand forecasting is an important basis for enterprise decision-making, which can not only help optimize resource allocation, but also enhance the market competitiveness of enterprises. To improve the accuracy of supply chain demand forecasting, a combined forecasting model considering univariate and multivariate variables is designed and embedded. On the univariate prediction model, the study considers optimizing the parameters of the backpropagation neural network through an improved whale optimization algorithm. On the multivariate prediction model, the study combines improved particle swarm optimization algorithm, convolutional neural network, and long short-term memory network. The findings showed that the accuracy, root mean square error, time consumption, and maximum memory usage of the univariate prediction model were 98.05%, 1.03%, 61ms, and 10.85%, respectively, which were significantly better than the comparison model. The maximum accuracy of the multivariate prediction model was 98.51%, the minimum was 96.02%, and the maximum root mean square error was 0.58. After embedded deployment, the maximum increase in time consumption of the combined prediction model was 47.76%, and the accuracy only decreased by 0.18%. The designed combination forecasting model has good performance and can provide model support for predicting supply chain demand.

Povzetek: Kombinirani napovedni model IWOA-BPNN in IPSO-CNN-LSTM, zasnovan za vgradne AI platforme, omogoča bolj kvalitetno napoved povpraševanja v dobavni verigi kot modeli ARIMA, RF, SVM in standardni CNN-LSTM, saj dosega višjo TOčnost, manjše napake ter učinkovitejšo porabo časa in pomnilnika.

1 Introduction

In the context of global economic integration and deep integration of digital technology, supply chain management has become a key battlefield for companies to enhance their core competitiveness. As the "nerve centre" of supply chain management, the forecasting of demand in the supply chain has been demonstrated to exert a direct influence on the management of inventory, the planning of production and the scheduling of logistics within enterprises. In addition, it is also related to customer satisfaction and the rate of response of markets [1-2]. According to McKinsey research, accurate demand forecasting can reduce inventory costs for businesses by 20%-30% and increase order fulfillment rates by 15%-25% [3-4]. Therefore, it is necessary to improve the accuracy of Supply Chain Demand Forecasting (SCDF). With regard to the issue in question, the extant literature proposes a number of methodologies for analysis. These include the application of statistical models based on historical data, human experience and judgement, time series analysis, causal models, and machine learning algorithms [5]. In addition, many researchers have already explored this issue.

Liu et al. developed a prediction model that integrates Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) algorithms to predict supply chain dynamics and address inventory optimization issues. The model was hyperparameter adjusted through Bayesian optimization. The findings denoted that this method could improve the operational efficiency and cost control effectiveness of enterprises [6]. Saad et al. developed a prediction algorithm that integrates an attention mechanism and bidirectional LSTM to precisely capture the dynamic fluctuations and patterns of supply chain logistics demand. This algorithm was trained through the utilization of the gradient descent method. Meanwhile, the study employed local outlier factors to remove outliers from the data. The findings demonstrated the efficacy of the algorithm in predicting demand within the supply chain logistics domain. Subsequent application of the algorithm resulted in an on-time delivery rate exceeding 95% on a monthly basis [7]. Zhang et al. developed a prediction method that utilizes the sequence-to-sequence and attention mechanisms to forecast demand for environmentally friendly electronic products within supply chains. The method has been adapted to address the dynamic and complex demands of the environmental

protection market by incorporating an end-to-end multi-step time prediction approach. Concurrently, the study utilized a linear regression approach founded on Huber Loss for the purpose of data processing. The findings demonstrated that the methodology outlined in this study exhibited superiority over conventional time series prediction models and machine learning regression models with regard to prediction accuracy [8]. Hasan et al. explored the feasibility of machine learning techniques in the field of SCDF, using five techniques including linear regression, elastic networks, random forests, multi-layer perceptron regressors, and extreme gradient boosting to comprehensively analyze a company's historical data. The results showed that linear regression was the most effective model with smaller errors [9].

However, current research and methods also have certain problems, such as the susceptibility of human experience judgments to individual cognitive biases, and the high computational cost of machine learning algorithms. To improve the accuracy of SCDF, a Univariate Prediction Model (UPM) considering Backpropagation Neural Network (BPNN) parameter optimization and a Multivariate Prediction Model (MPM) based on CNN and LSTM were designed and deployed as a combined prediction model on an embedded Artificial Intelligence (AI) platform. The research aims to improve the accuracy of SCDF, help enterprises plan inventory levels reasonably, avoid excessive inventory backlog or stockouts, and thereby reduce inventory holding costs and stockout costs. The research innovation combines the Improved Whale Optimization Algorithm (IWOA) and BPNN to improve prediction accuracy, reduce model time consumption, and ultimately lower the computational cost of the model [10].

2 Methods and materials

To predict supply chain demand, a combined forecasting model including UPM and MPM is designed starting from sales data. To improve the inference speed of the combined prediction model, it is quantitatively optimized and deployed on an embedded AI platform.

2.1 Construction of a univariate prediction model considering BPNN parameter optimization

E-commerce sales data is an important basis for SCDF. To predict supply chain demand, a combined forecasting model is designed considering the characteristics of e-commerce and retail sales data. Moreover, the combined prediction model includes a UPM based on BPNN parameter optimization and an MPM based on CNN-LSTM. To enable high-speed model inference of the research and design combination prediction model, it is deployed on an embedded AI platform. When constructing a UPM, time series data is used and normalized to raise the model's prediction accuracy. The expression of normalization processing is shown in equation (1) [11].

$$\bar{A}_B = \frac{A_B - A_{\min}}{A_{\max} - A_{\min}} \tag{1}$$

In equation (1), $\bar{A}_B$ indicates the normalized data value, and A_B represents the B th original data. $A_{\max}$ and $A_{\min}$ refer to the max and mini values of the original data, respectively. BPNN is a neural network based on error BP algorithm, which has the advantages of strong nonlinear mapping ability, adaptability, generalization ability, and easy implementation. Compared with other traditional time series prediction methods, it has higher prediction accuracy [12-13]. Therefore, the study uses it as the basis for UPMs. However, the initial weights and thresholds of the BPNN can affect the convergence speed and results of the algorithm. Therefore, to better utilize the role of BPNN, an IWOA is designed to optimize its parameters. The advantage of the WOA is its fast convergence speed, few parameters, and ease of adjustment, but it is also prone to getting stuck in local optima [14]. Therefore, the study makes three improvements to it. Improvement one is to introduce Circle mapping to strengthen the diversity of the initial population. Meanwhile, to make the chaotic value distribution of the Circle map more uniform, slight adjustments are made to it, and the adjusted Circle map is shown in equation (2).

$$C_{D+1} = \mathrm{mod}\left(3C_D + 0.4 - \left(\frac{0.5}{3\pi}\sin\left(3\pi \cdot C_D\right), 1\right)\right) \tag{2}$$

In equation (2), C_{D+1} represents the next chaotic variable value after mapping, and C_D is the current chaotic variable value. D means the dimension of the solution, which is the number of problem variables. $\mathrm{mod}(\cdot,1)$ represents modulo operation. The second improvement is to introduce adaptive inertia weight E to dynamically adjust the global and local search capabilities of the algorithm. The expression of E is shown in equation (3).

$$E = \sin\left(\frac{\pi t}{2t_{\max}} + \pi\right) + 1 \tag{3}$$

In equation (3), t denotes the current amount of iterations, and $t_{\max}$ refers to the max amount of iterations. At this point, the position update expression of the WOA is shown in equation (4).

$$\vec{F}(G+1) = \begin{cases} E \cdot \overrightarrow{F^*}(G) - H \cdot I & , J < 0.5 \\ \vec{I} \cdot \mathrm{e}^{JK} \cdot \cos(2\pi l) + E \cdot \overrightarrow{F^*}(G) & , J \geq 0.5 \end{cases} \tag{4}$$

In equation (4), $\vec{F}(G+1)$ represents the position update vector at time $G+1$, and $\overrightarrow{F^*}(G)$ is the position vector of the optimal whale individual at time G. H represents the coefficient vector, and J represents the random probability. K is a random vector, e^{JK} is a spiral equation. $\vec{I}$ represents the distance between the current whale and its prey, while I denotes the distance vector between the individual whale and its optimal position. The third improvement is the introduction of Levy flight

strategy to balance local search and global search, thereby avoiding falling into local optima. The probability density function $Levy(L)$ expression of Levy's flight strategy is shown in equation (5) [15].

$$Levy(L)=\frac{1}{\pi}\int_0^{+\infty}\left(-M\mid N\mid^P\right)\cos(NL)dN \tag{5}$$

In equation (5), L represents the step size variable, and L is the shape control parameter. N and P are integral variables and decay rate control parameters, respectively. dN is the derivative of N. The expression of Levy's flight random step size $Levy(M)$ is shown in equation (6).

$$Levy(M)=L=\frac{Q}{\mid R\mid^{1/M}} \tag{6}$$

In equation (6), both R and Q represent random variables. After updating the position in WOA, the study uses Levy flight to update the position again, and the expression is denoted in equation (7).

$$\vec{F}(G+1)=F(G)+S\oplus \text{Levy}(M) \tag{7}$$

In equation (7), $\oplus$ represents dot product operation. S is the step size factor. The main process of the IWOA is denoted in Figure 1.

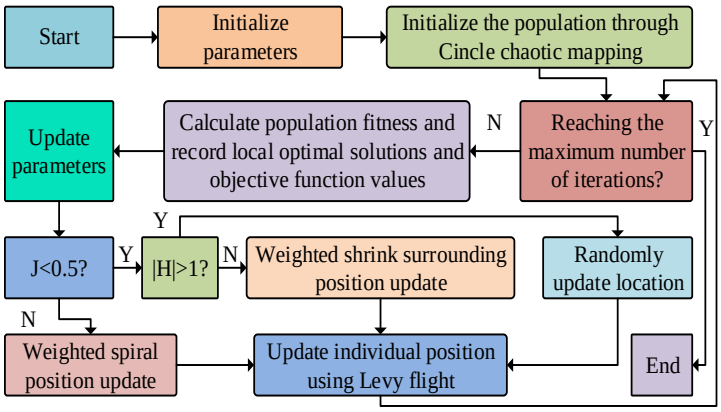


Figure 1: The main process of IWOA

From Figure 1, the main process of the IWOA includes initializing the population through Circle chaotic mapping, updating individual positions using Levy flight, determining whether the max amount of iterations has been reached, and updating parameters. The employment of the IWOA facilitates the optimization of the initial weights and thresholds of the BPNN, thereby enhancing its predictive efficacy. The IWOA-BPNN's main process is denoted in Figure 2.

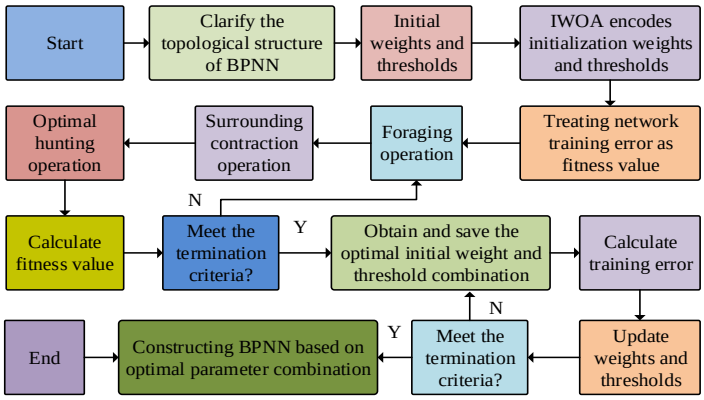


Figure 2: The main process of IWOA-BPNN

From Figure 2, the main process of IWOA-BPNN includes clarifying the topology structure of BPNN, encoding initial weights and thresholds, calculating fitness, obtaining and saving the optimal initial weight and threshold combination, and constructing BPNN based on the optimal parameter combination. After constructing a UPM based on optimal initial weights and thresholds using BPNN, the supply chain needs can be predicted based on normalized time series data. The pseudo-code of the IWOA-BPNN is shown in Figure 3.

```
Initialize BPNN structure and parameters.
Generate initial population via Circle mapping.
Evaluate fitness of each individual.
for each iteration to MaxIter do
    for each individual in population do
        if condition met then
            Update position using Levy flight.
        else
            Update position based on fitness.
        end if
    end for
    Update global optimal individual.
end for
Train BPNN with optimal weights and thresholds from IWOA.
Return trained BPNN model.
```

Figure 3: Pseudo-code of IWOA-BPNN

From Figure 3, the pseudo-code of the IWOA-BPNN clearly presents the core process of the IWOA-BPNN. First, the BPNN structure and parameters are initialized, and then the initial population is generated by Circle mapping and its fitness is evaluated. Subsequently, it enters the iteration loop, where in each iteration, based on conditional judgment, the individual position is updated using Levy flight or fitness based methods, and the global optimal individual is updated. Finally, the weights and thresholds optimized by IWOA are used to train BPNN, resulting in a well-trained model.

2.2 Construction of multivariate prediction model considering CNN-LSTM model

To predict supply chain demand, a UPM is designed for the combination forecasting model. Due to the non-stationary nature of e-commerce sales data and the presence of multiple variables (such as advertising investment, promotion strategies, etc.) that affect sales data, single variable prediction may not be accurate enough. To improve the accuracy of SCDF, an MPM based on CNN-LSTM model is studied and designed. Before constructing the prediction model, data preprocessing is necessary. In terms of missing values, the study uses the median to fill in. On outliers, the study directly removes them. To improve the processing speed of the model, the data is normalized and utilized as input for the MPM. Due to the non-stationary nature of sales value time series data, the study also uses natural logarithm to transform the data, to remove its non stationarity and facilitate subsequent feature acquisition. In addition, the study applies Pearson correlation coefficient to analyze the correlation between different features. After processing the data, the study constructs a CNN-LSTM prediction model, and the model's main process is denoted in Figure 4.

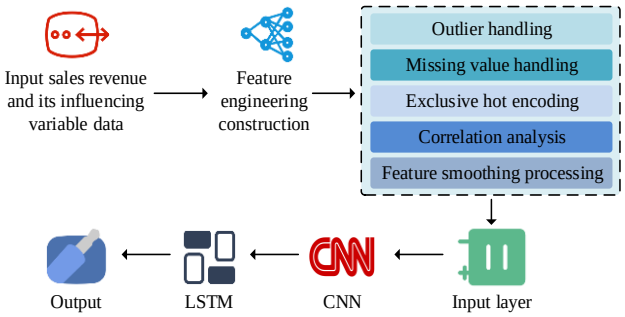


Figure 4: The main process of CNN-LSTM prediction model

From Figure 4, the CNN-LSTM prediction model requires input of sales revenue and its influencing variable data, and feature engineering construction of this data, including outlier handling, correlation analysis, and feature smoothing. The processed features will enter the input layer and obtain data features through the CNN layer. Afterwards, sequence prediction is achieved through an LSTM layer, and prediction data is output through a fully connected layer. CNN is a deep neural network mainly composed of convolutional layers, pooling layers, activation functions, and fully connected layers. It can not only automatically learn local features in input data, but also achieve multi-layer feature extraction [16-17]. In CNN, causal padding is used to ensure that only the input data of the current and previous time steps are used when calculating the output of the current time step, without considering the data of future time steps, to maintain the causal relationship of the time series. In CNN, the activation function used in the study is Sigmoid, which is expressed as equation (8).

$$Sigmoid = \frac{1}{1+e^{U}} \tag{8}$$

In equation (8), U represents the input value of the function. LSTM is a special type of recurrent neural network that can not only capture the dependencies of long time intervals in time series, but also dynamically adjust the flow and storage of information. LSTM is made up of a forgetting gate, an inputting gate, and an outputting gate, and it updates the feature information of the hidden layer through the forget gate [18-19]. In LSTM, the expression of the hidden layer V step state W_V is shown in equation (9).

$$W_V = f\left(XY_V + ZW_{V-1}\right) \tag{9}$$

In equation (9), $f(\cdot)$ represents the nonlinear activation function. X means the connection matrix of the input layer. Y_V represents the input of step Y_V. Z is the weight matrix from the previous hidden layer to the next hidden layer. W_{V-1} represents the state of the hidden layer at the previous moment. The expression of the output O_V at step V is shown in equation (10).

$$O_V = a\left(b * W_V\right) \tag{10}$$

In equation (10), a represents the activation function, and b means the weight matrix of the output layer. However, the selection of parameters such as

learning rate and hidden layers in CNN-LSTM models can affect the performance of the model. To find the best parameters for the CNN-LSTM model, an Improved Particle Swarm Optimization (IPSO) algorithm is created. PSO algorithm is a swarm intelligence optimization algorithm with advantages such as simplicity and ease of implementation. It is widely used in engineering optimization and machine learning parameter tuning fields [20]. However, the performance of PSO algorithm relies heavily on the initial particle swarm quality and is prone to getting stuck in local optima. Therefore, the study makes four improvements to the PSO algorithm.

Improvement one is to combine Sobol sequence and K-means clustering to make the initial population distribution of PSO more uniform. The Sobol sequence is a low-variance sequence. This means that it can generate sample points that are spread evenly across a high number of dimensions. The study uses a Sobol sequence generator to create a sample pool, evaluates and sorts the fitness of the samples, and then analyzes the sorted samples using K-means clustering. The second improvement is to introduce non-linear inertia weight d to balance the exploration and development capabilities of the algorithm in different periods. The expression of d is shown in equation (11).

$$d = d_{\max} - \frac{\left(d_{\max} - d_{\min}\right)}{1 + \exp\left(\frac{10t}{t_{\max}} - 5\right)} \tag{11}$$

In equation (11), $d_{\max}$ and $d_{\min}$ represent the max and mini-inertia weight values, respectively. Improvement three is to introduce an inertia weight direction adaptive change strategy to balance the global and local search capabilities of PSO and overcome the problem of local optima. The key to this strategy is the inertia direction change coefficient g_h, which is expressed as equation (12).

$$g_h = \begin{cases} 1 & j\left(m_h^t\right) > j\left(m_h^{t-1}\right) \\ -rand & j\left(m_h^t\right) > j\left(m_h^{t-1}\right) \end{cases} \tag{12}$$

In equation (12), h represents the particle number, $rand$ represents the random number within the range, $j\left(m_h^t\right)$ and $j\left(m_h^{t-1}\right)$ represent the fitness values of particle h at the t th and $t-1$th iterations, respectively. Under this strategy, the update of particle velocity is denoted in equation (13).

$$v_{hq}^t = g_h \cdot dv_{hq}^{t-1} + x_1 y_1 \left(\beta_{best-hq}^{t-1} - m_{hq}^{t-1}\right) + x_2 y_2 \left(\varphi_{best-q}^{t-1} - m_{hq}^{t-1}\right) \tag{13}$$

In equation (13), v_{hq}^t and v_{hq}^{t-1} denote the velocities of the h th particle in the q th dimension at the t th and $t-1$th iterations, respectively, while x_1 and x_2 indicate the individual learning factor and social learning factor, respectively. y_1 and y_2 are both random numbers, and $\beta_{best-hq}^{t-1}$ and $\beta_{best-hq}^{t-1}$ are the individual and global optimal positions in the $t-1$ th iteration, respectively. m_{hq}^{t-1} is the position of the h th particle in the q th dimension at the $t-1$ th iteration. The fourth improvement is the introduction of an adaptive hierarchical learning strategy to enhance the PSO optimization capability. This strategy requires dividing it into three layers based on particle fitness ranking and adopting different learning strategies for each layer. The main process of IPSO algorithm is denoted in Figure 5.

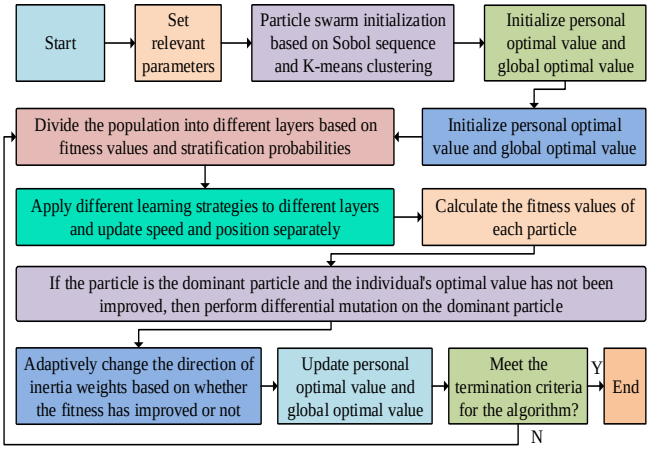


Figure 5: The main process of IPSO algorithm

From Figure 5, the main process of the IPSO algorithm involves particle swarm initialization based on Sobol sequences and K-means clustering, population stratification, particle fitness calculation, and adaptive change of inertia weight direction. Through IPSO, important parameters of CNN-LSTM can be optimized to raise the effectiveness of the prediction model. The main process of IPSO-CNN-LSTM is denoted in Figure 6.

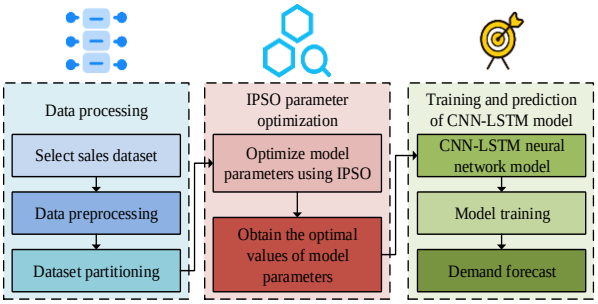


Figure 6: The main process of IPSO-CNN-LSTM

From Figure 6, the IPSO-CNN-LSTM prediction model is mainly divided into three parts: data processing, IPSO parameter optimization, and CNN-LSTM model training and prediction. Based on the prediction model after parameter optimization, the study is able to provide better prediction of supply chain demand. The pseudo-code of the IPSO-CNN-LSTM prediction model is shown in Figure 7.

```
Initialize BPNN structure and parameters.
Generate initial population via Circle mapping.
Evaluate fitness of each individual.
for each iteration to MaxIter do
  for each individual in population do
    if condition met then
      Update position using Levy flight.
    else
      Update position based on fitness.
    end if
  end for
  Update global optimal individual.
end for
Train BPNN with optimal weights and thresholds from
IWOA.
Return trained BPNN model.
```

Figure 7: Pseudo-code for IPSO-CNN-LSTM prediction model

From Figure 7, the pseudo-code fully demonstrates the key process of the IPSO-CNN-LSTM prediction model. First, the structure of the CNN-LSTM model is initialized, and then the initial population is generated and evaluated for fitness based on Sobol sequences and K-means clustering. During the iteration process, adaptive strategies are used to update individual speed and position, continuously adjusting the global optimal individual. Finally, the IPSO optimized parameters are used to train the CNN-LSTM model, resulting in an accurate prediction model.

A combination forecasting model for SCDF is designed. To improve the portability of the combined prediction model and reduce its cost and power consumption, an embedded design is studied. In the training of the model, two methods are adopted: model quantification and algorithm optimization to obtain a model that is more suitable for embedded AI platforms. In addition, the study selects NVIDIA Jetson TX2 as the embedded AI platform for deploying the combined prediction model. The NVIDIA Jetson TX2 has powerful computing and AI inference capabilities and low hardware costs. In the model quantization optimization, the NVIDIA inference acceleration scheme TensorRT is adopted. The TensorRT optimization process is shown in Figure 8.

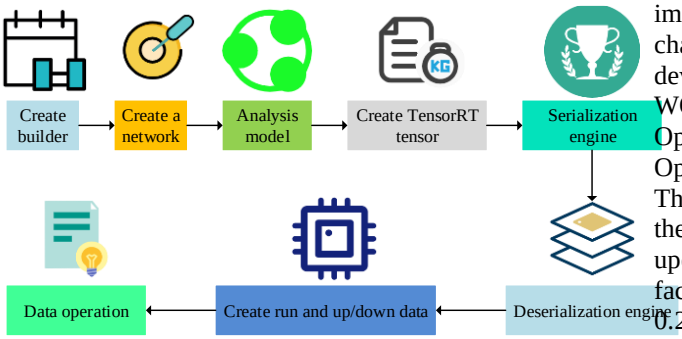


Figure 8: The optimization process of TensorRT

From Figure 8, the TensorRT optimization process mainly includes creating builders, creating TensorRT tensors, and creating runs and upper and lower data. In addition, TensorRT can also perform memory management on models to optimize their structure, making them better suited for embedded AI platforms and improving inference speed. The embedded design of the supply chain demand combination prediction model can improve the prediction speed and efficiency of the model.

3 Results

To assess the effect of the research-designed prediction model, the experimental environment and apparatus were described, and the data used in the experiment were set. In addition, the study also selected comparative models and analyzed them based on indicators such as prediction accuracy and error.

3.1 Performance verification of parameter optimization prediction model based on BPNN

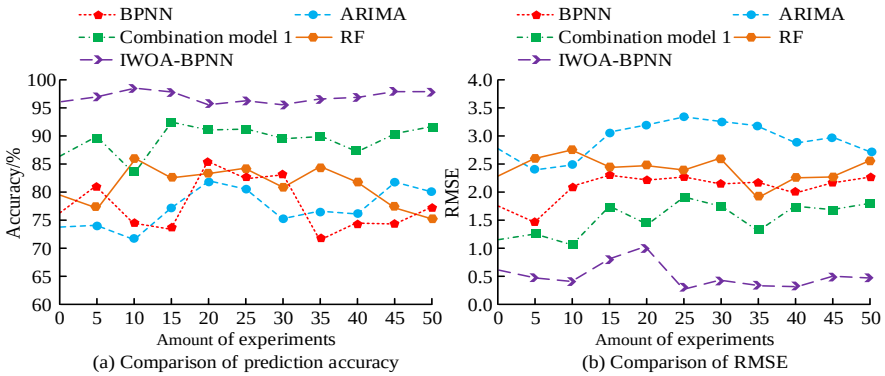
To validate the performance of a single factor prediction model based on BPNN parameter optimization, the study selected the monthly demand (sales data) of learning supplies in an e-commerce mall from January 2020 to December 2024 for empirical analysis, and preprocessed the obtained data. "Learning supplies" included various stationery items such as paper, pens, and notebooks, as well as commonly used learning tools for students such as calculators and folders. In addition, the study selected two test functions from the CEC2013 test function set to validate the performance of the IWOA, namely the unimodal function $f4$ and the multimodal function $f18$ [21]. The CEC2013 test function set is a benchmark function set widely used for optimizing algorithm performance evaluation, including 28 different types of functions, including unimodal functions, multimodal functions, and combination functions, used to simulate various complex optimization problems. Among them, unimodal functions such as spherical functions are used to test algorithm convergence speed, multimodal functions such as cross benchmark functions are used to evaluate algorithm exploration ability, and combination functions further increase complexity through translation, rotation, and other operations. It is widely used in research fields such as intelligent optimization algorithms, large-scale optimization problem solving, and algorithm improvement innovation, providing a standardized and challenging testing platform for the research and development of optimization algorithms. Meanwhile, WOA, Sparrow Search Algorithm (SSA), Seagull Optimization Algorithm (SOA), and Grey Wolf Optimization (GWO) were also selected for comparison. The population size of these comparative models was 40, the maximum iteration number was 400, and the spiral update parameter of WOA was 1.5, with a convergence factor of 2. In addition, the discoverer ratio of SSA was 0.2, and the number of seagulls and inertia weight of SOA were 40 and 0.5, respectively. The operating system used in the study was Windows 10, and the central processing unit was Intel Core i5-13500. In terms of prediction models, the study selected BPNN, Auto-Regression Integrated Moving Average (ARIMA), Random Forest (RF), WOA-BPNN, and a combination model 1 that combines Support Vector Machine (SVM) regression and LSTM for comparison. The population size of the IWOA was 40, the max amount of iterations was 400, and the

learning rate of the BPNN was 0.02. The comparison of the average values of different algorithms under different test functions is denoted in Table 1.

Table 1: Comparison of average values of different algorithms under different test functions

Algorithm	Unimodal function *f*4						
	Amount of experiments						
	1	2	3	4	5	6	7
SSA	3.71×10^3	2.38×10^3	2.98×10^3	2.81×10^3	3.10×10^3	3.10×10^3	2.59×10^3
SOA	1.96×10^3	2.07×10^3	2.30×10^3	2.55×10^3	1.98×10^3	1.88×10^3	2.62×10^3
GWO	1.63×10^3	2.05×10^3	1.44×10^3	1.59×10^3	1.52×10^3	1.81×10^3	1.91×10^3
WOA	1.41×10^3	1.37×10^3	1.33×10^3	1.26×10^3	1.43×10^3	1.34×10^3	1.42×10^3
IWOA (manuscript)	0.22×10^3	0.27×10^3	0.30×10^3	0.33×10^3	0.25×10^3	0.21×10^3	0.26×10^3
Algorithm	Multimodal *f*18						
	Amount of experiments						
	1	2	3	4	5	6	7
SSA	2.43×10^2	2.51×10^2	1.94×10^2	2.44×10^2	2.35×10^2	2.29×10^2	1.77×10^2
SOA	2.56×10^2	1.18×10^2	2.37×10^2	1.86×10^2	1.87×10^2	1.77×10^2	1.74×10^2
GWO	1.56×10^2	1.90×10^2	1.16×10^2	1.38×10^2	1.92×10^2	1.53×10^2	2.19×10^2
WOA	1.46×10^2	1.26×10^2	1.37×10^2	1.54×10^2	1.20×10^2	1.01×10^2	1.10×10^2
IWOA (manuscript)	0.32×10^2	0.28×10^2	0.34×10^2	0.26×10^2	0.29×10^2	0.28×10^2	0.25×10^2

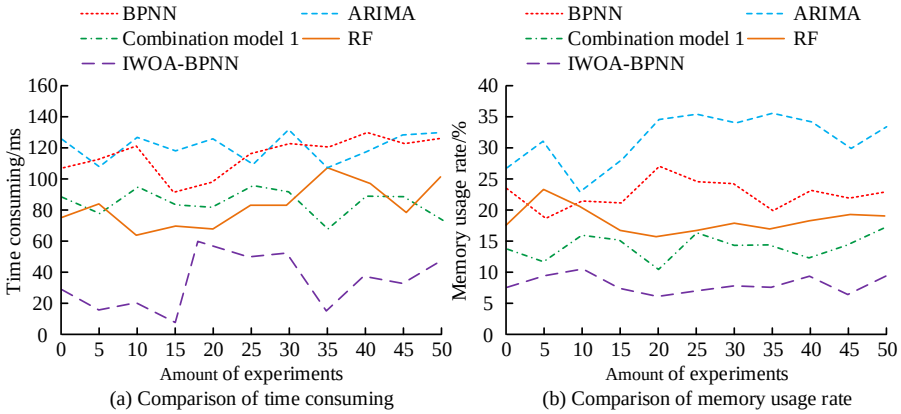
From Table 1, under the unimodal function *f*4 and multimodal function *f*18, the average values of the IWOA were significantly lower than those of the comparison algorithm. For example, under the unimodal function *f*4, the average values of the IWOA were all below 1.0×10^3, while the average values of SSA, SOA, GWO, and WOA were significantly greater than 1.0×10^3. This indicates that the IWOA has better optimization ability and stability under both test functions, and can better optimize the initial weights and thresholds of the BPNN. This may be because the IWOA introduces Circle chaotic mapping, adaptive inertia weight, and Levy flight strategy, which improves the algorithm's global search ability, local optimization ability, and ability to escape from local optima, thereby achieving better results in the optimization of unimodal and multimodal functions. The comparison of prediction accuracy and Root Mean Square Error (RMSE) of different UPMs is shown in Figure 9.



(a) Comparison of prediction accuracy (b) Comparison of RMSE

Figure 9: Comparison of prediction accuracy and RMSE of different univariate prediction models

From Figure 9 (a), in terms of comparison of prediction accuracy, the max value of the IWOA-BPNN model was 98.05%, and the mini value was 95.27%. The maximum prediction accuracy of BPNN, ARIMA, RF, and combination model 1 were 85.14%, 82.09%, 85.37%, and 92.68%, respectively, all of which were less than 98.05%. From Figure 9 (b), in the comparison of RMSE, the IWOA-BPNN model designed in the study had the smaller error, followed by the combination model 1. In terms of specific values, the maximum RMSE of the IWOA-BPNN model was 1.03 and the minimum was 0.25, which was much lower than the comparison model. This indicates that the IWOA-BPNN model designed for research can make more accurate predictions of supply chain demand. The comparison of time consumption and memory usage of different UPMs is shown in Figure 10.



(a) Comparison of time consuming (b) Comparison of memory usage rate

Figure 10: Comparison of time consumption and memory usage of different univariate prediction models

As shown in Figure 10 (a), in terms of time consumption comparison, the max values of the IWOA-BPNN model and the four comparison models were 61ms, 129ms, 132ms, 108ms, and 95ms, respectively. The time consumption of the IWOA-BPNN model was greatly lower than that of the comparison model. From Figure 10 (b), the IWOA-BPNN model performed better in terms of memory usage, with a max value of 10.85% and a mini value of 6.03%. Overall, the IWOA-BPNN model has lower time and memory consumption and better performance. To further validate the performance of the IWOA-BPNN model, the study selected Mean Absolute Error (MAE) and inference delay metrics. The comparison of MAE and inference delay for different models is shown in Table 2.

Table 2: Comparison of MAE and inference delay among different models

Algorithm	MAE					Inference delay/ms				
	Number of experiments					Number of experiments				
	1	2	3	4	5	1	2	3	4	5
BPNN	2.440	1.746	1.894	2.362	1.723	13.75	15.80	16.51	16.27	13.30
ARIMA	3.201	3.188	3.483	3.089	2.986	20.79	20.76	20.29	18.30	16.27
RF	2.236	2.080	2.017	2.783	2.104	14.67	16.43	18.29	16.83	15.45
Combination model 1	1.875	1.155	1.212	1.173	1.740	9.06	8.15	7.41	11.66	8.35
IWOA-BPNN	0.307	0.681	0.914	0.359	0.799	2.82	4.02	3.10	4.53	4.34

From Table 2, the proposed IWOA-BPNN model outperformed other compared algorithms in terms of MAE and inference delay. Specifically, the average MAE of the IWOA-BP neural network model was 0.612, while the average MAE of the BPNN model, ARIMA, RF, and combination model 1 were 2.033, 3.189, 2.244, and 1.431, respectively. The MAE of the IWOA-BPNN model was the smallest, indicating that its prediction error was the smallest and its stability was high, which could more accurately predict supply chain demand. In addition, in terms of inference delay, the IWOA-BPNN model had significantly lower values than other models, indicating that its inference efficiency was higher and could meet the high real-time requirements of supply chain prediction scenarios.

3.2 Performance validation of multivariate prediction model considering CNN-LSTM

To validate the effect of the MPM based on CNN-LSTM, the same operating system and central processing unit were used, and the NVIDIA Jetson TX2 embedded AI platform was configured. The study selected the sales volume of daily necessities in a certain shopping mall from January 2020 to December 2024 for analysis, preprocessed it, and analyzed the correlation coefficients between features. "Daily necessities" encompassed essential items for daily life, such as food, cleaning supplies, and personal care products. Since these correlation coefficients were not high, there was no need to further reduce their dimensionality. In addition, advertising investment was measured by the total budget of advertising activities, in units of yuan, and was obtained by summarizing the amount of advertising invested by enterprises in different advertising channels (such as television, internet, newspapers, etc.). The promotional strategy quantified the intensity of promotional activities, using discount intensity as an example, expressed as the ratio of the discounted price to the original price. Other influencing variables such as social media popularity were quantified by the search volume of relevant keywords or topic popularity, while market trends were quantified by referring to trend indicators in market research reports or industry analysis articles. In the validation of the IPSO algorithm, the same testing function was used, and PSO, Genetic Algorithm (GA), WOA, and a combination model 2 combining GA and BPNN were selected for comparison. In terms of prediction models, CNN-LSTM, SVM, PSO-CNN-LSTM, and a combination model 3 combining PSO and extreme gradient boosting were selected for comparison. The optimizer of the model was Adam, the learning factor of PSO was 1.5, the particle swarm size was 40, and the dimension was 30. To improve the transparency and practicality of the model, the SHapley Additive exPlans (SHAP) technology was introduced in

the study. SHAP values can quantify the contribution of each feature to the prediction results, helping to identify the factors that have the greatest impact on demand forecasting. The SHAP values of different variables are shown in Table 3.

Table 3: SHAP values of different variables

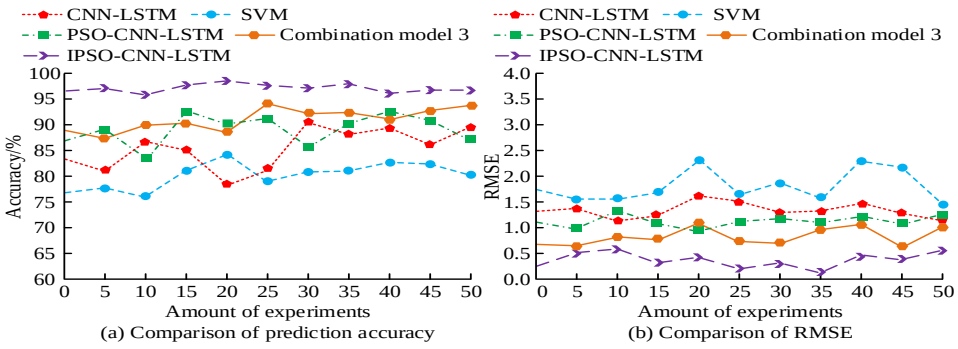
Characteristic variable	SHAP value	Characteristic variable	SHAP value
Sales volume	0.45	Store sales personnel	0.08
WeChat push frequency	0.15	Store poster display	0.05
Local TV advertising investment	0.25	Store promotion events_specific	0.30
Online advertising investment	0.20	\	\

From Table 3, sales volume had the greatest contribution to SCDF, with an SHAP value as high as 0.45. The store promotion event was followed by the special event, with an SHAP value of 0.30. These two are key factors that affect demand forecasting. The positive impact of local TV advertising investment (0.25) and online advertising investment (0.20) on demand should not be underestimated. In contrast, although WeChat push frequency (0.15), store sales personnel (0.08), and store poster display (0.05) had some impact, they were relatively small. If enterprises want to improve prediction accuracy, they can focus on data collection and analysis in sales, store promotions, and advertising investment. The standard deviation comparison of different algorithms under different test functions is shown in Table 4.

Table 4: Comparison of standard deviations of different algorithms under different test functions

Algorithm	Unimodal function *f*4						
	Amount of experiments						
	1	2	3	4	5	6	7
PSO	3.96×10^{3}	3.82×10^{3}	3.99×10^{3}	3.94×10^{3}	3.72×10^{3}	3.61×10^{3}	3.59×10^{3}
GA	2.50×10^{3}	3.27×10^{3}	2.95×10^{3}	2.92×10^{3}	2.87×10^{3}	2.74×10^{3}	2.76×10^{3}
WOA	2.76×10^{3}	2.25×10^{3}	2.37×10^{3}	2.95×10^{3}	2.35×10^{3}	2.01×10^{3}	2.81×10^{3}
Combination model 2	1.47×10^{3}	1.92×10^{3}	1.79×10^{3}	1.16×10^{3}	1.97×10^{3}	1.53×10^{3}	1.10×10^{3}
IPSO (manuscript)	0.85×10^{3}	0.82×10^{3}	0.80×10^{3}	0.77×10^{3}	0.81×10^{3}	0.80×10^{3}	0.76×10^{3}
Algorithm	Multimodal *f*18						
	Amount of experiments						
	1	2	3	4	5	6	7
PSO	5.72×10^{1}	6.45×10^{1}	5.55×10^{1}	5.86×10^{1}	5.63×10^{1}	5.93×10^{1}	6.08×10^{1}
GA	4.57×10^{1}	4.61×10^{1}	4.52×10^{1}	4.88×10^{1}	4.60×10^{1}	5.38×10^{1}	4.51×10^{1}
WOA	3.79×10^{1}	4.06×10^{1}	4.12×10^{1}	3.91×10^{1}	3.98×10^{1}	4.47×10^{1}	4.19×10^{1}
Combination model 2	2.25×10^{1}	2.22×10^{1}	1.59×10^{1}	1.98×10^{1}	1.93×10^{1}	1.60×10^{1}	1.71×10^{1}
IPSO (manuscript)	0.14×10^{1}	0.15×10^{1}	0.12×10^{1}	0.14×10^{1}	0.13×10^{1}	0.16×10^{1}	0.11×10^{1}

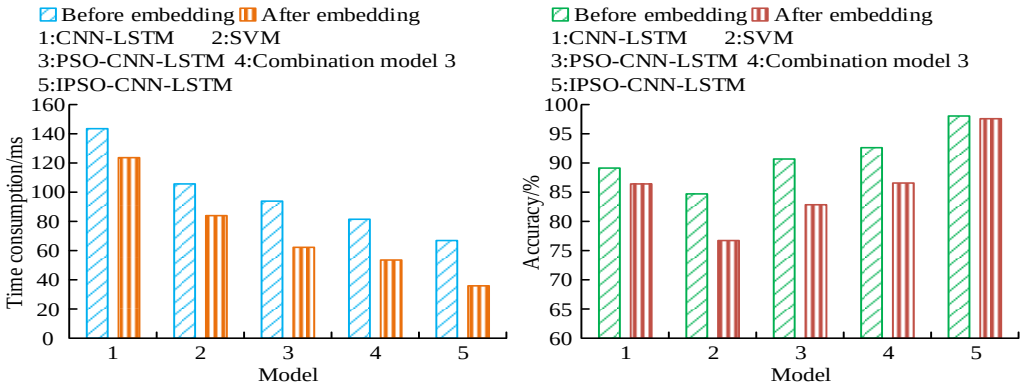
From Table 4, the standard deviation of IPSO algorithm was significantly lower than that of the comparison algorithm under unimodal *f*4 and multimodal *f*18. For example, under the multimodal function *f*18, the standard deviation values of the IPSO algorithm were all below 1.0×10^{1}, while the comparison algorithms were all above 1.0×10^{1}. This indicates that the optimization results of IPSO algorithm have higher stability and reliability. This may be because the IPSO algorithm has improved the initialization strategy by introducing nonlinear inertia weights and inertia weight direction adaptive change strategies. The accuracy and RMSE comparison of different MPMs are shown in Figure 11.



(a) Comparison of prediction accuracy (b) Comparison of RMSE

Figure 11: Comparison of accuracy and RMSE of different multivariate prediction models

According to Figure 11 (a), the maximum prediction accuracy of the IPSO-CNN-LSTM model was 98.51%, and the minimum was 96.02%. The maximum prediction accuracies of CNN-LSTM, SVM, PSO-CNN-LSTM, and combination model 3 were 90.55%, 84.25%, 92.75%, and 94.69%, respectively, all of which were less than 98.51%. According to Figure 11 (b), the IPSO-CNN-LSTM model had a smaller RMSE, with maximum RMSE values of 0.58, 1.59, 2.38, 1.37, and 1.15 compared to the four comparison methods. The IPSO-CNN-LSTM model performed better. The comparison of time consumption and accuracy before and after embedding the IPSO-CNN-LSTM model is shown in Figure 12.



(a) Comparison of time consumption before and after embedding

(b) Comparison of accuracy before and after embedding

Figure 12: Comparison of time consumption and accuracy before and after embedding IPSO-CNN-LSTM model

According to Figure 12 (a), in terms of time consumption, the maximum values before and after embedding the IPSO-CNN-LSTM model were 67ms and 35ms, respectively, with an improvement ratio of 47.76% between the two. According to Figure 12 (b), the maximum difference in accuracy between the IPSO-CNN-LSTM model before and after embedding was 0.18%. Deploying the combination prediction model on an embedded AI platform could reduce the model's time consumption with only a 0.18% decrease in accuracy.

4 Discussion and conclusion

A combination prediction model was designed to improve the accuracy of SCDF, which includes an IWOA-BPNN UPM and an IPSO-CNN-LSTM MPM. The findings denoted that under the unimodal function *f*4, the average values of the IWOA were all below 1.0×10^3, significantly lower than the comparison algorithms. Moreocver, the average value of the IWOA under the multimodal function *f*18 was also smaller. This indicated that the IWOA had better optimization ability and stability, and could better optimize the initial weights and thresholds of the BPNN. The max prediction accuracy of the UPM was 98.05%, which was 12.91%, 15.96%, 12.68%, and 5.37% higher than the max values of the comparison model, respectively. This indicates that the model can make more accurate predictions of supply chain demand, and this may be due to its use of parameter combinations optimized by the IWOA. The standard deviation of IPSO algorithm under unimodal *f*4 and multimodal *f*18 was significantly lower than that of the comparison algorithm. The max prediction accuracy of the IPSO-CNN-LSTM model was 98.51%, which was 7.96%, 14.26%, 5.76%, and 3.82% higher than the maximum values of CNN-LSTM, SVM, PSO-CNN-LSTM, and combination model 3, respectively. After deploying the combination prediction model on the embedded AI platform, its time consumption increased by a maximum of 47.76% without significantly affecting the model accuracy. Overall, the designed combination prediction model has good performance. However, the IPSO algorithm designed in this study shows a decline in performance when facing complex multimodal problems. Future research can further optimize it to enhance its ability to overcome local optima.

Funding information

This work was sponsored in part by "14th Five-Year Plan" Key First-Tier Discipline of Business Administration in Jiangsu Province (SJYH2022-2/285).

References

[1] Ducharme C, Agard B, Trépanier M. Improving demand forecasting for customers with missing downstream data in intermittent demand supply chains with supervised multivariate clustering. Journal of Forecasting, 2024, 43(5):1661-1681. https://doi.org/10.1002/for.3095

[2] Karthick B, Uthayakumar R. An optimal strategy for forecasting demand in a three-echelon supply chain system via metaheuristic optimization. Soft Computing, 2023, 27(16):11431-11450. https://doi.org/10.1007/s00500-023-08290-x

[3] Bai B. Acquiring supply chain agility through information technology capability: the role of demand forecasting in retail industry. Kybernetes, 2023, 52(10):4712-4730. https://doi.org/10.1108/K-09-2021-0853

[4] Subramanian B, Mishra A, Bharathi V R, Mandala G, Kathamuthu N D, Srithar S. Big data and fuzzy logic

for demand forecasting in supply chain management: A data-driven approach. Journal of Fuzzy Extension and Applications, 2025, 6(2):260-283. https://doi.org/10.22105/jfea.2025.488816.1703

[5] Mbonyinshuti F, Nkurunziza J, Niyobuhungiro J, Kayitare E. Health supply chain forecasting: a comparison of ARIMA and LSTM time series models for demand prediction of medicines. Acta Logistica, 2024, 11(2):269-280. https://doi.org/10.22306/al.v11i2.510

[6] Liu R, Vakharia V. Optimizing supply chain management through BO-CNN-LSTM for demand forecasting and inventory management. Journal of Organizational and End User Computing (JOEUC), 2024, 36(1):1-25. https://doi.org/10.4018/JOEUC.335591

[7] Saad N H M. A study of deep learning-based algorithms for supply chain logistics demand forecasting. Edelweiss Applied Science and Technology, 2025, 9(3):1640-1654. https://doi.org/10.55214/25768484.v9i3.5650

[8] Zhang C, Zhang H, Pu T, Pan J. Supply Chain Demand Forecasting Based on Data Mining Algorithm and Seq2Seq. International Journal of Control, Automation and Systems, 2025, 23(1):89-104. https://doi.org/10.1007/s12555-024-0141-8

[9] Hasan M D R, Islam M R, Rahman M A. Developing and implementing AI-driven models for demand forecasting in US supply chains: A comprehensive approach to enhancing predictive accuracy. Edelweiss applied science and technology, 2025, 9(1):1045-1068. https://doi.org/10.55214/25768484.v9i1.4308

[10] Liu Y, Guo X. Research on Influence Factors of Cooperatives Performance in the Agricultural Products Supply Chain Based on System Dynamics Malaysian E Commerce Journal. 2017; 1(2): 01-03. http//doi.org/10.26480/mecj.02.2017.01.03

[11] Newton M, Gravagne I, Jack D. Normalization and processing of rotational eddy current scans for layup characterization of CFRP laminates. Research in Nondestructive Evaluation, 2024, 35(6):337-355. https://doi.org/10.1080/09349847.2024.2406239

[12] Pal S, Roy A, Palaiahnakote S, Pal U. Adapting a swin transformer for license plate number and text detection in drone images. Artificial Intelligence and Applications. 2023, 1(3):129-138. https://doi.org/10.47852/bonviewAIA3202549

[13] Wang J, Li Y H, Wang D, Chai M. A reliability calculation method based on ISSA-BP neural network. International Journal of Structural Integrity, 2024, 15(6):1249-1267. https://doi.org/10.1108/IJSI-07-2024-0104

[14] Li L, Yang Y, Zhou T, Wang M. Data-Driven Combination-Interval Prediction for Landslide Displacement Based on Copula and VMD-WOA-KELM Method. Journal of Earth Science, 2025, 36(1):291-306. https://doi.org/10.1007/s12583-021-1555-3

[15] Vasundra S, Venkatesh D, Bella H K. Levy Flight Strategy-based Tasmanian Devil's Optimization and Autoencoder Approach for Intrusion Detection in Cloud Computing. International Journal of Intelligent Engineering & Systems, 2025, 18(2):548-561. https://doi.org/10.22266/IJIES2025.0331.40

[16] Liu Y. Remote Sensing Image Scene Classification Based on Convolutional Neural Networks. Informatica, 2025, 49(9):45-54. https://doi.org/10.31449/inf.v49i9.5912.

[17] Sung W T, Kang H W, Hsiao S J. Speech Recognition via CTC-CNN Model. Computers, materials & continua, 2023, 76(9):3833-3858. https://doi.org/10.32604/cmc.2023.040024

[18] Aggarwal D, Banerjee S. Forecasting of S&P 500 ESG index by using CEEMDAN and LSTM approach. Journal of Forecasting, 2025, 44(2):339-355. https://doi.org/10.1002/for.3201

[19] Xu H, Song L, Li Y, Zhang T, Shen J. Comparison of LSTM and QR models for predicting CPUE of albacore tuna in waters near the Cook Islands. Fisheries Science, 2025, 91(2):259-274. https://doi.org/10.1007/s12562-025-01851-z

[20] Hao Q, Huang H. Multi-hop Network Security Strategy Integrating ACO Algorithm and PSO Algorithm. Informatica, 2025, 49(17):81-94. https://doi.org/10.31449/inf.v49i17.7499

[21] Mihăilescu M, Stancu-Dumitru D, Teca A. On a Rayleigh-type quotient involving a variable exponent which depends on test functions. Archiv der Mathematik, 2025, 124(5):557-570. https://doi.org/10.1007/s00013-024-02097-4