Journal of Network Intelligence
Taiwan Ubiquitous Information

Volume **9**, Number **1**, February **2024**

Time Series Prediction Model for Big Data on Improved Bionic Population Intelligence Algorithm

Ling-Mei Kong*

Financial and Monetary Institute
Guangdong Nanhua Vocational College of Industry and Commerce
Guangzhou 510507, P. R. China
amy_go2023@163.com

Yan-Hong Zhang

College of information technology
St. Paul University Philippines
Tuguegarao City, Cagayan 3500, Philippines
zyhldy2010@163.com

*Corresponding author: Ling-Mei Kong
Received July 15, 2023, revised September 19, 2023, accepted December 7, 2023.

Abstract. *As the field of big data continues to evolve quickly, machine learning techniques have gradually been applied to financial time series forecasting models due to the large computational errors and single research content of traditional models. Such methods are model-based and increase the model's ability to accurately forecast data by adjusting hyperparameters. Although some progress has been made in stock stock price prediction and other aspects, the prediction accuracy of a single model can no longer meet the current financial time series requirements. To address this problem, this work proposes a combined forecasting model based on an improved bionic population intelligence algorithm. Firstly, the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) method in the field of signal processing is used to fully extract the fluctuation information implied by the big data time series. Secondly, the decomposed low frequency series, trend terms and processed high frequency series are fed into a support vector regression (SVR) model based on the sparrow search algorithm (SSA), and the prediction results are summed up to achieve the final stock market price forecast. Meanwhile, to address the shortcomings of SSA in the optimization process, chaos mapping, sine cosine search strategy, dynamic change strategy, stochastic backward learning strategy and Corsi perturbation are introduced to boost the capability of SSA's global search as well as its convergence capacity in order to further enhance its forecasting performance. Three evaluation errors, RMSE, MAE and MAPE, are employed in the process of providing a quantitative assessment of the prediction model's accuracy. The experimental results show that the proposed combined model outperforms other comparative models and possesses better forecasting results, moreover, it provides a new direction for financial time series data forecasting.*
Keywords: Big data; time series; sparrow search algorithm; chaos mapping; support vector regression

1. **Introduction.** Big data time series forecasting is an important research endeavour that uses large amounts of time series data and advanced analytical techniques to accurately predict future trends and changes [1,2]. By modelling and analysing complex

time series data, it can help us to gain insight into time-related phenomena and provide valuable predictive information.

In the field of equities, big data time series forecasting has great potential for application [3,4]. Firstly, by analysing historical stock price and trading data, the hidden patterns and trends can be revealed, leading to a better understanding of market behaviour. Secondly, by using big data techniques and machine learning algorithms [5,6], accurate stock price prediction models can be constructed to help investors make more informed decisions. In addition, big data time series forecasting can be applied to risk management, trading strategy optimisation and market monitoring to provide better support and guidance to investors and market participants.

Nowadays, the two most popular types of models in stock market forecasting are mainly machine learning models [7,8] and traditional forecasting models [9,10]. With the continuous improvement and rapid development of big data technology, machine learning techniques are gradually being applied to financial time series forecasting models due to the large computational errors and single research content of traditional models. Such methods are based on models, and the data prediction accuracy of the models is improved by adjusting the parameters [11,12]. Although some progress has been made in stock stock price prediction and other aspects, the prediction accuracy of a single model can no longer meet the current financial time series requirements. With the advancement of AI in the past decade, machine learning methods have been extensively used in various industries due to their powerful data processing and analysis capabilities. In big data time series prediction analysis, as financial time series are noisy [13,14] and non-smooth series, they are dynamically distributed with time. The accuracy obtained by forecasting financial time series using traditional statistical forecasting models is low [15], while machine learning models are more capable of capturing non-linear relationships and identifying structure, and their forecasting accuracy is higher. As a result, numerous scholars have conducted research on financial time series forecasting using machine learning models [16,17].

As a machine learning algorithm, although Support Vector Regression (SVR) [18] is widely used in various fields, such as finance, medicine, engineering, etc., it has good results for predicting problems with continuous variables. However, the presence of a large amount of noisy information in stock data reduces the accuracy of stock market prediction modelling, so separating the noise is also a difficult task in building prediction models. In this context, it is of great research interest to find out how to decompose such time series to build models with noise reduction. In addition, in SVR, appropriate kernel functions and parameters need to be selected, such as kernel function type, kernel function parameters, penalty parameters, etc [19]. The choice of these parameters has a significant impact on the results, but often needs to be tried and adjusted through methods such as experience or cross-validation. This work attempts to use a bionic population intelligence algorithm to search the space adaptively to find the optimal combination of parameters, reducing the workload of manual tuning of parameters.

1.1. **Related Work.** The two most popular types of models in financial time series forecasting include: machine learning and traditional statistical models. Traditional statistical models are mainly predictive models commonly used in statistics and economics, such as Generalized Autoregressive Conditional Heteroscedasticity Model (GARCH) [20], Autoregressive Moving Average Model (ARIMA) [21], and nonlinear models. Wang et al. [22] effectively combined the ARIMA model and GARCH model to achieve accurate PM 2.5 pollution prediction. Akpan et al. [23] proposed an ARMA-GARCH model based on differential information, and obtained higher accuracy of trend judgment and prediction than the single model. However, it is difficult to accurately model the stock

market because traditional time series forecasting models cannot handle the noise and non-smoothness of stocks.

In the past few years, machine learning-based artificial intelligence techniques have been highly sought after for applications in various fields, and the financial sector is no exception. Among the financial time series forecasting models based on machine learning, there are two broad categories, one is forecasting methods based on a single machine learning model, such as artificial neural network (ANN), random forest algorithm, SVR and other models, and the other is the use of hyperparametric optimisation algorithms to enhance the forecasting performance of a single machine learning model. Kristjanpoller and Minutolo [24] proposed an ANN-based method for forecasting stock market data, and the results showed that the ANN has a significant improvement in forecasting accuracy compared to ARMA and ARMA-GARCH models. Guo et al. [25] proposed an SVR-based method for forecasting stock market data and verified its feasibility on several stock market indices.

However, single machine learning models generally suffer from more painstaking parameter selection, so Pradeepkumar and Ravi [26] proposed a Random Forest (RF) algorithm based on the discrete Binary Particle Swarm (PSO) algorithm was tested using stock data from three companies. The results show that the improved RF algorithm has a significant improvement in accuracy compared to the traditional RF algorithm. Similarly, in SVR, it is also necessary to select the appropriate kernel function and parameters, such as kernel function type, kernel function parameters, penalty parameters, etc. Therefore, Liang et al. [27] proposed to use the PSO algorithm to automate the parameter search for SVR. Similar to the PSO algorithm, the Sparrow Search Algorithm (SSA) [28] is a bionic population intelligence optimisation algorithm inspired by the predatory behaviour of sparrows. SSA simulates the behaviour of sparrows when searching for food and optimises the solution to the problem through learning and adaptation.

Signals are usually composed of interference signals and real signals, and in order to reveal the real information contained in the signal, it is necessary to separate the signals by an effective signal decomposition method, so as to obtain the noise and real signals. In 1998, Huang et al. [29] first proposed the Empirical Mode Decomposition (EMD) method, which can decompose an arbitrary signal to form a finite Intrinsic Mode Function (IMF), which can represent the local characteristics of the signal. It is possible for the IMF to reflect the local properties of a signal. Yang and Lin [30] combined EMD with an Extreme Learning Machine (ELM) to form a new combinatorial model and applied it to exchange rates, therefore enhancing the precision of exchange rate forecasting. The Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) method [31] not only solves the problem of modal confounding that exists in EMD, but also reduces the decomposition error and efficiency of EEMD. The results of existing studies show that the combined model with decomposition algorithm and machine learning has lower prediction error.

1.2. **Motivation and contribution.** The above analysis reveals that the hyperparametric optimisation algorithm can improve the prediction performance of a single machine learning model to maximise accuracy. The PSO algorithm requires the velocity and position of the particles to be updated and uses random factors to perform the search. SSA, on the other hand, does not require velocity and position updates due to the specific nature of its design, which reduces the complexity of the algorithm and increases the efficiency of the search. Furthermore, the combined model combining the decomposition algorithm and machine learning has lower prediction errors and lags. Therefore, to further improve

the financial time series forecasting accuracy, this work proposes an improved SSA-based SVR model.

The main innovations and contributions of this work include:

(1) The CEEMDAN method in the field of signal processing is used to perform preliminary decomposition of the original data to fully extract the fluctuation information implied by the time series of large data, so as to improve the prediction accuracy of the subsequent model.

(2) To address the shortcomings of SSA in the optimization process, chaotic mapping, positive cosine search strategy, dynamic change strategy, stochastic backward learning strategy and Corsi perturbation are introduced to enhance the global search and convergence capability of SSA and further improve the prediction performance.

(3) Combining signal decomposition and machine learning models to build a combined forecasting model, which provides a new approach to financial time series nowadays and a new direction to think about non-smooth, non-linear time series data.

2. **Models and basic theory.**

2.1. **Decomposition models.**

EMD was initially applied mainly to signal decomposition in the atmosphere and ocean, but in recent years it has been introduced to the fields of economics and finance. The introduction of the EMD decomposition method provides a superior decomposition method for financial time series decomposition.

The EMD decomposition process is based on the following assumptions:

(1) The presence of at least one extreme value and one minimal value in the signal data sequence;

(2) Time domain characteristics are determined at extreme intervals;

(3) If the data series lacks extreme values but has inflection points, the extreme value points can be represented by one or more derivations and the final result can be obtained by integration.

The decomposition process of the EMD, i.e. the extraction of the IMF, is as follows [32]:

(1) Mark the local extremes;

(2) The upper envelope is formed by connecting the extreme value points through cubic spline interpolation and the extreme value points through the lower envelope;

(3) Calculate the upper and lower envelopes' mean value m;

(4) Deduct the upper and lower envelopes' mean value from the input signal $X(t)$:

$$X(t) - m_1 = h_1 \tag{1}$$

One iteration of the above process does not guarantee that h_1 is an eigenmode function and the above process needs to be repeated until h_1 is an eigenmode function.

The stopping criterion determines the number of screening processes performed to produce an eigenmode function, and the stopping criterion can be expressed as

$$SD = \sum_{t=0}^{T} \frac{|h_{k-1}(t) - h_k(t)|^2}{h_{k-1}^2(t)} \tag{2}$$

where SD denotes the standard deviation of the screening results, T is the length of the time series and $h_k(t)$ denotes the kth difference signal. When the value of SD is less than the given threshold, the screening process stops and $h_k(t)$ is considered to satisfy the second condition of the eigenmode function.

Except for the first sequence, the same steps will be performed for the remaining sequences. First a new sequence $r_i(t) = X(t) - c_i(t)$ needs to be constructed. Next, the

above procedure is repeated for the new sequence to obtain the rest of the sequence. Finally, the decomposition is stopped when $r_n(t)$ is a constant or a monotonic function.

In this case the final $r_n(t)$ is called the residual sequence. Therefore, the input signal $X(t)$ can be written in combinatorial form.

$$X(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \tag{3}$$

For example, suppose the input signal $X(t)$ consists of an amplitude-modulated component, a sine component and a cosine component.

$$\begin{cases} x_1(t) = 2\sin(60\pi t) \times (1 + 0.5\sin(2\pi t)) \\ x_2(t) = \sin(120\pi t) \\ x_3(t) = 0.5\cos(10\pi t) \\ x(t) = x_1(t) + x_2(t) + x_3(t) \end{cases} \tag{4}$$

The result of the decomposition of the signal $X(t)$ using EMD is shown in Figure 1.

2.2. **Support vector regression models.** Most multiple linear regression models suffer from slow convergence, low accuracy and weak non-linear fitting ability. Through the study, it is found that SVR are better than BP neural network, genetic algorithm, RBF neural network, etc.

When the sample is non-linear, a non-linear regression analysis is used to map the sample data into a higher dimensional space.

$$f(x) = \omega \cdot \varphi(x) + b \tag{5}$$

where $\phi(\mathbf{x})$ denotes the regression function, b denotes the bias and ω denotes the weight. The optimal problem for SVR is shown below:

$$\min_{\omega,b,\xi} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{6}$$

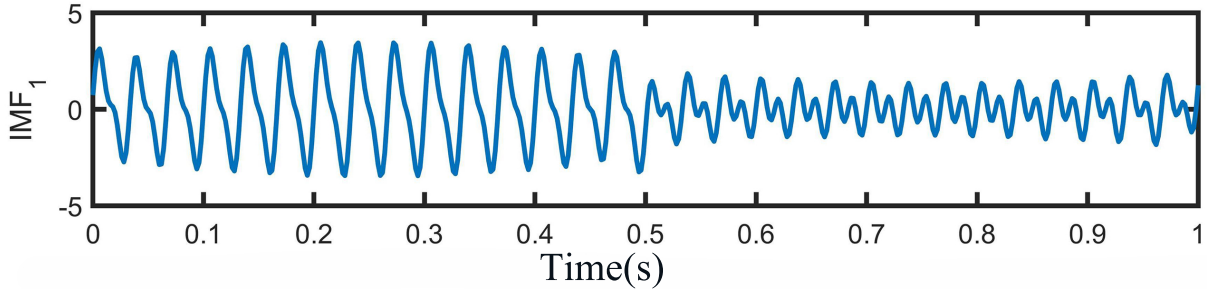
where ξ_i and ξ_i^* denote slack variables and C denotes the penalty factor.

The appropriate inner product function is used in the SVR algorithm according to the algorithm needed for engineering practice [**?**]. Due to the small amount of features in the financial time series data in this work and the non-linear component, the Gauss radial basis kernel function is chosen as the kernel function for the SVR.

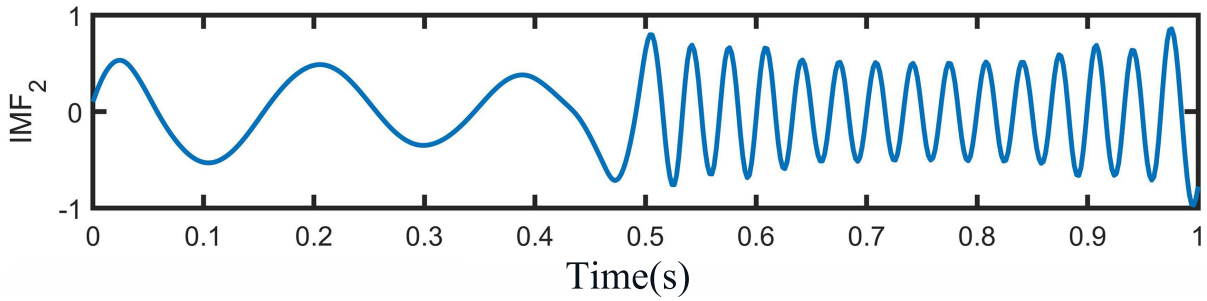
$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \sigma > 0 \tag{7}$$

where σ denotes the radial base bandwidth.

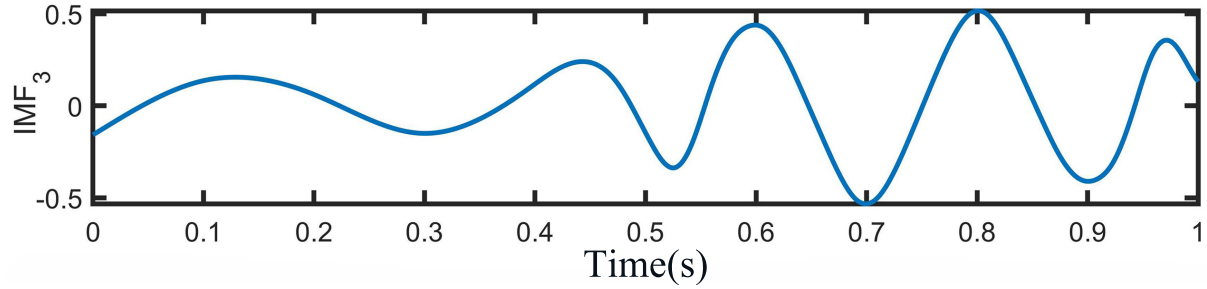
SVR increases with the training samples, but the computation time does not increase significantly with the number of samples. Only a small number of sample points play a role in the solution of SVR, and SVR uses a loss function to automatically ignore useless samples, so a suitable loss function needs to be selected for the calculation. Also, different loss functions will have an impact on the regression line and robustness of the SVR. In this paper, ε is chosen as the loss function.



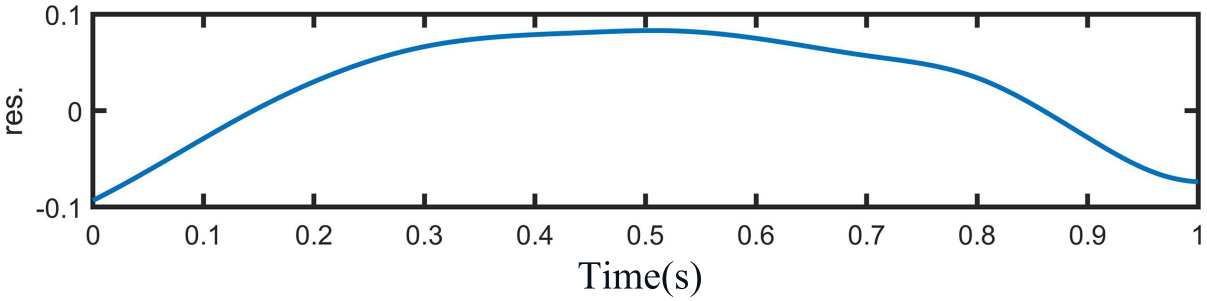
(a) IMF_1



(b) IMF_2



(c) IMF_3



(d) res

Figure 1. Results of EMD decomposition

2.3. **Sparrow search algorithm.** The sparrow algorithm is an algorithm obtained from the observation of sparrow foraging, in which the sparrow population is divided into two types of finders and followers, which can be transformed into each other. The discoverer provides the direction and location of food for the group, while the follower follows the discoverer to obtain food. Peripheral sparrows are more susceptible to invasion by alien species, so they need to adjust their position to ensure their safety and move closer to the safety zone.

Assume that the initial size of the sparrow search algorithm is n. Throughout the solution space, the location of the finder search for food is updated at each iteration of the search and is computed as follows [34]:

$$X_{i,j}\left(t+1\right)=\begin{cases} X_{i,j}\left(t\right)\cdot\exp\left(-\frac{i}{\alpha\cdot t_{\max}}\right), R<ST \\ X_{i,j}\left(t\right)+Q\cdot L, R\geq ST \end{cases} \tag{8}$$

where $X_{i,j}$ denotes the position information of the i-th sparrow in the solution space in the j-th dimension, $t_{\max}$ denotes the maximum number of iterations, α denotes a random value of (0,1], R denotes the warning value of a sparrow chirping when it encounters danger (taking values in the range of [0,1]), and ST denotes the safety coefficient (taking values in the range of [0.5,1]).

The foraging process is modelled for a population of sparrows. a population of n sparrows can be expressed as follow:

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix} \tag{9}$$

where n denotes the number of sparrow populations and d is the dimension of the solution space.

In the sparrow search algorithm, sparrows with higher fitness will have priority in obtaining food, increasing their own survival rate. The fitness function for the sparrow population is shown as follows:

$$F(x)=\begin{bmatrix} f([x_{1,1},x_{1,2},\ldots,x_{1,d}]) \\ f([x_{2,1},x_{2,2},\ldots,x_{2,d}]) \\ \vdots \\ f([x_{n,1},x_{n,2},\ldots,x_{n,d}]) \end{bmatrix} \tag{10}$$

where f indicates the fitness value.

In sparrow populations, there is a balanced shift in status between finders and joiners who occupy food areas. In addition to discoverers and joiners, a certain percentage of individuals in the population are randomly selected as vigilantes, who give early warning signals when the population is in danger. The location of the vigilantes is updated as follows:

$$X_{i,j}^{t+1}=\begin{cases} X_{\text{best}}^{t}+\beta|X_{i,j}^{t}-X_{\text{best}}^{t}|, f_{\text{ibest}}>f_{\text{gbest}} \\ X_{i,j}^{t}+K\left(\frac{|X_{i,j}^{t}-X_{\text{worst}}^{t}|}{(f_{\text{ibest}}-f_{\text{worst}})+\xi}\right), f_{\text{ibest}}=f_{\text{gbest}} \end{cases} \tag{11}$$

where X_{best}^{t} denotes the t-th iteration global optimal position, β denotes the step control parameter, K is a random number between $[-1,1]$, ξ is a minimal constant (to avoid the case where the denominator is 0). f_{best}, f_{gbest}, and f_{worst} denote the current adaptation, the best adaptation and the worst adaptation, respectively.

3. Time series forecasting model based on CEEMDAN-ISSA-SVR.

3.1. **Data pre-processing.** The CSI 300 index was selected as the basis of data modelling. The CSI 300 index is the best and most liquid 300 companies in the Shanghai and Shenzhen markets, and these companies basically represent the trend of A shares. It can also facilitate investors to make investment portfolio and provide important investment direction for investors. The time period for the initial CSI 300 data was chosen to run from 4 January 2011 to 31 May 2021 and the raw data is shown in Table 1. As this paper

Table 1. CSI 300 raw data

Date	Opening Price	Highest price	Lowest Price	Closing Price
2011-01-04	3155.557	3194.358	3143.604	3189.682
2011-01-05	3170.181	3193.779	3158.871	3175.662
2011-01-06	3177.835	3198.052	3152.572	3159.643
⋮	⋮	⋮	⋮	⋮
2021-05-27	5311.373	5378.476	5286.05	5338.233
2021-05-28	5338.728	5360.277	5288.652	5321.089
2021-05-31	5318.083	5331.629	5281.685	5331.57

only forecasts for the closing price of the stock market, only the date and closing price are retained. By sorting and filtering the initial data, a total of 3,129 samples were selected as the raw data for the forecasts in this paper. The first 70 % of the data were selected as the training set, 15 % of the data were selected as the validation set, and the remaining 15 % were used as the test set to analyse the effect of the predictions.

3.2. **CEEMDAN Modal Decomposition.** As there are many factors influencing the stock indices, their interactions are superimposed and difficult to be separated. In this paper, the CEEMDAN decomposition method can be used to decompose the raw stock data into IMF series and trend terms of different scales.

The IMF series of different scales are screened to obtain high frequency series using the statistical method of t-test. The high frequency series represents a short-term random fluctuation impact, but the impact of the high frequency series is relatively small in both time and intensity, and its mean value can be approximated as zero, or the high frequency series can be interpreted as noise. It is then summed up to obtain the new sequence. For the low-frequency series, it mainly reflects the periodic influence of each important factor on the original data. res trend term reflects a long-term slow trend of the original data.

Compared with wavelet decomposition, EMD decomposition is not affected by the choice of wavelet basis function, number of decomposition layers, threshold and threshold function, but it also has many problems, one of which is the mixing of modalities. The steps of the method are shown in detail as follows:

Assuming that the input signal sequence is added with T times of white noise, the signal sequence for the i-th time is

$$S_i(t) = S(t) + \varepsilon_0 \omega_i(t)(i = 1, 2, ..., T) \tag{12}$$

where ε_0 denotes the standard deviation, $\omega(t)$ denotes white noise and $S(t)$ denotes the initial sequence.

The first IMF sequence after CEEMDAN decomposition can be obtained by first decomposing $S_i(t)$ using EMD.

$$IMF_1 = \frac{1}{T} \sum_{i=1}^{T} IMF_{i1} \tag{13}$$

Arithmetic averaging of the first part leads to the second IMF sequence after CEEMDAN decomposition.

$$IMF_2 = \frac{1}{T} \sum_{i=1}^{T} E_1(r_1(t) + \varepsilon_1 E_1(\omega_i(t))) \tag{14}$$

$$r_1(t) = S(t) - IMF_1 \tag{15}$$

Then, the second stage residuals were calculated.

$$r_2(t) = r_1(t) - IMF_2 \tag{16}$$

And so on to obtain the n-th stage residual.

$$r_n(t) = r_{n-1}(t) - IMF_n \tag{17}$$

After performing T times EMD decomposition, the $(n+1)$-th IMF sequence after CEEMDAN decomposition is obtained.

$$IMF_{n+1} = \frac{1}{T}\sum\nolimits_{i=1}^{T} E_1(r_n(t) + \varepsilon_n E_n(\omega_i(t))) \tag{18}$$

The above steps are repeated until the sequence is not decomposable, then the decomposition of the original signal of CEEMDAN is formulated.

$$S(t) = r(t) + \sum\nolimits_{j=1} IMF_j \tag{19}$$

3.3. **The proposed ISSA algorithm.** An ISSA algorithm is proposed by introducing chaotic mappings, three optimization strategies and Corsi perturbations to enhance the global search and convergence capability of SSA in response to the shortcomings of SSA in the optimization process.

Logistic mapping as a typical chaotic system. In most of the existing bionic population intelligence algorithms, this is used to generate random solutions to determine the initial distribution state of the population. However, Logistic mappings do not have good ergodicity, thus affecting the initialisation of the population to some extent. In general, Kent mapping is isomorphic to logistic mapping, so it is practical to generate random solutions using Kent mapping. Kent mapping is shown as follow:

$$Z_{k+1} = \begin{cases} Z_k/\mu & 0 < Z_k < \mu \\ (1 - Z_k)/1 - \mu & \mu \le Z_k < 1 \end{cases} \tag{20}$$

where k is the number of mappings, p is the chaos control parameter and Z is the k-th mapping function value.

Compared to Logistic mappings, Kent mappings form chaotic systems that are less dependent on the values of the control parameters, have better ergodicity and are more conducive to function finding. Therefore, the Kent mapping is used to initialise the sparrow population.

$$X_i = X_{\text{tb}} + 0.5(X_{\text{tb}} - X_{\text{ub}})(y_i + 1) \tag{21}$$

where X_{ub} and X_{lb} are the upper and lower boundaries of the dimension in which individuals in the sparrow population are located.

The lack of a mechanism in the sinusoidal search strategy that effectively promotes local optimality-seeking capability leads to a rapid decline in the search capability of the SSA algorithm. The search mechanism of the original joiner in SSA is replaced using the sine and cosine search strategy so as to exploit the oscillation property for the search of the optimum in order to reduce the risk of falling into a local optimum solution. The improved joiner's position will produce some changes, and its position is calculated as follows:

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^t + S_1 \sin(S_2) \cdot \left| S_3 \cdot X_{\text{best}}^t - X_{i,j}^t \right| & , S_4 < 0.5 \\ X_{i,j}^t + S_1 \cdot \cos(S_2) \cdot \left| S_3 \cdot X_{\text{best}}^t - X_{i,j}^t \right| & , S_4 \ge 0.5 \end{cases} \tag{22}$$

where S_1, S_2, S_3, and S_4 are random numbers that follow a uniform distribution.

S_1 controls the direction of search and S_2 controls the distance the sparrow moves each time. When $S_3 > 1$, the original optimal position in the population will influence the position of the next generation of individuals.

In sparrow populations, the addition of vigilantes enhances the global search capability of the algorithm, but if the number of vigilantes is maintained at a constant value throughout the iteration cycle, it reduces the convergence capability of the algorithm at a later stage. Therefore, this work uses a dynamic change strategy to control the number of vigilantes throughout the iteration cycle. At the beginning of the iteration, the search space needs to be searched extensively to find the region where the optimum response is located, requiring an increase in the number of vigilantes in the early part of the iteration. Later in the iteration, decreasing the number of vigilantes enhances the search capability in the local area, thus increasing the convergence speed of the SSA algorithm later in the iteration. The decreasing number of vigilantes is done in the following manner:

$$num_i = \text{int}((1 - \frac{t}{t_{\max}})e) + 1 \tag{23}$$

where num_i is the number of vigilantes in the current iteration and e is the initial number of vigilantes.

The inverse learning strategy is a novel strategy for improving bionic population intelligence algorithms. A reverse solution is generated based on the current optimum response, and the optimum response is compared with the newly generated reverse solution, combined with the adaptation corresponding to the position, and selected for the subsequent iterations. The reverse solution F^t_{best} is calculated as shown as follow:

$$F^t_{\text{best}} = X_{\text{lb}} + (X_{\text{ub}} - X^t_{\text{best}}) \tag{24}$$

The backward learning strategy can improve the algorithm's local search for superiority to some extent, but the lack of random perturbation by the subsets makes the population prone to homogenisation in the late iterations. Therefore, a stochastic backward learning strategy is required. The calculation of the reverse solution is shown as follow:

$$H^t_{\text{best}} = X_{\text{lb}} + c(X_{\text{ub}} - X^t_{\text{best}}) \tag{25}$$

where c is a single row d-dimensional matrix consisting of random numbers between 0 and 1.

The probability function of the Cauchy distribution is as follows.

$$f(x) = \frac{1}{\pi}\frac{t}{t + x^2}, x \in (-\infty, +\infty) \tag{26}$$

When $t = 1$, the Corsi distribution is the standard Corsi distribution. At this point, the probability density curve for the Gaussian and Corsi distributions is shown in Equation 2:

The distribution characteristics show that the Corsi distribution has a smaller peak and a smoother decline from the peak to the nadir, with a uniform range of variation and better extension at both ends of the distribution, and individuals within the population have a higher probability of making a better position available; therefore, the Corsi analysis outperforms the Gaussian distribution in terms of perturbation variability. The use of the Corsi perturbation for globally optimal individuals facilitates the SSA algorithm to escape the extremes of the region, increasing the convergence ability of the SSA algorithm in the vicinity of the local extremes. Therefore, this work uses the Corsi distribution perturbation strategy to generate new individuals for SSA in the following manner:

$$G^t_{\text{best}} = X^i_{\text{best}}[1 + Cauchy(0, 1)] \tag{27}$$

where G^t_{best} is the position of the individual after the Cauchy variation and Cauchy$(0, 1)$ is the standard Cauchy distribution.

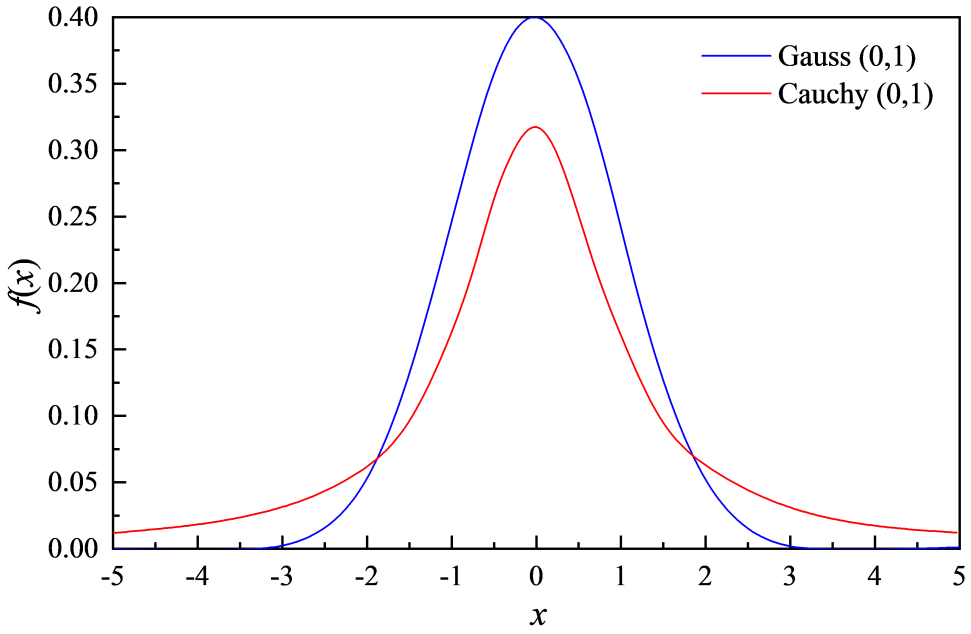


Figure 2. Gaussian and Corsi distribution probability density curves money

Finally, the global optimal individual is updated by reasonably allocating the operation of these two strategies, which further enhances the performance of SSA in finding the optimal individual. The strategy for updating the globally optimal individual is determined by the selection probability p, which can be expressed as follow:

$$p = 0.6 - \frac{0.1\,(t_{\max} - t)}{t_{\max}} \tag{28}$$

When $rand < p$, the update is performed using a stochastic backward learning strategy; otherwise, the update is performed using a Corsi variant. $rand$ is a random number between 0 and 1.

3.4. **SVR prediction model.** The IMFH sequences, low frequency sequences and trend terms obtained above are used as the input quantities of the SVR model to predict the original time series. In order to make the SVR have better generalization capability, the ISSA algorithm is used in this paper to optimize the parameters (C, σ, ε) of the SVR. The search position and speed of ISSA are determined by the 3D parameters (C, σ, ε).

Three metrics were used for the predictive model evaluation: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{29}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{30}$$

$$MAPE = \frac{100\%}{n} \times \sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{31}$$

where y_i is the true value of the sample, $\hat{y}_i$ is the predicted value of the sample and n is the number of samples.

4. Experimental results and analysis.

4.1. Experimental platform and dataset.

The experimental platform was based on a Windows system and the CEEMDAN-ISSA-SVR model was built using Python 3.7 software.

The data set is derived from the daily closing price data of the CSI 300 Index. After all data were eliminated from the null values as well as reordered, a total of 3129 samples were screened as the raw data for the predictions in this paper. The experimental hardware configuration is shown in Table 2.

Table 2. Main computer hardware configuration

Configuration	Parameters
Processor	Inter(R) Core(TM) i3 M 370 @2.4GHz 2.39GHz
Memory	8 G
Hard disk capacity	1 TB
System type	64-bit operating systems
Operating systems	Microsoft Window 10

The parameters related to the ISSA algorithm in the experiments are set as shown in Table 3.

Table 3. Parameter information

Parameters	Numerical values
safety coefficient ST	0.8
Population of sparrows N	50
the number of discovers PD	0.5
Standard deviation SD	0.05
max_iteration t_{max}	200

4.2. Prediction results.

The prediction error values for each series obtained after CEEMDAN decomposition and processing are shown in Table 4.

Table 4. Prediction errors by series

Serials	MAE	RMSE	MAPE (%)
IMF1	41.95	13.58	11.09
IMF2	32.19	16.07	6.16
IMF3	29.16	18.47	9.37
IMF4	28.63	15.18	3.23
IMF5	35.32	46.88	51.57
IMF6	0.89	1.03	1.29
IMF7	15.75	26.96	2.67
res	5.92	6.51	0.13

The trend term res series reflects the time series of the long term slowly changing trend of the market and also has a long period characteristic which determines the long term trend of the market and has a decisive impact. the forecast errors of the CEEMDAN-ISSA-SVR model are shown in Table 5.

Table 5. CEEMDAN-ISSA-SVR model prediction errors

Models	MAE	RMSE	MAPE (%)
CEEMDAN-ISSA-SVR	31.87	39.12	0.81

It can be seen that the MAE, RMSE and MAPE of the CEEMDAN-ISSA-SVR model on the test set are 31.87, 39.12 and 0.81 % respectively. The prediction result curves for each day from 4 January 2011 up to 31 May 2021 are shown in Figure 3. It can be seen that the predicted values of the CEEMDAN-ISSA-SVR model are almost identical to the true values.

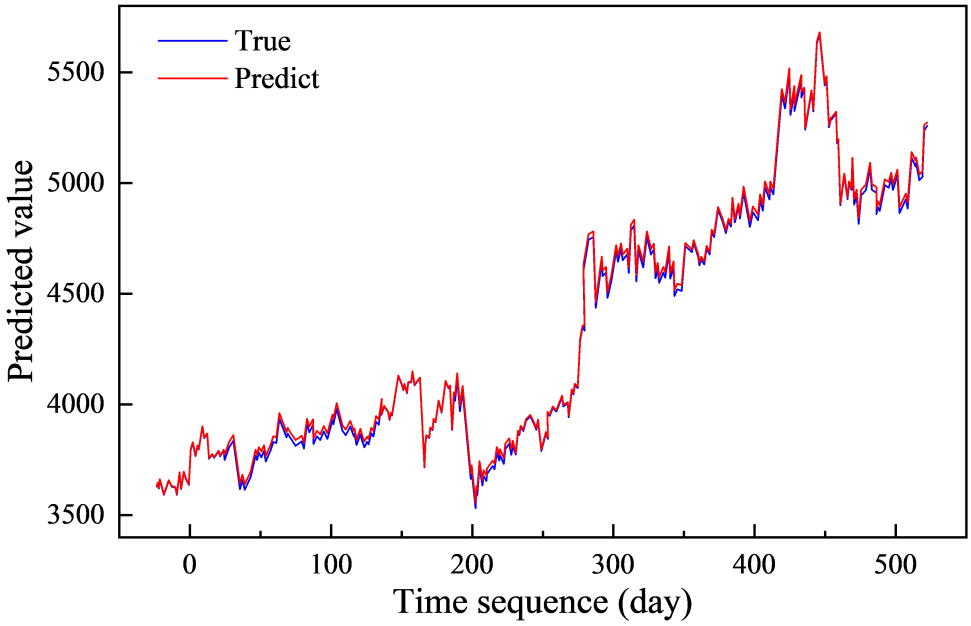


Figure 3. Daily predicted outcome curves

4.3. **Comparison of the prediction results of different models.** In order to further verify the effectiveness of the CEEMDAN-ISSA-SVR model proposed in this paper in dealing with financial time series, ARIMA, SVR, SSA-SVR and EMD-ISSA-SVR were selected for comparison experiments, and the results are shown in Table 6.

Table 6. Comparison of prediction errors of different models

Models	MAE	RMSE	MAPE (%)
CEEMDAN-ISSA-SVR	31.87	39.12	0.81
EMD-ISSA-SVR	32.91	40.09	0.95
SSA-SVR	37.72	45.94	1.09
SVR	39.54	47.08	1.11
ARIMA	44.37	51.25	1.18

It can be seen that the CEEMDAN-ISSA-SVR model has a MAE of 31.87, an RMSE of 39.12 and a MAPE of 0.81 % on the CSI 300 index test set. All the indicators are better than the four models, ARIMA, SVR, SSA-SVR and EMD-ISSA-SVR, proving that the CEEMDAN-ISSA-SVR model proposed in this paper has better fitting effect and lower prediction error. The main reason is that the CEEMDAN decomposition algorithm has

better denoising and analysis capability than the EMD decomposition algorithm. Compared with single machine learning models (SSA-SVR, SVR) and traditional time series ARIMA models, this work is a combined model combining CEEMDAN decomposition and machine learning models, which is more advantageous in several predictors.

5. **Conclusion.** This work builds a new combined CEEMDAN-ISSA-SVR forecasting model based on decomposition algorithms and machine learning models. CEEMDAN decomposes stock index data into high-frequency series, low-frequency series and trend terms. Among them, the high frequency series reflects the impact of short-term random factor disturbances in the market. The low-frequency series reflects policies with significant far-reaching effects and cyclical volatility characteristics. The trend term reflects the time series of the long term slowly changing trend of the market. By introducing chaotic mapping, sine cosine search strategy, dynamic change strategy, stochastic reverse learning strategy and Corsi perturbation, the ISSA algorithm is proposed with strong global search and convergence capability. Meanwhile, an ISSA-based SVR forecasting model for financial time series is constructed. In follow-up research we will try to apply the proposed ISSA-SVR model to more practical engineering areas.

Acknowledgements. The study was supported by Guangdong Nanhua Vocational College of Industry and Commerce quality engineering project " Teacher Teaching Innovation Team for Financial Services and Management Major " in 2023.

REFERENCES

[1] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, "A review on time series forecasting techniques for building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 74, pp. 902-924, 2017.

[2] Y. Ma, Y. Peng, and T.-Y. Wu, "Transfer learning model for false positive reduction in lymph node detection via sparse coding and deep learning," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 2, pp. 2121-2133, 2022.

[3] F. Zhang, T.-Y. Wu, Y. Wang, R. Xiong, G. Ding, P. Mei, and L. Liu, "Application of Quantum Genetic Optimization of LVQ Neural Network in Smart City Traffic Network Prediction," *IEEE Access*, vol. 8, pp. 104555-104564, 2020.

[4] S.-M. Zhang, X. Su, X.-H. Jiang, M.-L Chen, T.-Y. Wu, "A traffic prediction method of bicycle-sharing based on long and short term memory network," *Journal of Network Intelligence*, vol. 4, no. 2, pp. 17-29, 2019.

[5] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed, "DeepAnT: A deep learning approach for unsupervised anomaly detection in time series," *IEEE Access*, vol. 7, pp. 1991-2005, 2018.

[6] A. Tealab, "Time series forecasting using artificial neural networks methodologies: a systematic review," *Future Computing and Informatics Journal*, vol. 3, no. 2, pp. 334-340, 2018.

[7] S. Aminikhanghahi, and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and Information Systems*, vol. 51, no. 2, pp. 339-367, 2017.

[8] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific Reports*, vol. 8, no. 1, pp. 6085, 2018.

[9] X. Qiu, Y. Ren, P. N. Suganthan, and G. A. Amaratunga, "Empirical mode decomposition based ensemble deep learning for load demand time series forecasting," *Applied Soft Computing*, vol. 54, pp. 246-255, 2017.

[10] W. Bao, J. Yue, and Y. Rao, "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," *PloS One*, vol. 12, no. 7, e0180944, 2017.

[11] O. B. Sezer, and A. M. Ozbayoglu, "Algorithmic financial trading with deep convolutional neural networks: time series to image conversion approach," *Applied Soft Computing*, vol. 70, pp. 525-538, 2018.

[12] L. Chen, W. Gan, Q. Lin, S. Huang, and C.-M. Chen, "OHUQI: Mining on-shelf high-utility quantitative itemsets," *The Journal of Supercomputing*, vol. 78, no. 6, pp. 8321-8345, 2022.

[13] C.-M. Chen, L. Chen, W. Gan, L. Qiu, and W. Ding, "Discovering high utility-occupancy patterns from uncertain data," *Information Sciences*, vol. 546, pp. 1208-1229, 2021.

[14] W. Gan, L. Chen, S. Wan, J. Chen, and C.-M. Chen, "Anomaly Rule Detection in Sequence Data," *IEEE Transactions on Knowledge and Data Engineering*, 2022. [Online]. Available: https://doi.org/10.1109/TKDE.2021.3139086.

[15] S. Wang, J. Cao, and S. Y. Philip, "Deep learning for spatio-temporal data mining: A survey," *IEEE Transactions On Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3681-3700, 2020.

[16] W. Haoxiang, and S. Smys, "Big data analysis and perturbation using data mining algorithm," *Journal of Soft Computing Paradigm (JSCP)*, vol. 3, no. 01, pp. 19-28, 2021.

[17] V. Plotnikova, M. Dumas, and F. Milani, "Adaptations of data mining methodologies: a systematic literature review," *PeerJ Computer Science*, vol. 6, e267, 2020.

[18] T. Trzciński, and P. Rokita, "Predicting popularity of online videos using support vector regression," *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2561-2570, 2017.

[19] G. Santamaría-Bonfil, A. Reyes-Ballesteros, and C. Gershenson, "Wind speed forecasting for wind farms: a method based on support vector regression," *Renewable Energy*, vol. 85, pp. 790-809, 2016.

[20] A. H. Dyhrberg, "Bitcoin, gold and the dollar-A GARCH volatility analysis," *Finance Research Letters*, vol. 16, pp. 85-92, 2016.

[21] S. Khan, and H. Alghulaiakh, "ARIMA model for accurate time series stocks forecasting," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, pp. 31-42, 2020.

[22] P. Wang, H. Zhang, Z. Qin, and G. Zhang, "A novel hybrid-Garch model based on ARIMA and SVM for PM2.5 concentrations forecasting," *Atmospheric Pollution Research*, vol. 8, no. 5, pp. 850-860, 2017.

[23] E. A. Akpan, K. Lasisi, A. Adamu, and H. B. Rann, "Application of Iterative Approaches in Modeling the Efficiency of ARIMA-GARCH Processes in the Presence of Outliers," *Applied Mathematics*, vol. 10, no. 3, pp. 138-158, 2019.

[24] W. Kristjanpoller, and M. C. Minutolo, "Forecasting volatility of oil price using an artificial neural network-GARCH model," *Expert Systems with Applications*, vol. 65, pp. 233-241, 2016.

[25] Y. Guo, S. Han, C. Shen, Y. Li, X. Yin, and Y. Bai, "An adaptive SVR for high-frequency stock price forecasting," *IEEE Access*, vol. 6, pp. 11397-11404, 2018.

[26] D. Pradeepkumar, and V. Ravi, "Forecasting financial time series volatility using particle swarm optimization trained quantile regression neural network," *Applied Soft Computing*, vol. 58, pp. 35-52, 2017.

[27] H. Liang, J. Zou, Z. Li, M. J. Khan, and Y. Lu, "Dynamic evaluation of drilling leakage risk based on fuzzy theory and PSO-SVR algorithm," *Future Generation Computer Systems*, vol. 95, pp. 454-466, 2019.

[28] V. Sharma, D. Reina, and R. Kumar, "HMADSO: a novel hill Myna and desert Sparrow optimization algorithm for cooperative rendezvous and task allocation in FANETs," *Soft Computing*, vol. 22, pp. 6191-6214, 2018.

[29] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A. Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903-995, 1998.

[30] H.-L. Yang, and H.-C. Lin, "Applying the hybrid model of EMD, PSR, and ELM to exchange rates forecasting," *Computational Economics*, vol. 49, no. 1, pp. 99-116, 2017.

[31] J. Cao, Z. Li, and J. Li, "Financial time series forecasting model based on CEEMDAN and LSTM," *Physica A: Statistical Mechanics and Its Applications*, vol. 519, pp. 127-139, 2019.

[32] R. G. Thangarajoo, M. B. I. Reaz, G. Srivastava, F. Haque, S. H. M. Ali, A. A. A. Bakar, and M. A. S. Bhuiyan, "Machine learning-based epileptic seizure detection methods using wavelet and EMD-based decomposition techniques: A review," *Sensors*, vol. 21, no. 24, 8485, 2021.

[33] C. Luo, B. Keshtegar, S.-P. Zhu, and X. Niu, "EMCS-SVR: Hybrid efficient and accurate enhanced simulation approach coupled with adaptive SVR for structural reliability analysis," *Computer Methods in Applied Mechanics and Engineering*, vol. 400, 115499, 2022.

[34] C. Zhang, and S. Ding, "A stochastic configuration network based on chaotic sparrow search algorithm," *Knowledge- Based Systems*, vol. 220, 106924, 2021.