ICIC Express Letters ICIC International ©2021 ISSN 1881-803X
Volume **15**, Number **10**, October **2021** pp. 1075–1082

RESEARCH AND APPLICATION OF A NOVEL HYBRID FORECASTING SYSTEM FOR GOLD PRICE FORECASTING

Meng Du* and Yixin Chen

School of Economics and Management
Dalian University
No. 10, Xuefu Street, Jinzhou New District, Dalian 116622, P. R. China
*Corresponding author: dumeng@dlu.edu.cn; 2831301454@qq.com

Received February 2021; accepted April 2021

Abstract. *Forecasting gold price is very vital for the investments and decisions for mining projects and related companies. This paper proposes a hybrid forecasting framework for gold price forecasting. In this paper, Ensemble Empirical Mode Decomposition (EEMD) is applied to dividing the original gold price data into a set of components. Then, to achieve high accuracy, the whale optimization algorithm is used to optimize the initial weights between layers and the thresholds of the least square SVM. Finally, several criteria are introduced to make a comprehensive evaluation of this forecasting system, and the empirical studies reveal the proposed hybrid method is more accurate than traditional methods for gold price forecasting.*

Keywords: Gold price forecasting, Hybrid method, LSSVM, Whale optimization algorithm

1. Introduction.

Commodity price forecasting is crucial for risk management, commodity pricing and decision-making, and asset allocation. It has been widely paid attention to by academics and the physical world for many years [1]. Gold is a strategic resource. Among all precious metals, gold is the most common investment choice. It is the target of most central banks' international reserves. At the same time, gold is also an important raw material and widely used in the fields of electronics, chemistry and medicine. Therefore, small improvements in the prediction accuracy of gold price can bring huge profits [2]. However, how to accurately predict the trend of the gold price is a difficult problem.

Over the past few decades, several forecasting models have been developed. Traditional economic and statistical models such as Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), Generalized Autoregressive Conditional Heteroskedasticity (GARCH) and other GARCH family models are used to predict crude oil prices. However, the data used by these models need to be stable, and in order to overcome the shortcoming of these models, the references recommend using machine learning methods to forecast the price of gold. Artificial Neural Networks (ANN), the SVM, deep learning [3] to deal with nonlinear, unstable and complicated data structure are better than the traditional time series model.

Now, researchers are more likely to use hybrid models to predict commodity price trends. Hybrid methods with linear and nonlinear modeling capabilities can become the core strategy to solve such problems [4]. Usually, artificial neural networks are used to predict commodity price fluctuations in combination with other models. Kristjanpoller and Minutolo [5] used the mixed model combining ANN and GARCH model to predict the gold price, and found that the mixed model had a higher prediction accuracy than the GARCH model. Risse [6] combined discrete wavelet transform with support vector regression to predict the gold price dynamics, and the results showed that the long and

DOI: 10.24507/icicel.15.10.1075

short term trends of gold price were not stable through wavelet decomposition. Alameer et al. [7] used the Whale Optimization Algorithm (WOA) to train the multi-layer sensing Neural Network (NN) to predict the gold price. Compared with NN, PSO-NN, GA-NN, GWO-NN and ARIMA, this hybrid prediction model obtained the minimum mean square error.

Most studies focused on the impact of specific factors on the gold price, rather than the prediction or classification ability of the model. In contrast, our goal is to develop a forecasting model that can be used by investors as a decision support tool. Generally speaking, the development of forecasting model is more challenging than the estimation of explanatory model. The main reason is that the accuracy of super sample and super time prediction and the ability to generate return on investment become very important. This paper aims to extract and forecast the price trend of gold through hybrid EEMD and WOA-LSSVM. The basic concept of this method is that the hybrid forecasting model can reduce the generalization error in forecasting.

2. **Data Source and Processing.** This paper uses 360 months gold prices from July 1990 to June 2020 as the sample data, the data are from the Wind Economic Database, the gold price data trend is shown in Figure 1.

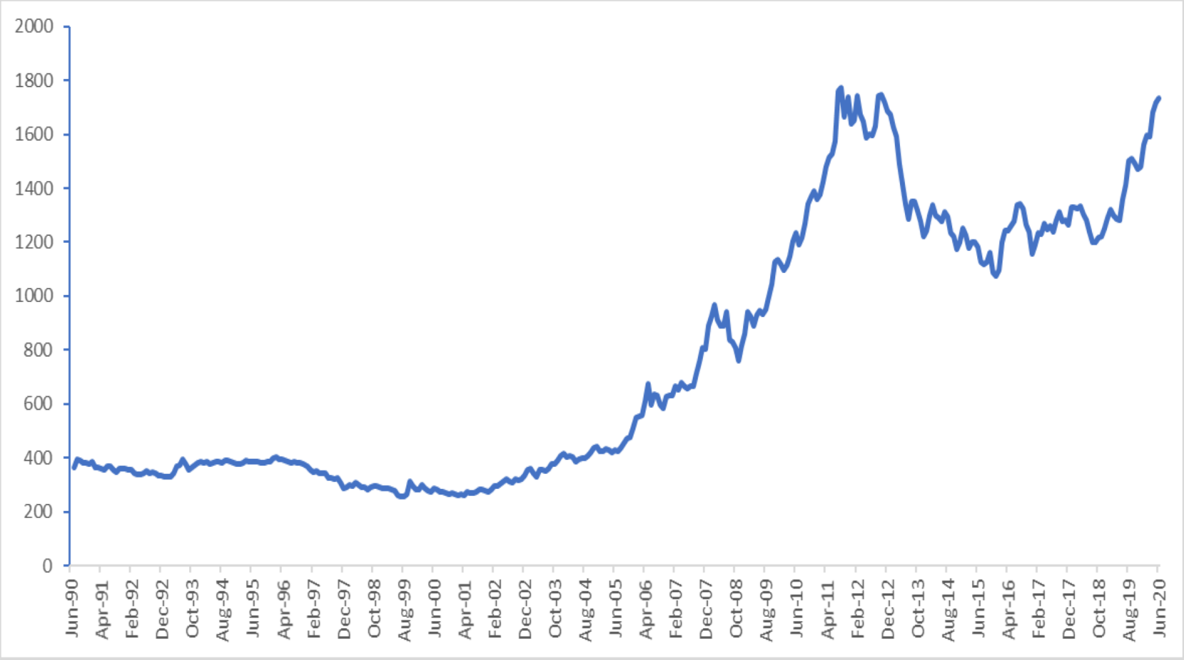

Figure 1. The monthly gold price from July 1990 to June 2020

Before statistical modeling, the data must be preprocessed. Data preprocessing includes the review, screening and sorting before data classification or grouping. In this paper, EEMD decomposition technology is used to remove noise, and it adds a specific noise at each stage of decomposition and calculates the unique residuals to get each mode.

1) Constructing a new sequence by adding the white noise sequence with normal distribution $w^i(t) \sim N\left(0, \sigma^2\right)$ to the original time series $y(t)$, it shows as follows:

$$y^i(t) = y(t) + w^i(t) \tag{1}$$

2) We use EMD model to decompose $y(t)$ into n IMF components $c_j^i(t)$ $(j = 1, 2, \ldots, n)$ and a residue $s^i(t)$, and the equation becomes as follows:

$$y^i(t) = \sum_{j=1}^{n} c_j^i(t) + s^i(t) \tag{2}$$

where $c_j^i(t)$ is the jth IMF component in the ith test.

3) Repeat n times steps 1) and 2), and caculate the average of the IMF component in n times, and then get the ultimate intrinsic mode function:

$$c_j(t) = \frac{1}{N}\sum_{i=1}^{N} c_j^i(t) \tag{3}$$

The original time series can be represented as a linear combination of IMFs and a residue, as follows:

$$y(t) = \sum_{j=1}^{n} c_j(t) + s(t) \tag{4}$$

where $c_j(t)$, $(t = 1, 2, \ldots, T)$ is the jth IMF component extracted in the jth separation process at time t, $s(t)$ is the final residual, and n is the number of IMF components.

3. **Model Establishment and Analysis.** In order to improve the prediction accuracy of gold price, this paper constructs a hybrid model by adding the whale optimization algorithm into LSSVM model, and the empirical results show that the combined prediction model proposed in this paper can significantly improve the prediction accuracy of gold price, which can well capture the change information of gold price.

(i) Whale optimization algorithm

Whale optimization algorithm is a new swarm intelligence optimization algorithm based on the unique hunting behavior of humpback whales. The advantages of this algorithm are simple operation and fewer parameters need to be adjusted. In the whale algorithm, the position of each whale represents a feasible solution, and the position of each whale in the n-dimensional solution space can be expressed as:

$$X = (x_1, x_2, \ldots, x_N) \tag{5}$$

There are two strategies for each whale: one is to surround the prey, and the other is to drive the prey through spiral position update. The encircling strategy can be divided into two parts: swimming to the optimal position and swimming to the random position. The probability that each whale chooses to surround or drive is equal, namely, P (surrounded) = P (driving) = 0.5, in order to describe the whale algorithm based on the strategy of feeding process, we will describe in detail the mathematical model.

1) The whale swims toward the optimal position

The position update formula of whales under this strategy is as follows:

$$\overrightarrow{x^{t+1}} = \overrightarrow{x_{best}^{t}} - \overrightarrow{A} * \left|\overrightarrow{C} * \overrightarrow{x_{best}^{t}} - \overrightarrow{x^{t}}\right| \tag{6}$$

where t is the current iteration number, $\overrightarrow{x^{t+1}}$ is the next updated position vector of the whale, $\overrightarrow{x_{best}^{t}}$ is the current optimal position vector of the whale, $\overrightarrow{x^{t}}$ represents the current position vector of the whale, and $\overrightarrow{A}$ represents the coefficient vector, in which $\overrightarrow{C}$ can be obtained by the following formula:

$$\overrightarrow{A} = 2a\overrightarrow{r_1} - a \tag{7}$$

$$\overrightarrow{C} = 2\overrightarrow{r_2} \tag{8}$$

$$a = 2 - \frac{2t}{T_{\max}} \tag{9}$$

where r_1 and r_2 are the random number of $(0, 1)$, C is a random number evenly distributed within $(0, 2)$, the value of a decreases linearly from the initial value 2 to 0 with the number of iterations, t represents the current number of iterations, and T represents the maximum number of iterations.

2) Swim to the location of random whales

The position update formula of whales under this strategy is as follows:

$$\overrightarrow{x^{t+1}} = \overrightarrow{x_{rand}^{t}} - \overrightarrow{A} * \left|\overrightarrow{C} * \overrightarrow{x_{rand}^{t}} - \overrightarrow{x^{t}}\right| \tag{10}$$

where $\overrightarrow{x_{rand}^{t}}$ is the location of randomly selected whales in the current population, the range of A possibly increases with the reduction of a, the range of A decreases along with the fall of a, when $|A| < 1$, the whales choose to swim toward the whales of the best location, when $|A| >= 1$, it swims to the random position of the whale, which can strengthen the detection ability of the algorithm and make WOA conduct global search.

3) Prey repelling strategy

When whales hunt, they constantly update their position and spiral to swim to prey. Under this strategy, the updating formula of whales' position is as follows:

$$\overrightarrow{x^{t+1}} = \left|\overrightarrow{x_{*}^{t}} - \overrightarrow{x^{t}}\right| * e^{bl} * \cos(2\pi l) + \overrightarrow{x_{*}^{t}} \tag{11}$$

where $\left|\overrightarrow{x_{*}^{t}} - \overrightarrow{x^{t}}\right|$ is the distance between the whale and the prey, b is a constant used to define the spiral shape, and l is a random number evenly distributed within $[-1, 1]$.

4) Hunting behavior

In the actual process of predation, whales swim to prey in a spiral shape and also take measures of contracting encircling. Therefore, in this behavior model, it is assumed that the probability of whales choosing the strategy of encircling prey is Pi, and then the probability of choosing the strategy of encircling prey is $1 - Pi$. In this case, the position updating formula of whales is as follows:

$$\overrightarrow{x^{t+1}} = \begin{cases} \overrightarrow{x_{best}^{t}} - \overrightarrow{A} * \left|\overrightarrow{C} * \overrightarrow{x_{best}^{t}} - \overrightarrow{x^{t}}\right| & |A| < 1,\ p < Pi \\ \overrightarrow{x_{rand}^{t}} - \overrightarrow{A} * \left|\overrightarrow{C} * \overrightarrow{x_{rand}^{t}} - \overrightarrow{x^{t}}\right| & |A| >= 1,\ p < Pi \\ \left|\overrightarrow{x_{*}^{t}} - \overrightarrow{x^{t}}\right| * e^{bl} * \cos(2\pi l) + \overrightarrow{x_{*}^{t}} & p \geq Pi \end{cases} \tag{12}$$

(ii) Least Squares Support Vector Machine (LSSVM)

Support Vector Machine (SVM) is a kind of machine learning algorithm based on statistical learning theory, the purpose of which is in high-dimensional feature space, in the different types of data to find the most optimal hyperplane to classify and forecast, the algorithm principle is through the application of kernel function and high-dimensional data reduction plan, to minimize the structural risk and realize the data classification or regression. In the application of SVM, there are problems such as the selection of hyperplane parameters and the influence of the number of training samples on the matrix size in the solution, which lead to the solution scale being too large. Suykens and Vandewalle [8] proposed the Least Squares Support Vector Machine (LSSVM), which starts from a machine learning loss function and uses two paradigms in the objective function of the optimization problem. The inequality constraints in SVM are replaced by the equality constraints, and the sum loss function of the square error is regarded as the experience loss of the training set, so the optimization problem is transformed into the solution of a set of linear equations, so it has a faster solving speed and accuracy. The basic equation of LSSVM is as follows:

$$\min J(\omega, \xi) = \frac{1}{2}\omega^{T}\omega + \frac{1}{2}\gamma \sum_{i=1}^{n} \xi_i^2 \tag{13}$$

$$\text{s.t. } Y_i = \omega^{T}\varphi(x_i) + b + \xi_i, \quad i = 1, 2, \ldots, n \tag{14}$$

where J is the risk boundary of structure, ω is the weight matrix, $\sum_{i=1}^{n}\xi_i^2$ is the error control function, b is the deviation, and $\varphi(x_i)$ denotes a kernel function. Lagrange function is used to solve this optimization problem, and it shows as follows:

$$L(\omega, b, \xi; \alpha) = J(\omega, \xi) - \sum_{i=1}^{n} \alpha_i \left\{\omega^T \varphi(x_i) + b + \xi_i - Y_i\right\} \tag{15}$$

Here is the Lagrange multiplier, by taking the derivative of ω, b, ξ, α respectively:

$$\frac{\partial L}{\partial \omega} = \frac{\partial L}{\partial b} = \frac{\partial L}{\partial \xi_i} = \frac{\partial L}{\partial \alpha_i} = 0 \tag{16}$$

$$\omega = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \tag{17}$$

$$\sum_{i=1}^{n} \alpha_i - ab = 0 \tag{18}$$

$$\omega^T \varphi(x_i) + b + \xi_i - Y_i = 0, \quad i = 1, 2, \ldots, n \tag{19}$$

The kernel function is defined as $K(x_i, x_j) = \varphi^T(x_i)\varphi(x_i) = x_i^T x_j$, $i, j = 1, 2, \ldots, n$, $K(x_i, x_j)$ is a symmetric function that satisfies Mercer's condition. According to the above formula, the optimization problem is transformed into solving a linear equation:

$$\begin{bmatrix} -a & 1 & \cdots & 1 \\ 1 & \varphi^T(x_1)\varphi(x_1) + \dfrac{1}{\gamma} & \cdots & \varphi^T(x_1)\varphi(x_i) \\ \cdots & \cdots & \cdots & \cdots \\ 1 & \varphi^T(x_1)\varphi(x_1) & \cdots & \varphi^T(x_i)\varphi(x_i) + \dfrac{1}{\gamma} \end{bmatrix} \begin{bmatrix} b \\ \alpha_1 \\ \vdots \\ \alpha_i \end{bmatrix} = \begin{bmatrix} 0 \\ Y_1 \\ \vdots \\ Y_i \end{bmatrix} \tag{20}$$

The least square method is used to solve the above equation and the least square regression function is derived

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x_j) + b \tag{21}$$

(iii) Hybrid forecasting model

As the penalty factor a and kernel function K in the constraints of LSSVM model affect the performance of the model, this paper uses the WOA to optimize them to improve the prediction accuracy of the model.

Step 1: Using EEMD to decompose the original data and normalize them, the sample data is divided into a training set and sample set according to 8 : 2.

Step 2: Initialize WOA and LSSVM parameters.

Step 3: Calculate the fitness of the individual whale population, and determine the current optimal whale individual according to the optimal fitness value; the algorithm iterates, updates the position of individual whale, and finally outputs the optimal individual result.

Step 4: Determine the a and K in LSSVM according to the optimal results, and then classify the test set to realize the evaluation and prediction of gold price series.

The flow chart of the hybrid prediction model is shown in Figure 2.

4. **The Simulation Results.**

Before the network training, this paper adopts EEMD decomposition and denoising technology for data preprocessing. After removing the high-noise data from the original sequence, the denoised data column is obtained, which is used to forecast the gold price, and the EEMD decomposition graph is as Figure 3.

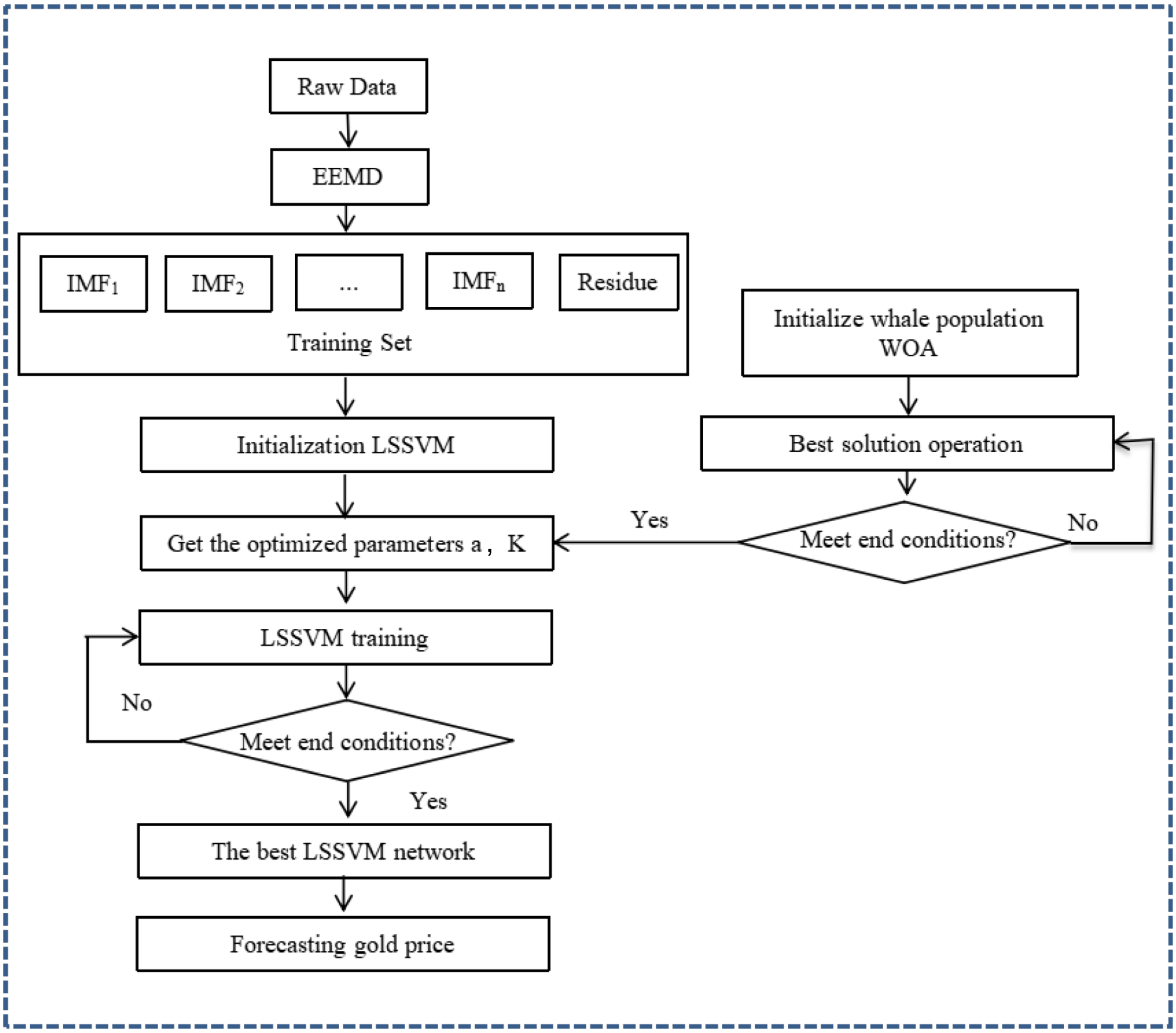

Figure 2. The flow chart of the hybrid prediction model

The first 288 items of the 7 IMF components formed by EEMD decomposition of the original data were taken as the training set, and the last 72 items were taken as the test set. WOA-LSSVM regression prediction algorithm was used to predict the test data set. In order to prove the prediction performance of the proposed model EEMD-WOA-LSSVM, several other models were selected for comparison, which are EEMD-GA-LSSVM, EEMD-POS-LSSVM, EEMD-LSSVM and LSSVM. The selection of these models is based on the basic model and similar model of the proposed model. EEMD-LSSVM and LSSVM are the basic models, and GA and POS are optimized algorithms that can be compared with WOA. In this paper, we choose Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) to evaluate the prediction performance; generally speaking, the smaller each error value is, the more accurate the prediction of the model is. Table 1 describes the evaluation indexes of different prediction models.

5. Conclusion.

In this paper, a hybrid forecasting method combining EEMD and WOA-LSSVM is proposed to solve the forecasting problem of gold price time series. The original data of gold price were decomposed by EEMD method to obtain IMF component and residual, which were input as initial parameters into WOA-LSSVM model for simulation prediction. Through empirical analysis and comparison with the prediction results of other mixed models, it is proved that this method has certain reliability and effectiveness. The specific conclusions are as follows:

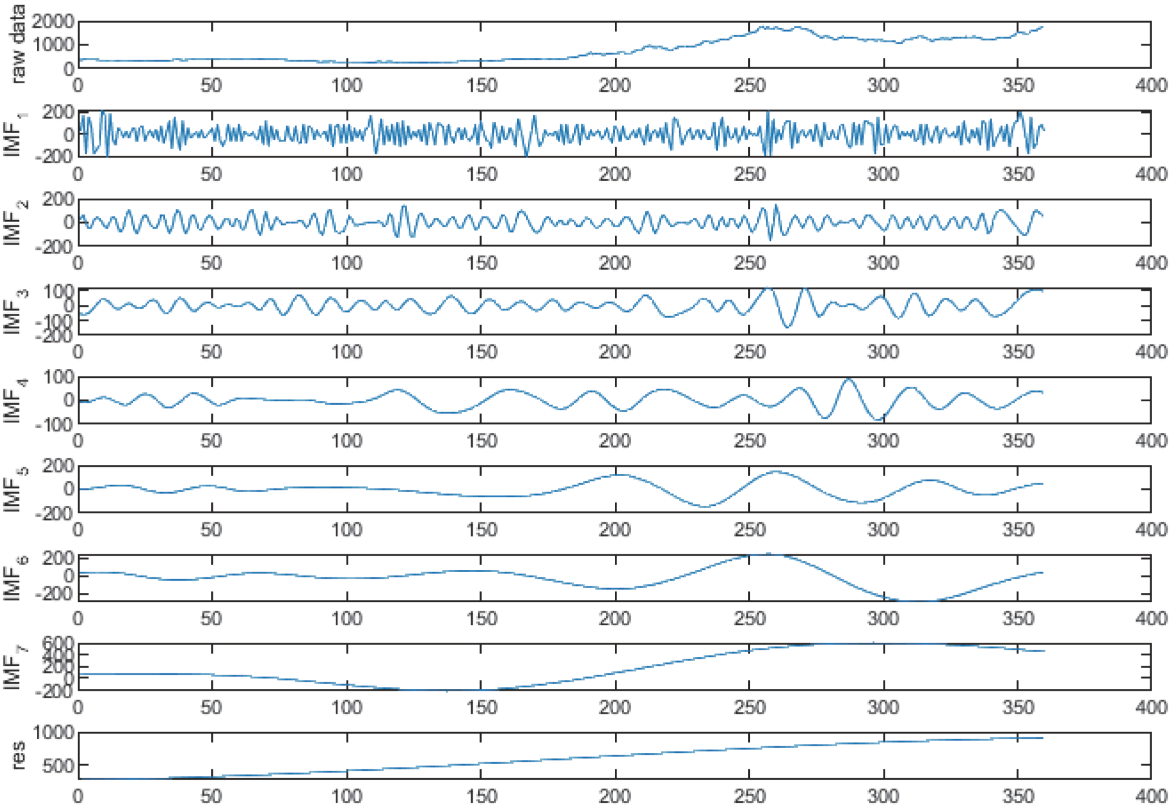

Figure 3. The EEMD decomposition graph

Table 1. Performance measures for the mixed prediction model

MODEL	RMSE	MAE	MAPE
EEMD-WOA-LSSVM	91.6964	70.5484	5.136%
EEMD-GA-LSSVM	96.2404	81.0573	5.912%
EEMD-POS-LSSVM	91.7691	70.8687	5.188%
EEMD-LSSVM	109.268	81.3748	6.112%
LSSVM	108.1749	81.5791	6.422%

1) The use of EEMD decomposition method can effectively overcome the modal mixing problem and further improve the accuracy of subsequent price prediction;

2) Compared with other prediction models, the prediction results of EEMD-WOA-LSSVM are better than those of EEMD-GA-LSSVM, EEMD-POS-LSSVM, EEMD-LSS-VM and LSSVM. It shows that the EEMD-WOA-LSSVM hybrid model has better prediction performance and can effectively predict the trend of the gold price.

The model will help investors and market participants to make accurate financial decisions in the financial investment market. In future work, the EEMD-WOA-LSSVM model will be applied to financial product prices to determine their forecasting accuracy.

REFERENCES

[1] L. Fang, B. Chen, H. Yu et al., The importance of global economic policy uncertainty in predicting gold futures market volatility: A GARCH-MIDAS approach, *Journal of Futures Markets*, vol.38, no.3, pp.413-422, 2018.

[2] D. Liu and Z. Li, Gold price forecasting and related influence factors analysis based on random forest, in *Proceedings of the Tenth International Conference on Management Science and Engineering Management. Advances in Intelligent Systems and Computing*, J. Xu, A. Hajiyev, S. Nickel and M. Gen (eds.), Singapore, Springer, 2017.

[3] T. Zhao, Y. Wang, Q. Guo and R. Zeng, A novel method based on numerical fitting for oil price trend forecasting, *Applied Energy*, vol.220, pp.154-163, 2018.

[4] M. Khashei and M. Bijari, A novel hybridization of artificial neural networks and ARIMA models for time series forecasting, *Applied Soft Computing*, vol.11, no.2, pp.2664-2675, 2011.

[5] W. Kristjanpoller and M. C. Minutolo, Gold price volatility: A forecasting approach using the artificial neural network – GARCH model, *Expert Systems with Applications*, vol.42, no.20, pp.7245-7251, 2015.

[6] M. Risse, Combining wavelet decomposition with machine learning to forecast gold returns, *International Journal of Forecasting*, vol.35, no.2, pp.601-615, 2019.

[7] Z. Alameer, M. A. Elaziz, A. A. Ewees, H. Ye and J. Zhang, Forecasting gold price fluctuations using improved multilayer perceptron neural network and whale optimization algorithm, *Resources Policy*, vol.61, pp.250-260, 2019.

[8] J. A. K. Suykens and J. Vandewalle, Least squares support vector machine classifiers, *Neural Processing Letters*, vol.9, pp.293-300, 1999.

[9] O. M. F. Montemayor, A. L. Rojas, S. L. Chavarria, M. M. Elizondo, I. R. Vargas and J. F. G. Hernandez, Mathematical modeling for forecasting the gross domestic product of Mexico, *International Journal of Innovative Computing, Information and Control*, vol.14, no.2, pp.423-436, 2018.