Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-14497-9

A deep learning-based nonlinear ensemble approach with biphasic feature selection for multivariate exchange rate forecasting

Jujie Wang[1] · Maolin He[1] · Wenjie Xu[1] · Feng Jing[1]

Received: 15 July 2021 / Revised: 20 June 2022 / Accepted: 31 January 2023

Abstract
Exchange rate prediction is a challenging task for investors and policymakers due to its nonstationary and nonlinear characteristics. This study develops a novel deep learning-based nonlinear ensemble approach with biphasic feature selection for multivariate exchange rate forecasting. The novel hybrid model has three modules, the preprocessing of target series, feature engineering, and forecasting module. The first module aims to extract more regular signals and reduce the dimension of the target series for better feeding into the forecasting module. The second part is the selection and reconstruction of twelve external variables, which outputs three sequences that contain the most important information. The forecasting part consists of Bi-directional long short-term memory and attention mechanism, which has better performance than the basic artificial intelligence algorithms. Three evaluation indicators are adopted to assess the hybrid model's performance. The results show that the new model performs better than other compared models.

Keywords Exchange rate forecasting · Biphasic feature selection · Deep learning · Nonlinear ensemble framework

1 Introduction

In recent years, the need for hedging exchange rate risk is growing steadily by the day. The accurate prediction of exchange rate can even have an impact on international transactions and the global currency market. For investors, policymakers and entrepreneurs especially in multinational companies, it is critical to forecast exchange rates as accurately as possible, because they can take advantages of it to design trading strategies and hedge currency risks in advance. Additionally, commercial banks hold the largest amount of foreign exchange assets

✉ Jujie Wang
jujiewang@126.com

[1] School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China

Published online: 20 February 2023

in China. If commercial banks cannot nicely predict foreign exchange rates, they cannot timely respond to the upcoming foreign exchange risk, which will bring huge economic losses. Moreover, central banks need utilize it to supervise markets and analyze macroeconomic situations [2]. There is a general recognition of the increasingly need for finding new ways to predicting exchange rates.

According to existing literature, the forecasting models for exchange rates are categorized into four main types: fundamental theories, statistical models, artificial intelligence (AI) and hybrid forecasting approaches. Fundamental models emphasize macroeconomic factors' effect in explaining exchange rate movements such as purchasing power parity, monetary models, and the portfolio balance model [17]. However, Meese and Rogoff examined that all the forecasts of exchange rates yielded by fundamental models could not outperform the random walk results [19]. Until now, the conclusion has not been overturned. In spite of this, the economic value of fundamental models cannot be denied and their drawbacks are overstated [1]. All in all, fundamental models can provide some theoretical guidelines for related workers to help them think about the mechanism of exchange rates, but they are not suitable for forecasting the numerical values of exchange rates, especially the short-term prediction which is an important and challenging task.

Statistical models tend to be conventional econometric models such as autoregressive integrated moving average (ARIMA), random walk (RW), and generalized autoregressive conditional heteroscedasticity (GARCH) [5, 34]. In addition, there are a lot of their extensions from the benchmark models such as combining the decomposition with GARCH [4], fractionally integrated GARCH [6], and exponential smooth transition autoregressive (ESTAR) [23]. They underline full use of the historical data to predict. Thus the prediction accuracy improves a lot and they also got extensive applications with respect to financial analysis. However, the statistical models need to obey the statistical principles (such as the linear form's assumption), which limits the application of these models [43]. It is well known that the exchange rate series is nonstationary and nonlinear. Therefore, statistical models cannot well capture the nonlinear patterns within the exchange rate time series, hindering further improvement in prediction performance [33].

In order to solve the problem, artificial intelligence (e.g. neural networks) is an alternative approach to statistical methodologies [17]. Neural network (NN) is a data-driven model which trains a certain amount of past observations with the purpose of discovering the underlying momentary relationship between the past and present [7]. It is proved that both linear and nonlinear models in artificial intelligence outperform the conventional statistical models such as the simple random walk in terms of exchange rate forecasting [8, 10, 18, 22, 25–27, 30, 41]. At first, researchers thought highly of NNs because of their flexibility and high prediction accuracy in the short term compared with AR and RW [7]. Unfortunately, NNs easily trapped into local optimal and the internal relationship among the inputs is out of consideration [14]. In order to solve it, recurrent neural network (RNN) is developed to capture the sequence information in the time series. But RNN still has the gradient explosion problem which restricts it a wider spreading [14]. Thus, scholars shifted their attention to the variants of RNN such as long short-term memory (LSTM) and gate recurrent unit (GRU) which can greatly alleviate gradient explosion through the addition of gate mechanism. Wang et al. employed GRU to forecast short-term financial time series and achieved a splendid performance [31]. Many researchers also attempted LSTM as the forecasting part in their models in the financial field in recent years [13, 38]. Zhou, Huang and Zhang summarized the short-term forecasting and concluded that LSTM is a mature and stable model compared to GRU according to the empirical results [42]. Moreover, based on LSTM, a new variant, Bi-directional LSTM (Bi-LSTM), is proposed to fully utilize the known information by architecturally introducing two forward and backward LSTM layers [15, 20].

The AI approaches have a potential for progress due to their weaknesses in how to settle the optimal parameters and how to handle the problems of misspecification, under-fitting, and over-fitting [33]. Thus, the hybrid forecasting models are proposed in the hope of further improving the accuracy of forecasting [11, 17, 28, 29, 32, 33, 37, 39]. In the literature, the most common framework in hybrid models is decomposition-ensemble approaches. They decompose the original data and then many basic models are used for forecasting. At last, an integrated result can be obtained according to averaging or voting mechanism. Thus, the decomposition-ensemble model can generate advantages from different individual models and can also reduce the risk of large forecast bias [17]. There are some frequently used decomposition techniques such as the wavelet transform, seasonal-trend decomposition procedure based on Loess (STL), and empirical-mode decomposition (EMD). But the wavelet transform's performance is dependent on the selection of basic functions and orders to a large extent and STL cannot handle the data with great randomness well [14]. EMD has the mode mixing problem that a single intrinsic mode function (IMF) contains several time scales or the same time scale appears in different IMFs. Singular spectrum analysis (SSA) as the decomposition method can deal with the nonstationary and nonlinear exchange rates pretty well and is good at extracting information from the time series containing noise. Researchers always use SSA with the de-noise method, which may discard some information. The combination between decomposition and forecasting can well solve the shortcomings of single model and improve prediction performance.

However, in the existing literature, most researches are solely dependent on historical data of the target series, and the external information from other markets is ignored even though these factors can offer useful signals to the fluctuations of exchange rates. In fact, feature engineering arises increasing attention in many fields and becomes a non-negligible part in forecasting. Two points need to consider carefully in feature engineering, which are the number of total features and the selection of features. In this paper, support vector machine recursive feature elimination (SVM-RFE) is employed to rank the importance of features and aims to discass the irrelevant variables and filter ones which can offer useful information in the prediction of exchange rates. In addition, the purpose of using stacked auto encoder (SAE) is to reduce the number of inputs to forecasting algorithm, which can speed up the computation and increase the efficiency.

In the abovementioned studies, forecasting part can be improved through the introduction of attention mechanism which can assign larger weights in more vital parameters so as to improve the prediction accuracy of the model. Wang et al. proposed a new model based on LSTM with attention mechanism to forecast carbon prices, which proves the usefulness of attention layer [31]. In this study, we combined the attention layer with Bi-LSTM to predict the exchange rate and from empirical results it shows better performance than other benchmark models.

In this paper, a new hybrid model proposed here may solve the problems above, which combined SSA, stacked auto encoder (SAE), LSTM and support vector machine recursive feature elimination (SVM-RFE). The original data is first decomposed through SSA and this operator can make predictors easier capture the nonlinear and nonstationary characteristics of data, well restrain the effects of noise on predictive accuracy, and save more computational time. After that SAE is used for data reconstruction, which can not only reduce the features' dimension but also save the relatively complete information compared to de-noise technology. Among the exogenous variables the SVM-RFE elected part of economic factors which were reconstructed to fewer series through SAE. This kind of processing to external variables can satisfy the two targets in feature selection mentioned above. At last the reconstructed data were inputted into Bi-LSTM with attention layer and it would output the final predicting results. The Bi-LSTM with attention layer has better ability to extract the long-term and short-term information without gradient explosion and its efficiency also improved with the weigh assignment which is calculated based on the signals' importance. In summary, this article has the following contributions.

(a) The new proposed model takes other economic factors into consideration in the novel hybrid model proposed in this paper, which are not common in the artificial intelligence and ensemble learning models in the exchange rate forecasting.
(b) The decomposed components and selected features were reconstructed by SAE, which can filter the redundant noise, reduce the inputs' dimension and improve both the performance and computation time. The results also show it is a good attempt.
(c) There is a horizontal comparison among the proposed model and benchmark models built by other researchers. According to the results, our proposed model is splendid and stable in predicting exchange rates in three datasets.

The remainder of paper has been organized as follows. Section 2 introduces the main procedures of the proposed model and related methods used in this study. Section 3 carries out the empirical study as well as some analysis and discussion. The final part provides the conclusion.

2 Model establishment and evaluation indicators

In this study, the proposed model has three modules which are the preprocessing of target series, feature engineering, and forecasting module. The flowchart of the proposed approach is illustrated in Fig. 1. The detailed steps are described as follows:

(1) Preprocessing of target series: The original exchange rate time series $x_t(t = 1, 2, \ldots T)$ is decomposed into N components, $c_t(k)$, $k = 1, 2, \ldots N$. The N components, $c_t(k)$, as the inputs of SAE, pass through several encoder-decoder processes and transform to M_1 features, $\widehat{c}_t(p), p = 1, 2, \ldots M_1$.
(2) Feature engineering: first, other external variables need to be collected. Assuming there is Z exogenous variables $x_t^z(t = 1, 2, \ldots T; z = 1, 2, \ldots Z)$, SVM-RFE is applied to rank their importance in the prediction process. Then the Q variables x_t^q $(t = 1, 2, \ldots T; q = 1, 2, \ldots Q; Q < Z)$ which get a high score in SVM-RFE are picked out from the Z variables. They will be transformed into M_2 features, $\widehat{x}_t(p), p = 1, 2, \ldots M_2$, by SAE method.
(3) Forecasting module: integrating the reconstructed data, $\widehat{c}_t(p)$ and $\widehat{x}_t(p)$, as a whole $\widehat{g}_t(p), p = 1, 2, \ldots (M_1 + M_2)$. After that, they are partitioned into training data and test data. The training set was used to fit the model and then the test data were forecasted through it. The Bi-LSTM with attention layer uses the features, $\widehat{g}_t(p)$, as the inputs, and

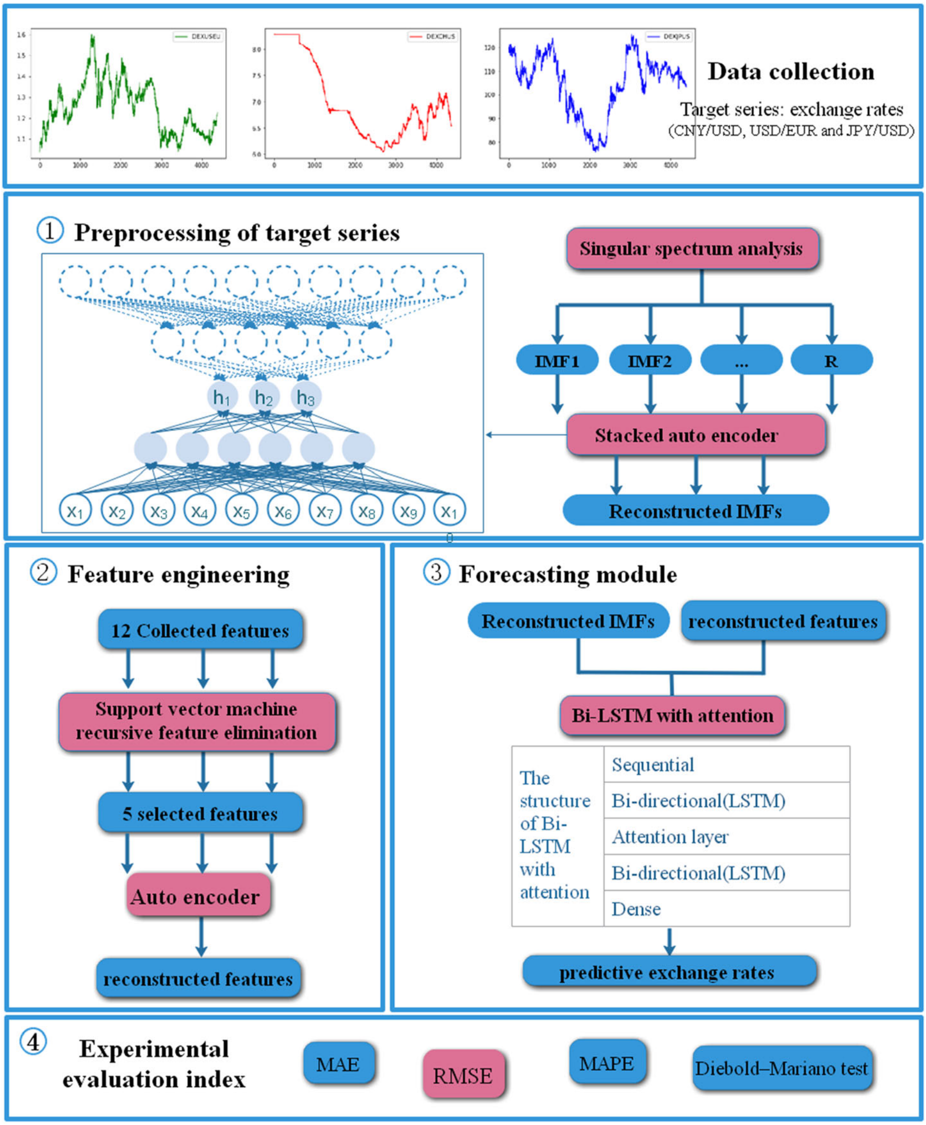

Fig. 1 The flowchart of proposed model

outputs the prediction results h_t. The internal details of the transform process can be seen in the methodology. Fig. 3

2.1 Preprocessing of target series

There are two operators to target series, decomposition and reconstruction. Decomposition of target series can extract signals from the original time series, such as long-term trend signal, periodic signal, and noise signal, which makes predictors easier to capture the nonlinear and non-stationary characteristics of data when they are decomposed. SSA is the decomposition method in the proposed model because SSA can convert data into meaningful components with realistic.

We used reconstruction instead of directly dismissing the noise component because some useful information may be included. SAE was adopted to process the components from SSA, which can remove the correlation, retain the serviceable information and reduce the number of features.

2.1.1 Singular spectrum analysis (SSA)

The SSA is a nonparametric method for time series analysis, no matter whether of short or long series, stationary or non-stationary, and linear or non-linear [24]. It can productively obtain the periodic and trend signals of the original time series [16]. In this paper, it is selected to decompose the exchange rate time series into several modes which will be reconstructed again by stacked auto encoder (SAE). The detailed steps of the SSA are as follows:

(1) Embedding. Suppose the original time series is $Y = [y_1, y_2, \ldots, y_N]$. The first parameter needed to choose is window length L, which is 10 here. The L-lagged vectors are defined as $X = [X_1, X_2, \ldots, X_K]$, where $X_i = [y_i, y_{i+1}, \ldots, y_{i+L-1}]^T$ and $K = N - L + 1$. The matrix X is the trajectory matrix which can be written as follows:

$$X = \begin{bmatrix} y_1 & y_2 & \cdots & y_K \\ y_2 & y_3 & \cdots & y_{K+1} \\ \vdots & \vdots & \cdots & \cdots \\ y_L & y_{L+1} & \cdots & y_N \end{bmatrix} \tag{1}$$

(2) Singular Value Decomposition (SVD). This step is used to decompose the trajectory matrix X. The eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_L$ and eigenvectors $U_1, U_2, \ldots, U_L$ are needed. For convenience, the covariance matrix XX^T will be calculated first. Using $XX^T = U\Sigma\Sigma^T U^T$ where Σ is the matrix only having values (eigenvalues) on the diagonal, we can get the left eigenvector U_i and all eigenvalues by squaring $\Sigma^T\Sigma$. Then the right eigenvector V_i can be obtained through $V_i = X^T U_i \sqrt{\lambda_i}$. The matrix X can be transformed as:

$$X = X_1 + X_2 + \ldots + X_d \text{ where } d = max\{i\}(\lambda_i > 0) \tag{2}$$

$$X_i = \sqrt{\lambda_i} U_i V_i^T \text{ where } X_i \text{ is elementary matrix} \tag{3}$$

(3) Grouping. The interval $\{1, \ldots, d\}$ is separated into discrete subsets $\{I_1, \ldots, I_m\}$ where m is the number of groups determined by the model's performance. The matrix X can be written as:

$$X = X_{I_1} + \ldots + X_{I_m} \tag{4}$$

(4) Diagonal Averaging. This step can produce m time series with N length according to X_I. The elementary matrix X_I is a $L \times K$ matrix with elements x_{ij}. In this paper, L is far smaller than K. The restructured series $(x^1, x^2, \ldots, x^m)$ is defined as follows:

$$x_k^p = \left\{ \begin{array}{c} \frac{1}{k}\sum_{i=1}^{k} x_{i,\ \ k-i+1}\ for\ 1 \le k \le L \\ \frac{1}{L}\sum_{i=1}^{k} x_{i,k-i+1}\ for\ L \le k \le K \\ \frac{1}{N-k+1}\sum_{i=k-K+1}^{N-K+1} x_{i,k-i+1}\ for\ \mathrm{K} \le k \le N \end{array} \right\} for\ p = 1, 2, \dots, m \tag{5}$$

2.1.2 Stacked auto encoder (SAE)

The SAE is a promising approach which can extract more abstract and nonlinear information from original data benefitting from its layer-wise learning framework [3]. Compared with dimensionality reduction techniques and wavelet transform methods, the SAE can save more valuable information without targeted data since it is a unsupervised learning method [36].

A hybrid of multiple auto encoders (AEs) constitutes SAE where AEs are executed one after another [21]. The AE aims to reconstruct the inputs accurately, and it is the data in the hidden layer that we need instead of outputs. The structure of an AE is shown in Fig. 2. It can be seen that the AE has only a hidden layer; all nodes are full connected and the numbers of inputs and outputs are the same. There are two processes in the AE's execution: encoding and decoding. The Eqs. (6) and (7) of these processes are as follows:

$$\mathbf{h} = f_1(\boldsymbol{W_1 x} + \boldsymbol{b_1}) \tag{6}$$

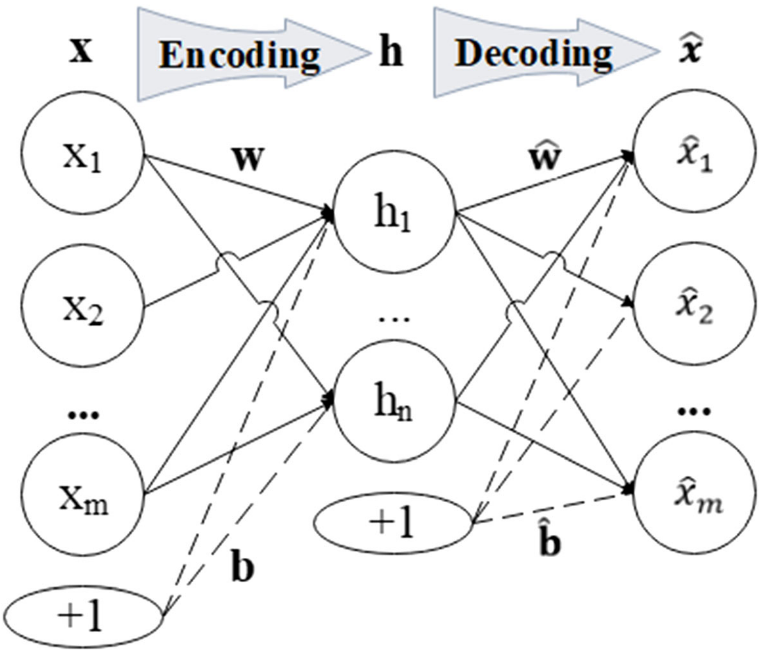

Fig. 2 The structure of AE

$$\widehat{\boldsymbol{x}} = f_2(\boldsymbol{W_2 h} + \boldsymbol{b_2}) \tag{7}$$

where $f_1(.)$ and $f_2(.)$ are active functions ——the sigmoid function, $\boldsymbol{W_1}$ is the encoding matrix, $\boldsymbol{W_2}$ is the decoding matrix, $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ are the bias vectors. The squared loss function is adopted to measure the reconstruction error. By minimizing the loss function $\mathcal{L}(.)$, the model parameters, denoted as θ, can be obtained as in Eq. (8).

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \widehat{x_i}\right)^2 \text{ where N is the number of inputs} \tag{8}$$

Train the first AE on the raw inputs to get the primary feature $\mathbf{h}^{(1)}$ as the second AE's inputs. Similarly, all AEs will be linked up, which is called SAE. The Fig. 3 shows the whole process.

2.2 Feature engineering

The collected features need to be selected and reconstruction for the purpose of removing irrelevant variables, reducing the inputs' dimension and increasing the efficiency. It has already proven that the performance of forecasting model will improve and then decline with the increasing number of relative features. Therefore, when the collected features are filtered by SVM-RFE, reconstructing the selected features instead of discarding the least important ones can obtain more information with the lower dimension. The reconstruction method is AE which is introduced in the SAE part.

2.2.1 Support vector machine recursive feature elimination (SVM-RFE)

SVM-RFE can rank the importance of features and then remove the relatively insignificant features for higher efficiency and accuracy. The data structure complexity and computational time can be reduced if doing so [12].

In the beginning, SVM is applied to solve the binary problem [35]. Suppose there are m features and n observations ($\boldsymbol{x_i}$, y_i), where $y_i \in \{-1, 1\}$, $\boldsymbol{x_i}$ is the feature vector in the ith observation and is the shape of (m × 1). The decision function of SVM can be written as f($\mathbf{x}$)

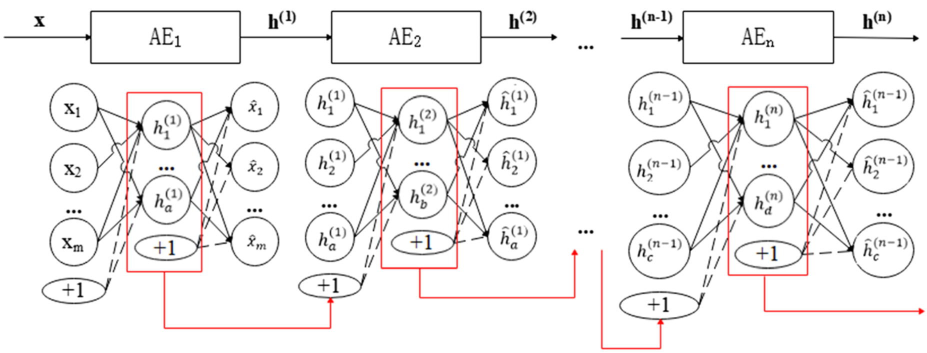

Fig. 3 The structure of SAE

$= \boldsymbol{w}^T\boldsymbol{x} + b$ where $\boldsymbol{w}$ is the weight vector and b is the bias. By using the Lagrange Theory and kernel trick, the optimization of SVM can be expressed as:

$$\min \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k\left(\boldsymbol{x_i}\boldsymbol{x_j}\right) - \sum_{i=1}^{n} \alpha_i \text{s.t.} \sum_{i=1}^{n} \alpha_i y_i = 0 \ \ (0 \le \alpha_i \le C) \tag{9}$$

where C decided by analyst represents the trade-off between accuracy and model complexity. The linear SVM uses the linear kernel $k(\boldsymbol{x_i}\boldsymbol{x_j}) = \boldsymbol{x_i}^T\boldsymbol{x_i}$. For the nonlinear SVM, lots of options can be selected such as Gaussian kernel.

SVM-RFE is a sequential backward feature elimination method and sorted in the back of a ordered feature list [35]. Since $\boldsymbol{w}$ is the linear combination of $\boldsymbol{x_i}$, we can obtain:

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x_i} \tag{10}$$

where α_i is solved by Eq. (9).

The ranking score of the pth feature can be computed as:

$$c_p = w_p^2 \tag{11}$$

The features with least score were discarded. In this paper, an order list is given in Figs. 4, 1 means the most important and the 12 represents the least.

2.3 Forecasting module

After data including the target series and feature are preprocessed, they are inputted into the forecasting model and here it is the Bi-LSTM. For given larger assignments to the more important part in Bi-LSTM, attention layer is added to forecasting model. Finally, the Bi-LSTM with attention mechanism outputs the final prediction results.

2.3.1 Bi-directional long short-term memory model (bi-LSTM)

Bi-LSTM is developed from LSTM which is originated in recurrent neural network (RNN). However, the Bi-LSTM solves the gradient disappearance problem and the lack of long-term memory which cannot be well handled by RNN [40].

Several memory units consist the LSTM neural network. In each memory unit, there is a gating mechanism to determine the information's going or remaining through three gate controllers, namely, "input gate", "forget gate" and "output gate". Within this kind of structure

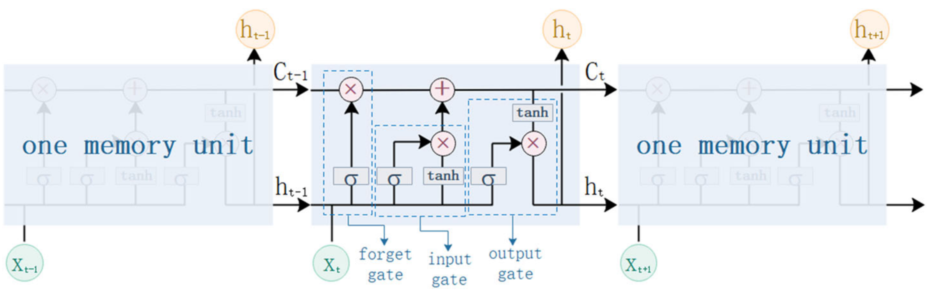

Fig. 4 The structure of LSTM cells: The cross points without arrows mean copy. The first horizontal line denotes the cell state

shown in Fig. 4, the LSTM processes the previous unit output (h_{t-1}), current unit input (x_t), and previous unit status (C_{t-1}) by discarding and adding to get the current unit output (h_t) and status (C_t).

Basic calculations of LSTMs can be separated to four steps as follows:

(1) The input of unit, x_t, and the output of previous unit, h_{t-1}, firstly pass through "forget gate" (f_t). The output of "forget gate" is obtained through the "sigmoid" activation function, represented as σ in Fig. 3. The value of f_t is between zero and one. When f_t is 1, all the information of the inputs will be remained. When f_t is 0, the whole information of x_t and h_{t-1} will be given up. The formula can be seen in Eq. (12) where W_f is the linear relationship coefficient matrix and b_f is the bias vector.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{12}$$

(2) The "input gate", i_t, is also generated by a sigmoid function. Besides, a new candidate cell state $\widetilde{C}_t$ is produced through a hyper tangent function, denoted as "tanh":

$$\widetilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{13}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{14}$$

(3) Some information in previous cell state C_{t-1} and the new candidate cell state $\widetilde{C}_t$ is chosen by "forget gate" and "input gate" respectively. The cell state is updated, denoted as C_t. The * means elements multiplication.

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{15}$$

(4) The "output gate", o_t, controls the amount of information filtered out from the cell state, which multiply the activated C_t to obtain the output h_t.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{16}$$

$$h_t = o_t * \tanh(C_t) \tag{17}$$

Bi-LSTM is consisted of two forward and backword LSTM layers and its structure is shown in Fig. 5. The two hidden layers run in opposite directions and are concatenated to the same output layer so the neural network can extract bidirectional sequential relationships that exist in the time series data.

2.3.2 Attention mechanism

The attention model can use limited attention resources to obtain key information from a large amount of data. Integrating the attention mechanism into Bi-LSTM model can allocate limited resources more to key parts of the input sequence, which can weaken the redundant information and emphasize the vital part.

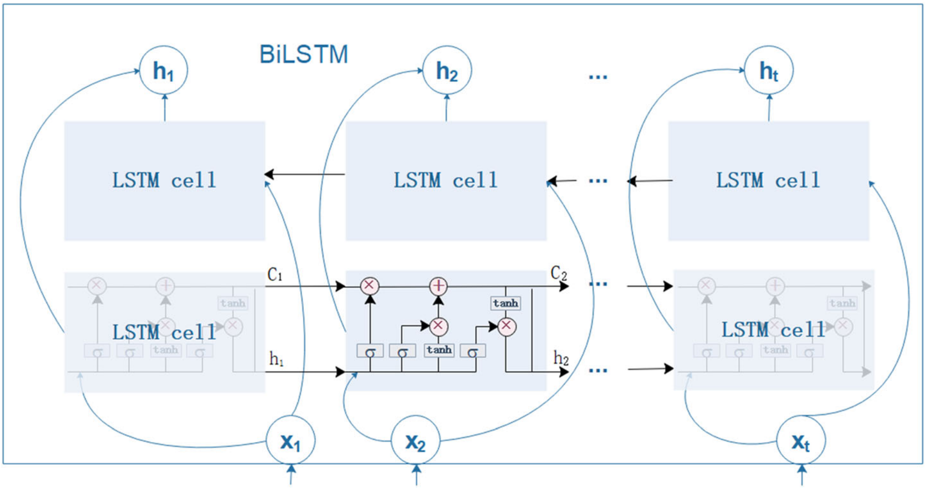

Fig. 5 The structure of Bi-LSTM

$x^{(t)} = \{x_1, x_2, \dots, x_t\}$ presents the input data $y^{(t)} = \{y_1, y_2, \dots, y_t\}$ denotes the output $w^{(n,t)}$ is the attention weight value of each feature, $w^{(n,t)}$ is calculated as follow:

$$w^{(n,t)} = \frac{exp\left(e^{(n,t)}\right)}{\sum_{t=1}^{T} exp\left(e^{(n,t)}\right)} \tag{18}$$

2.4 Evaluation indicators

This paper uses three evaluation indicators to measure the performance. They are MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error) and RMSE (Root Mean Squares Error). Their definitions are shown as below:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{19}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100\% \tag{20}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{21}$$

where n is the number of train or test samples, y_i is the actual value and $\hat{y}_i$ is the prediction value.

Diebold–Mariano (DM) test is also conducted to verify the significant difference between the proposed system in this study and other prediction models. If the null hypothsis is refused, then the two models have significant difference. The steps of DM test is as follows: first two error sequences (the difference between the actual values and the predictive values) are calculated by using the proposed model (e^1) and the comparative model (e^2). Then the loss function f(·) can use one of the Eqs. 19–22. Finally, DM statistics is calculated by the Eq. 22 where S^2 is the variance estimation of f(e^1) − f(e^2) and T is the numble of samples.

$$1/T\sum_{n=1}^{T}\left(\boldsymbol{f}\left(e^1\right)-\boldsymbol{f}\left(e^2\right)\right)/\sqrt{S^2/T} \tag{22}$$

3 Empirical analysis

3.1 Data analysis

In this study, the data set of daily CNY/USD, USD/EUR and JPY/USD exchange rates are obtained from FRED ECONOMIC DATA (https://fred.stlouisfed.org/). The entire data set covers from January 2, 2003 to December 18, 2020, yielding a total of 4380 observations, which excludes missing values. The data was separated into two sets, training and test data, which accounts for 80% and 20% respectively. The training set begins from January 2, 2003 to May 9, 2017, generating a total of 3504 observations, while the test set starts from May 10, 2017 to December 18, 2020 with a number of 876 observations. The Fig. 6 illustrate the original daily exchange rates respectively. It is obvious that the three exchange rates have different trends and volatility. The USD/EUR and JPY/USD exchange rates both show relatively larger fluctuations as opposed to CNY/USD exchange rate. In the beginning of this data set the CNY/USD exchange rate reached its peak (8.28 Chinese yuan/U.S. dollar) and kept almost unchanged for around six hundred days, while Japanese Yen and euro had been appreciated relative to the dollar during the same period. The CNY/USD exchange rate had been significantly decreased to the bottom from July 22, 2005 to January 14, 2014, while the JPY/USD and EUR/USD both experienced a trough and then return the origin during the same time. Form that time on, there are a large amount of fluctuations for CNY/USD exchange rate not only in the general trend but in the short term. One of the reasons is perhaps that China was prepared to attend IMF and also intended to carry forward the marketization of Chinese yuan.

The descriptive statistics of exchange rates is shown in Table 1. It can be seen that the standard deviation is pretty small for all of them, indicating the less volatile exchange rates. According to the results of Jarque-Bera Normality Test, the empirical distribution of daily exchange rates studied herein does not exhibit the appearance of normal distribution. The ADF test is exhibited in Table 2, which can measure whether a sequence is stable or not. All the critical values are smaller than 1%, 5% and 10% T-statistic and the p value is approximately 44%, 13% and 31% respectively, so we cannot reject the null hypothesis that the series is non-

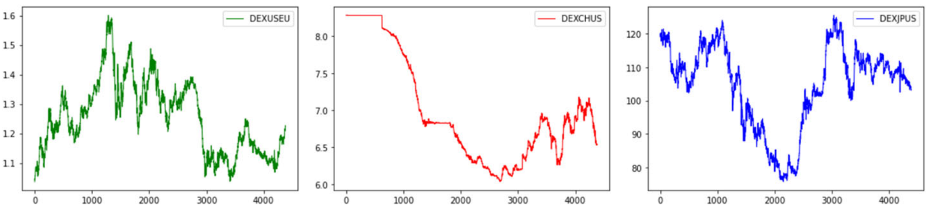

Fig. 6 The original daily exchange rates

Table 1 Descriptive statistics of daily exchange rate

	Mean	Median	Min	Max	Std.	Skewness	Kurtosis	Normality Test
CNY/USD	7.0007	6.8273	6.0402	8.2800	0.72150	0.685524	−0.89046	487.6466
USD/EUR	1.2539	1.2504	1.0361	1.6010	0.12297	0.381569	−0.55479	162.5337
JPY/USD	104.66	107.79	75.720	125.58	12.5182	0.747838	−0.37680	434.0226

Std. means standard deviation; the normal test is Jarque-Bera test

Table 2 The ADF test results of exchange rate series

	T-statistic	Prob.	1% T-statistic	5% T-statistic	10% T-statistic
CNY/USD	−1.675	0.443	−3.431	−2.862	−2.567
USD/EUR	−2.429	0.133	−3.431	−2.862	−2.567
JPY/USD	−1.937	0.314	−3.431	−2.862	−2.567

stationary. The results of BDS test are shown in the Table 3 and we can see that all the p-values are nearly zero when the embedding dimension is 2–6. Thus it can support that the exchange rate series is nonlinear. Additionally, in this paper 12 exogenous variables' data were collected from FRED ECONOMIC DATA and they take part in the predictive process, as well. The Table 4 shows their short names and brief introductions.

3.2 Parameter setting

In the proposed model, SVM-RFE selected five import variables added into the prediction process. The SAE for decomposed exchange rates' reconstruction has two hidden layers (one has five neurons and the other has three neurons.) The SAE for external variables' reconstruction has one hidden layers with three neurons. Time step is always ten. The forecasting model has two Bi-LSTM layers and one attention layer. The neurons in two Bi-LSTM layers are eight and four respectively and the activation function in the attention layer is relu.

3.3 Empirical results and discussion

In this study, the multiscale proposed model was applied to predict the CNY/USD, USD/EUR, and JPY/USD exchange rates. In order to measure its performance, other six models were proposed for comparison. Three of them are single models (ARIMA, BPNN and LSTM),

Table 3 The BDS test results of exchange rate series

	CNY/USD		USD/EUR		JPY/USD	
Dimension	BDS-statistics	Prob.	BDS-statistics	Prob.	BDS-statistics	Prob.
2	17.283	0.0	13.7520	0.0	12.543647	0.0
3	21.504	0.0	16.8256	0.0	15.21852	0.0
4	27.144	0.0	20.7705	0.0	18.59601	0.0
5	34.761	0.0	25.87594	0.0	22.8781	0.0
6	45.099	0.0	32.54341	0.0	28.36295	0.0

Table 4 The variables' names appeared in this paper and their brief introductions

No.	Trading code	Description
1	DEXCHUS	China / U.S. Foreign Exchange Rate
2	DCOILBRENTEU	Crude Oil Prices: Brent - Europe
3	DCOILWTICO	Crude Oil Prices: West Texas Intermediate (WTI) - Cushing, Oklahoma
4	GOLDAMGBD228NLBM	Gold Fixing Price 10:30 A.M. (London time) in London Bullion Market, based in U.S. Dollars
5	DHHNGSP	Henry Hub Natural Gas Spot Price
6	T5YIE	5-Year Breakeven Inflation Rate
7	T10YIE	10-Year Breakeven Inflation Rate
8	T5YIFR	5-Year, 5-Year Forward Inflation Expectation Rate
9	DFII10	10-Year Treasury Inflation-Indexed Security, Constant Maturity
10	DGS10	10-Year Treasury Constant Maturity Rate
11	USD1MTD156N	1-Month London Interbank Offered Rate (LIBOR), based on U.S. Dollar
12	DGS1	1-Year Treasury Constant Maturity Rate
13	DFF	Effective Federal Funds Rate

which can represent the traditional statistical model, erlay single layer neural network, and the deep learning model respectively. The rest comparative models were proposed by the other researchers, which were used here as benchmark models to verify the superiority of our proposed model. The performance of these models are illustrated in Table 5. The best results are shown in bold. Figures 7, 8 and 9 depict the actual exchange rates and the one-step predictive values forecasted by the proposed model in three datasets. In order to clearly exhibit the prediction ability, we zoomed part of the predictive values in the big picture and the small picture shows the whole testing set.

Table 5 The performance of different models for daily CNY/USD exchange rate

Datasets	Model	MAE	RMSE	MAPE(%)	DW statistics
Daily CNY/USD exchange rate	ARIMA	0.19825	0.2615	2.9994	−19.29
	BPNN	0.04836	0.0585	0.7285	−19.66
	LSTM	0.03122	0.0371	0.4571	−10.07
	Bootstrap LSTM [29]	0.02711	0.0332	0.3977	−6.34
	WT-SAE-LSTM [21]	0.02624	0.0347	0.3877	−6.98
	Multi-feature GRU [9]	0.26124	0.3238	3.8429	−20.83
	Proposed model	**0.02117**	**0.0268**	**0.3123**	
Daily USD/EUR exchange rate	ARIMA	0.06294	0.0745	5.3604	−27.35
	BPNN	0.01353	0.0145	1.1861	−42.33
	LSTM	0.00396	0.0052	0.3435	−15.87
	Bootstrap LSTM [29]	0.00660	0.0084	0.5767	−17.6
	WT-SAE-LSTM [21]	0.00512	0.0067	0.4453	−15.02
	Multi-feature GRU [9]	0.12833	0.1512	11.2025	−29.25
	Proposed model	**0.00107**	**0.0012**	**0.0935**	
Daily JPY/USD exchange rate	ARIMA	1.78959	2.2582	1.6511	−19.81
	BPNN	0.35790	0.4858	0.3277	−3.59
	LSTM	0.36511	0.4997	0.3348	−3.88
	Bootstrap LSTM [29]	0.54074	0.7075	0.4958	−10.44
	WT-SAE-LSTM [21]	3.05190	3.8496	2.8019	−17.58
	Multi-feature GRU [9]	4.34771	5.3211	3.9973	−21.25
	Proposed model	**0.32708**	**0.4182**	**0.2999**	

The bold number represents the lowest error level in each dataset

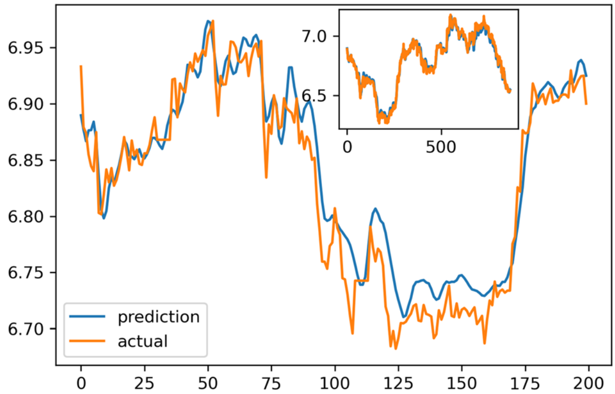

Fig. 7 The prediction results of the proposed model in daily CNY/USD exchange rates

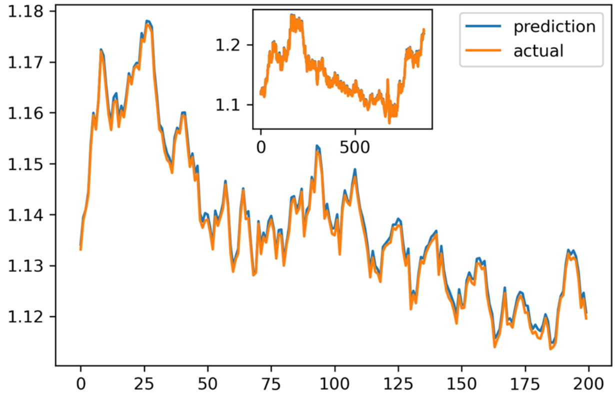

Fig. 8 The prediction results of the proposed model in daily USD/EUR exchange rates

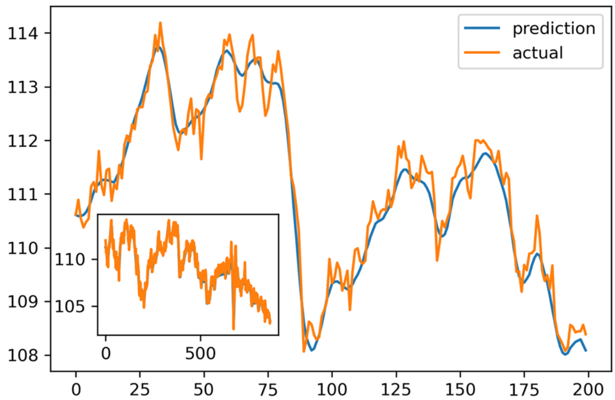

Fig. 9 The prediction results of the proposed model in daily JPY/USD exchange rates

It can be obviously seen that in three datasets our proposed model always has the smallest error and the best prediction performance, and this not only presents its splendid prediction ability and also verify its stability and robustness. Taking the CNY\USD exchange rates as the example, the MAE, RMSE, and MAPE of the proposed model are 0.02117, 0.0268, and 0.3123% accordingly, which are improved almost 90% compared to the multi-feature GRU, the worst prediction model in all benchmark models in this dataset. In addition, according to the DW statistics, all the comparative models have the significant difference to the proposed model because none of them are larger than −2.76 (the t statistic at 1% significant level).

As we expected, the representation of traditional statistical models, ARIMA, shows the worst ability to predict short-term exchange rates because exchange rates cannot satisfy its inherent assumptions which requires data are linear and stationary. The MAPEs of ARIMA in three datasets are 2.99%, 5.36%, and 1.65% respectively, while those are 0.72%, 1.18%, and 0.32% in BPNN, a representative AI algorithm. From that it can be discovered that BPNN can well capture the nonlinear characteristic of exchange rates compared to ARIMA. However, it is not enough for a single neural network to capture the change of exchange rates. There is a slight increase of prediction accuracy for LSTM which solves the gradient explosion problem and utilizes the long-term historical information opposite to simple neural network. In the prediction of USD/CNY exchange rates, LSTM improved the accuracy by over 35% compared to BPNN, and in USD/EUR dataset, the improved percentage even achieves 70%.

The various hybrid models were built in recent years and some of them which were proposed by others are employed in this study for examining the effectiveness of our proposed model. The bootstrap LSTM is built several LSTM models with the bootstrap technique to expand the training set and then average the results of all LSTM. According to the results in Table 5, this framework can improve the prediction performance to some extent. For example, in USD/CNY exchange rates, the three evaluation indicators of LSTM are 0.031, 0.037, and 0.45%, while those of bootstrap LSTM illustrate an obvious decrease (0.027, 0.033, and 0.39% accordingly). However, in other two datasets, the bootstrap LSTM has an inferiorer performance than LSTM, which demonstrates its unstablity.

WT-SAE-LSTM is a decomposition reconstruction and then forecasting framework where WT is wavelet transform. The prediction performance of WT-SAE-LSTM is suboptimal among the hybrid methods conducted in this study in USD/CNY and USD/EUR exchange rates (its MAPEs are 0.38% and 0.44%), only inferior to our proposed model (0.31% and 0.09%). It shows that decomposition and reconstruction by SAE can achieve better effects for extracting signals, reducing noise signal residual, saving more useful information and reducing the dimension. The similar steps (decomposition and reconstruction by SAE) are also used in preprocessing of target series part in our proposed model, and the empirical results can also prove the importance of feature engineering which exist in the proposed model but not exist in WT-SAE-LSTM.

Multi-feature GRU is used three feature selection algorithms to select the useful external variables and then the selected features are fed into GRU as the forecasting part. The historical data of target series are not fed into multi-feature GRU and as we known the historical data contains a large amount of useful information. This is why the performance of multi-feature GRU is even not as good as ARIMA.

All in all, the new model proposed in this study show the lowest error level in three datasets and statistical test also verifies the significant difference between the proposed model and comparative models at 1% level.

4 Conclusions

This paper proposes a new hybrid model to predict the exchange rate accurately and efficiently. There are three modules in the new hybrid model, which are preprocessing of the target series, feature engineering and forecasting module. In the preprocessing of the target series, the SSA method is combined with the SAE tool for uniting their respective advantages. Feature engineering consists of SVM-RFE for selection and SAE for reconstruction, which can filter the related external variables, reduce the features' dimension, and save the useful information. The forecasting module is employed Bi-LSTM with attention mechanism, an emerging algorithm. The empirical results also show its superior than single LSTM or GRU. According to the results our new model outperforms other benchmark models in all samples.

This study has several implications. Firstly, apart from historical data, feature engineering is also vital to influence the performance of model. Appropriate feature engineering can largely increase the accuracy but inappropriate one can even make the performance worse. Secondly, the SSA combined with SAE is a worth trying since it can make original data regular, filter the noise, reduce the demision, and still save the notable information. The empirical experiments also show the outperformance of this combination in this study. Thirdly, the hybrid model proposed in this study produces the best performance among all the compared models, which proves the feasibility and potential of the novel model.

The novel hybrid model provides policymakers and investors with a new exchange rate predicting tool. For bankers and merchants, it is also worthy of attention if they want expand their overseas business. However, the number of compared models mentioned in this study is not too much. There is probably larger potential to develop the proposed model. In addition, even though exchange rate is an old topic, it will not be out dated since it is closely related to our economic life and is also an important indicator. Thus, it is necessary for us to keep attention to it. Moreover, even though the novel hybrid model proposed here has a good prediction performance, there are some improvable potential. For example, the fluctuations of exchange rates are dependent on the central bank's policy to some extent. Thus, the big event and policies can also be considered when we predict exchange rates.

Acknowledgments This research was supported by the National Natural Science Foundation of China (Grant No. 71971122 and 71501101).

Data availability The datasets analysed during the current study are openly available in the FRED ECONOMIC DATA, [https://fred.stlouisfed.org/].

Declarations

Competing interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Abhyankar A, Sarno L, Valente G (Jul. 2005) Exchange rates and fundamentals: evidence on the economic value of predictability. J Int Econ 66(2):325–348. https://doi.org/10.1016/j.jinteco.2004.09.003

2. Amo Baffour A, Feng J, Taylor EK (Nov. 2019) A hybrid artificial neural network-GJR modeling approach to forecasting currency exchange rate volatility. Neurocomputing 365:285–301. https://doi.org/10.1016/j.neucom.2019.07.088
3. Bai Y, Bezak N, Sapač K, Klun M, Zhang J (2019) Short-term streamflow forecasting using the feature-enhanced regression model. Water Resour. Manag, p. 15
4. Barunik J, Krehlik T, Vacha L (May 2016) Modeling and forecasting exchange rate volatility in time-frequency domain. Eur J Oper Res 251(1):329–340. https://doi.org/10.1016/j.ejor.2015.12.010
5. Cai Z, Chen L, Fang Y (2012) A new forecasting model for USD/CNY exchange rate. Stud Nonlinear Dyn Econom 16(3):1878. https://doi.org/10.1515/1558-3708
6. Chortareas G, Jiang Y, Nankervis JC (2011) Forecasting exchange rate volatility using high-frequency data: is the euro different? Int J Forecast 27(4):1089–1107. https://doi.org/10.1016/j.ijforecast.2010.07.003
7. de O Santos Júnior DS, de Oliveira JFL, de Mattos Neto PSG (2019) An intelligent hybridization of ARIMA with machine learning models for time series forecasting. Knowl-Based Syst 175: 72–86. https://doi.org/10.1016/j.knosys.2019.03.011.
8. Dunis CL, Huang X (2002) Forecasting and trading currency volatility: an application of recurrent neural regression and model combination. J Forecast 21(5):317–354. https://doi.org/10.1002/for.833
9. Fan G-F, Yu M, Dong S-Q, Yeh Y-H, Hong W-C (2021) Forecasting short-term electricity load using hybrid support vector regression with grey catastrophe and random forest modeling. Util Policy 73:101294. https://doi.org/10.1016/j.jup.2021.101294
10. Galeshchuk S (Jan. 2016) Neural networks performance in exchange rate prediction. Neurocomputing 172: 446–452. https://doi.org/10.1016/j.neucom.2015.03.100
11. He K, Chen Y, Tso GKF (2018) Forecasting exchange rate using Variational mode decomposition and entropy theory. Phys Stat Mech Its Appl 510:15–25. https://doi.org/10.1016/j.physa.2018.05.135
12. Huang M-L, Hung Y-H, Lee WM, Li RK, Jiang B-R (2014) SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. Sci World J 2014:10. https://doi.org/10.1155/2014/795624
13. Huang Y, Shen L, Liu H (Feb. 2019) Grey relational analysis, principal component analysis and forecasting of carbon emissions based on long short-term memory in China. J Clean Prod 209:415–423. https://doi.org/10.1016/j.jclepro.2018.10.128
14. Li J, et al. (2021) A novel hybrid short-term load forecasting method of smart grid using MLR and LSTM neural network. IEEE Trans Ind Inform 14(4):2443
15. Li Y, Jiang S, Li X, Wang S (Mar. 2021) The role of news sentiment in oil futures returns and volatility forecasting: data-decomposition based deep learning approach. Energy Econ 95:105140. https://doi.org/10.1016/j.eneco.2021.105140
16. Liu H (2019) Smart wind speed deep learning based multi-step forecasting model using singular spectrum analysis, convolutional Gated Recurrent Unit network and Support Vector Regression. Renew Energy, p. 13
17. Lubecke TH, Nam KD, Markland RE, Kwok CCY (1998) Combining foreign exchange rate forecasts using neural networks. Glob Finance J 9(1):5–27. https://doi.org/10.1016/S1044-0283(98)90012-6
18. Medeiros MC, Veiga A, Pedreira CE (Jul. 2001) Modeling exchange rates: smooth transitions, neural networks, and linear models. IEEE Trans Neural Netw 12(4):755–764. https://doi.org/10.1109/72.935089
19. Meese RA, Rogoff K (Feb. 1983) Empirical exchange rate models of the seventies. J Int Econ 14(1–2):3–24. https://doi.org/10.1016/0022-1996(83)90017-X
20. Peng T, Zhang C, Zhou J, Nazir MS (Apr. 2021) An integrated framework of bi-directional long-short term memory (BiLSTM) based on sine cosine algorithm for hourly solar radiation forecasting. Energy 221: 119887. https://doi.org/10.1016/j.energy.2021.119887
21. Qiao W, Yang Z (2020) Forecast the electricity price of U.S. using a wavelet transform-based hybrid model. Energy 193:116704. https://doi.org/10.1016/j.energy.2019.116704
22. Raji CG, Chandra SSV (2017) Long-term forecasting the survival in liver transplantation using multilayer perceptron networks. IEEE Trans Syst Man Cybern Syst 47(8):2318–2329. https://doi.org/10.1109/TSMC.2017.2661996
23. Rapach DE, Wohar ME (Apr. 2006) The out-of-sample forecasting performance of nonlinear models of real exchange rate behavior. Int J Forecast 22(2):341–361. https://doi.org/10.1016/j.ijforecast.2005.09.006
24. Rodrigues PC, Mahmoudvand R (2020) A new approach for the vector forecast algorithm in singular spectrum analysis. Commun Stat - Simul Comput 49(3):591–605. https://doi.org/10.1080/03610918.2019.1664578
25. Sermpinis G, Dunis C, Laws J, Stasinakis C (Dec. 2012) Forecasting and trading the EUR/USD exchange rate with stochastic neural network combination and time-varying leverage. Decis Support Syst 54(1):316–329. https://doi.org/10.1016/j.dss.2012.05.039

26. Sermpinis G, Stasinakis C, Theofilatos K, Karathanasopoulos A (Dec. 2015) Modeling, forecasting and trading the EUR exchange rates with hybrid rolling genetic algorithms—support vector regression forecast combinations. Eur J Oper Res 247(3):831–846. https://doi.org/10.1016/j.ejor.2015.06.052
27. Shen F, Chao J, Zhao J (Nov. 2015) Forecasting exchange rate using deep belief networks and conjugate gradient method. Neurocomputing 167:243–253. https://doi.org/10.1016/j.neucom.2015.04.071
28. Sun S, Wang S, Wei Y, Zhang G (2018) A clustering-based nonlinear ensemble approach for exchange rates forecasting. IEEE Trans Syst Man Cybern Syst:1–9. https://doi.org/10.1109/TSMC.2018.2799869
29. Sun S, Wang S, Wei Y (Oct. 2020) A new ensemble deep learning approach for exchange rates forecasting and trading. Adv Eng Inform 46:101160. https://doi.org/10.1016/j.aei.2020.101160
30. Tyree EW, Long JA (1995) Forecasting currency exchange rates: Neural networks and the random walk model, in Forecasting Currency Exchange Rates: Neural Networks and the Random Walk Model, Wall Street New York, p. 11
31. Wang J, Cheng Q, Sun X (2021) Carbon price forecasting using multiscale nonlinear integration model coupled optimal feature reconstruction with biphasic deep learning. Environmental Science and Pollution Research, https://doi.org/10.1007/s11356-021-16089-2
32. Wang C, Wang X, Zhang J, Zhang L, Bai X, Ning X, Zhou J, Hancock E (Apr. 2022) Uncertainty estimation for stereo matching based on evidential deep learning. Pattern Recogn 124:108498. https://doi.org/10.1016/j.patcog.2021.108498
33. Wei Y, Sun S, Ma J, Wang S, Lai KK (2019) A decomposition clustering ensemble learning approach for forecasting foreign exchange rates. J Manag Sci Eng 4(1):45–54. https://doi.org/10.1016/j.jmse.2019.02.001
34. West KD, Cho D (Oct. 1995) The predictive ability of several models of exchange rate volatility. J Econom 69(2):367–391. https://doi.org/10.1016/0304-4076(94)01654-I
35. Xue Y, Zhang L, Wang B, Zhang Z, Li F (2018) Nonlinear feature selection using Gaussian kernel SVM-RFE for fault diagnosis. Appl Intell. https://doi.org/10.1007/s10489-018-1140-3
36. Yang H-F, Chen Y-PP (2019) Representation learning with extreme learning machines and empirical mode decomposition for wind speed forecasting methods. Artif Intell 277:103176. https://doi.org/10.1016/j.artint.2019.103176
37. Yang H-L, Lin H-C (2017) Applying the hybrid model of EMD, PSR, and ELM to exchange rates forecasting. Comput Econ 49(1):99–116. https://doi.org/10.1007/s10614-015-9549-9
38. Yang S, Chen D, Li S, Wang W (2020) Carbon price forecasting based on modified ensemble empirical mode decomposition and long short-term memory optimized by improved whale optimization algorithm. Sci Total Environ 716:137117. https://doi.org/10.1016/j.scitotenv.2020.137117
39. Yu L, Lai KK, Wang S (2008) Multistage RBF neural network ensemble learning for exchange rates forecasting. Neurocomputing 71(16):3295–3302. https://doi.org/10.1016/j.neucom.2008.04.029
40. Zhang Q, Zhang J (2020) Short-term load forecasting method based on EWT and IDBSCAN. J Electr Eng Technol 15(2):635–644. https://doi.org/10.1007/s42835-020-00358-0
41. Zhang G, Eddy Patuwo B, Hu MY (1998) Forecasting with artificial neural networks: The state of the art. Int J Forecast 14(1):35–62. https://doi.org/10.1016/S0169-2070(97)00044-7
42. Zhou F, Huang Z, Zhang C (Apr. 2022) Carbon price forecasting based on CEEMDAN and LSTM. Appl Energy 311:118601. https://doi.org/10.1016/j.apenergy.2022.118601
43. Zhu J, Liu J, Wu P, Chen H, Zhou L (2019) A novel decomposition-ensemble approach to crude oil price. Int J Mach Learn Cybern. no. 10, pp. 3349–3362. https://doi.org/10.1007/s13042-019-00922-9.