Development of a Stock Price Prediction Model Integrating LSTM, SVR in Deep Learning and BLS

Hao Wang

Mathematics Department, Southwestern University of Finance and Economics, Chengdu Sichuan, 610000, China
20232660@bistu.edu.cn

Abstract. Stock market aids investors in making wiser choices. However, the complexity and uncertainty of the financial market result in existing prediction models falling short of expectations. Existing Long Short-Term Memory (LSTM) models commonly suffer from lagging issues, while Support Vector Regression (SVR) models tend to overfit. To enhance prediction accuracy, this paper introduces a novel approach that utilizes the advantages of LSTM and SVR separately to handle low-frequency and high-frequency components, respectively, after denoising using Complete Ensemble Empirical Mode Decomposition (CEEMD). By applying Singular Spectrum Analysis (SSA) to remove noise from high-frequency components obtained from CEEMD, prediction precision is improved. Furthermore, Broad Learning System (BLS) is introduced to mitigate the overfitting risk in LSTM, thereby enhancing model stability and generalization capability. This enables the model to excellently predict turning points and high-frequency fluctuations. The effectiveness of the proposed CEEMD-SSA-LSTM-SVR-BLS model in stock price prediction is demonstrated. This comprehensive approach addresses challenges posed by stock market volatility, thereby enhancing the reliability and applicability of stock prediction models. The model proposed in this paper can provide valuable advice to a wide range of investors, offering significant assistance in stock selection and trading decisions.

Keywords: Complete Ensemble Empirical Mode Decomposition, Broad Learning System, Support Vector Regression.

1 Introduction

Stocks which serve as a barometer can intuitively reflect the economic conditions of a country. Meanwhile the operational profitability and future development of companies which are in various industries can be showcased in the stock market. After World War II, the unprecedented prosperity of the US economy [1] led its stock market to naturally evolve into the world's largest and most actively traded stock market. Coupled with investments offering high risks and returns, attracted an

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_10

increasing number of investors and financial institutions. Predicting stock market fluctuations is necessary for investors, as reasonable predictions can significantly reduce investor risks.Therefore, how to obtain higher returns in the stock market has been a topic of discussion among investors.

In the early days of stock prediction research, researchers mostly relied on fundamental analysis and trendline analysis, which later evolved into the use of statistical models to predict stock trends, such as Charles Dow first applied Dow theory in 1884 to create an index comprising the averages of 11 stock prices [2],and Shiab utilizes Auto-Regressive Moving Average Model to analyze the predictive feasibility of Amman stock [3]. Yet, with the continuous expansion of the number of stocks and the stock market, traditional techniques struggled to efficiently handle vast amounts of data to meet market demands. With the rise of computers, new computer technologies like machine learning, deep learning, etc., have been applied to stock prediction, such as Long Short-Term Memory (LSTM) [4], Support Vector Regression (SVR) [5], Recurrent Neural Network (RNN) [6], etc. However, facing the non-linearity [7], non-stationary, and other characteristics of stocks, different methods have their pros and cons.

LSTM networks have been widely applied in financial forecasting, but the presence of noise [8] in financial sequences makes it difficult to achieve precise predictions. Many scholars have derived more accurate stock prediction models based on LSTM, such as the Empirical Mode Decomposition (EMD)-LSTM [9], which optimizes the LSTM single model well. However, due to the limitations of EMD in handling complex stock price data and its inability to adapt well to the dynamic characteristics of data, it affects the accuracy of predictions. ZHANG and his colleagues [10] found through experiments that the Complete Ensemble Empirical Mode Decomposition (CEEMD)-LSTM model can better predict stock index returns compared to EMD-LSTM. However, even with CEEMD, noise may still exist in high-frequency components, This algorithm needs further improvement.

In this research context, the paper aims to explore the potential application of combining CEEDM, Singular Spectrum Analysis (SSA) [11], and Broad Learning System (BLS) [12] algorithms in stock prediction. This paper proposes a combination prediction model of CEEMD-SSA-LSTM-SVR-BLS for stocks, using the opening prices of 1300 stocks from various industries listed on the New York Stock Exchange for experimental verification. This paper applies SSA to remove noise from several high-frequency components obtained from CEEMD to improve prediction accuracy. Meanwhile, due to the overfitting risk of LSTM in predictions, causing instability in future predictions, this paper introduces BLS to reduce overfitting risk while enhancing the overall applicability and generalization ability of the model. Additionally, due to the outstanding robustness of BLS, when there is significant price volatility in stocks, BLS can handle it better, improving the stability and reliability of the model.

2 Research Methodology

2.1 Data Source

The dataset selected for this study comprises price data of 2000 stocks listed on the New York Stock Exchange obtained from the Kaggle database. The time frame ranges from July 21, 2008, to October 11, 2017. The dataset contains the opening price, closing price, trading volume, as well as the highest and lowest prices during the trading session for stocks. Non-trading days were excluded from the dataset, resulting in a total of 1267 data sets. Subsequently, 90% of the data was chosen as the training set, and 10% as the testing set. To further improve the predictive accuracy of the models, normalization preprocessing was conducted before empirical analysis. The normalization formula is as follows:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Where x represents the original sequence, x_{max} represents the maximum value of the original sequence, and x_{min} represents the minimum value of the original sequence.

2.2 Fundamental Theory

To better accommodate the dynamic features of the data, this paper initially utilizes the CEEMD-LSTM model. Subsequently, in order to mitigate the noise effects of high-frequency residual components, the SSA model is integrated to bolster model performance. Following this, the SVR and BLS models are introduced to enhance overall generalization capability, code execution speed, and prediction accuracy.

CEEMD is a technique used for signal processing and data analysis, aimed at handling nonlinear and non-stationary signals. It involves adding a series of white noise signals with opposite signs to the original signal to alter the distribution of extreme points in the signal, followed by EMD of the signal. The steps of CEEMD decomposition are as follows:

Adding multiple groups of white noise signals with opposite signs to the original signal, each group having the same amplitude. Specifically, for each group, positive and negative noise signals are added as follows:

$$m_i^+(t) = x(t) + n_i^+(t) \tag{2}$$

$$m_i^-(t) = x(t) + n_i^-(t) \tag{3}$$

Performing EMD decomposition on m_i^+ and m_i^- to obtain a series of Intrinsic Mode

Function (IMF) components. The IMF obtained from the decomposition must satisfy two conditions:

The number of extremum points (including maximum and minimum values) in each IMF should be equal or differ by no more than one. This ensures relative stability in the number of extremum points in each IMF.

The mean of each IMF over the entire time series should be close to zero, indicating oscillations in both positive and negative directions without significant DC offset. The decomposition process is based on a method called "ensemble removal iteration," where in each iteration, CEEMD identifies and extracts the main oscillatory patterns from the signal and removes them to obtain the IMF. These oscillatory patterns, represented by local extremum points (maxima and minima), are removed in each iteration. The resulting IMFs are then categorized into two groups based on whether positive or negative noise was added. IMF1 represents the ensemble average result with positive noise added, while IMF2 represents the ensemble average result with negative noise added.

The ensemble average of the two sets of IMF components constitutes the final decomposition result.

LSTM networks are based on the principle of "gate mechanisms," which control the flow of information to address the issue of long-term dependencies in traditional RNNs. Hu has provided an excellent explanation of the working principle of LSTM and its outstanding capability in capturing long-term dependency information in sequential data [13]. Meanwhile, Sun's research demonstrates the high efficiency of LSTM in successfully predicting trends in the stock market index prices, further validating the superior performance of LSTM[14].

SSA is a non-parametric method used for signal processing and data analysis, aimed at decomposing sequence data and extracting components such as trend, periodicity, and noise. The basic steps of SSA are:

Embedding Data: Embedding a time series of length N into a matrix X, where K is the embedding window size, and $L = N - K + 1$ is the number of columns of the embedded matrix.

$$\begin{matrix} x_1 & x_2 & \cdots & x_L \\ x_2 & x_3 & \cdots & x_{L+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_K & x_{K+1} & \cdots & x_N \end{matrix} \tag{4}$$

Singular Value Decomposition (SVD): Performing Singular Value Decomposition on the embedded matrix X to obtain eigenvalues and eigenvectors, λ_i and v_i, i=1,2,...,L .

$$X = U \sum V^T \tag{5}$$

Constructing Subspaces: Selecting appropriate eigenvalues and eigenvectors to construct subspaces. Typically, the top r largest eigenvalues and corresponding eigenvectors are chosen, where r is the number of effective components in the signal.

Reconstructing Signals: Reconstructing the signal using the selected eigenvectors to obtain the decomposed signal.

$$\hat{x}_t = \sum_{i=1}^{r} \lambda_i \langle v_i, X_t \rangle v_i \tag{6}$$

Where $\hat{x}_t$ is the reconstructed signal, $\langle \cdot, \cdot \rangle$ denotes the inner product operation, and X_t is the row vector of the data embedding matrix corresponding to time point t.

SVR is a regression method based on Support Vector Machines (SVMs), used to solve regression problems. It has several characteristics such as non-linear modeling capability, robustness, control of model complexity, and margin bounds. In SVR, the key elements are the support vectors, which determine the final result of the model. Due to its robustness, SVR utilizes support vectors as critical elements in the model,

thereby reducing the influence of noisy data and outliers. The basic idea of SVR is to find an optimal hyperplane to fit the data, maximizing the margin around the hyperplane while limiting the errors within the margin. In SVR, the article define a boundary, known as "support vectors," to accommodate most of the points in the training data instead of trying to fit each data point accurately to the hyperplane. The objective of SVR is to find a function f(x) such that for a given input x, its predicted value f(x) has the smallest possible error compared to the actual value y while keeping the complexity of the function as low as possible. The typical objective function of SVR is formulated as:

$$min_{\omega,b,\xi,\xi^*} \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{7}$$

Subject to the constraints:

$$y_i - \omega \cdot \phi(x_i) - b \leq \epsilon + \xi_i \tag{8}$$

$$\omega \cdot \phi(x_i) + b - y_i \leq \epsilon + \xi_i^* \tag{9}$$

$$\xi_i, \xi_i^* \geq 0 \tag{10}$$

Where ω is the weight vector, b is the bias term, C is the feature mapping function, $\phi(x_i)$ is the regularization parameter, ϵ is the epsilon-tube, and ξ_i and ξ_i^* are the slack variables. By solving the above optimization problem, the article can obtain the optimal weight vector ω and bias term b, thereby determining the final predictive function.

BLS is a machine learning method aimed at utilizing a large number of shallow neural networks to achieve learning, with each network responsible for learning a portion of the data's features. In width learning, neural networks are typically used for modeling. A typical neural network model includes two stages: forward propagation and backpropagation, each containing many parameters to be optimized. During the training stage, optimization algorithms such as gradient descent are commonly used to minimize the loss function. During the inference stage, the input data is passed through the trained network to generate predictions.

2.3 Evaluation Metrics

This paper adopts three performance indicators, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2_score), to evaluate the performance of the proposed model algorithms. The calculation formulas are as follows:

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-y_i^*)^2} \tag{11}$$

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i - y_i^*| \tag{12}$$

$$R^2\_score=1 - \frac{\sum_{i=1}^{n}(y_i-y_i^*)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \tag{13}$$

Where n is the number of samples, y_i is the true label (observation value), y_i^* is the predicted label (prediction value), and $\overline{y}$ is the mean value of the true labels. The smaller the values of the first two indicators, and the closer they are to 1, the better the model fit.

3 Results Analysis and Discussion:

3.1 Code Process

This paper proposes the CEEMD-SSA-LSTM-SVR-BLS model, outlining the specific process as follows.

To begin with, the data preprocessed with normalization is decomposed into high and low-frequency components using the CEEMD module. In Fig. 1, IMF1 to IMF6 represent the high-frequency components, reflecting the decisions and transactions of stock traders in the short term. IMF7 to IMF11, on the other hand, depict the low-frequency components, which capture the overall trends of stock traders influenced by government policies, laws, regulations, and long-term market dynamics.

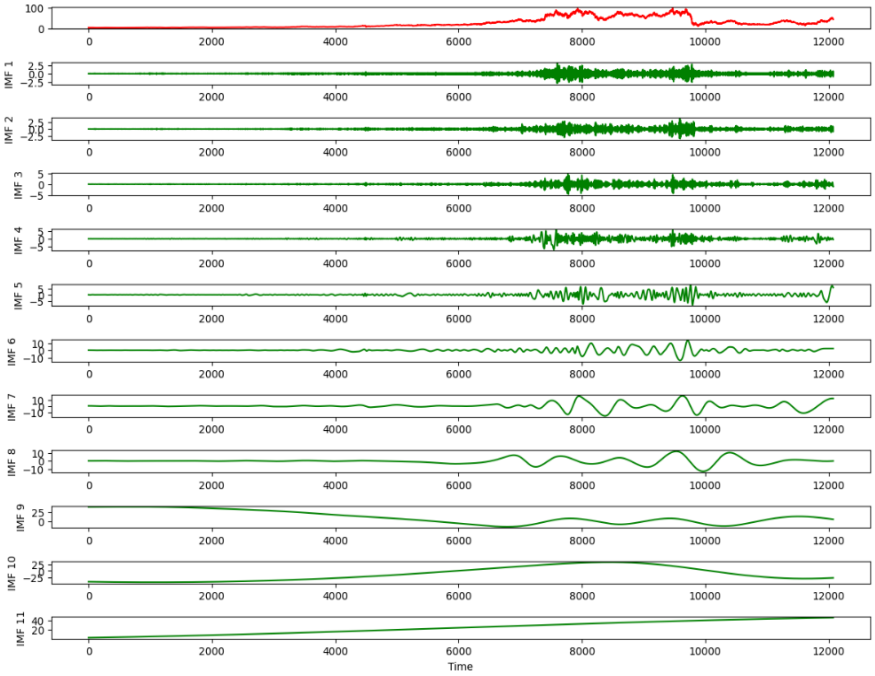

Fig.1. CEEMD decomposition

Fig.2 shows the methods and processes. Then utilizing LSTM's excellent ability to capture long-term dependencies, the article fitted the low-frequency component (i.e., long-term trends, also known as low-frequency IMF). Meanwhile, leveraging SVR's robustness and generalization capabilities, the article fitted the high-frequency component (i.e., short-term fluctuations, also known as SSA). Finally, by employing BLS for its robustness to noise and outliers, the article further improved the model's performance and significantly reduced prediction time.

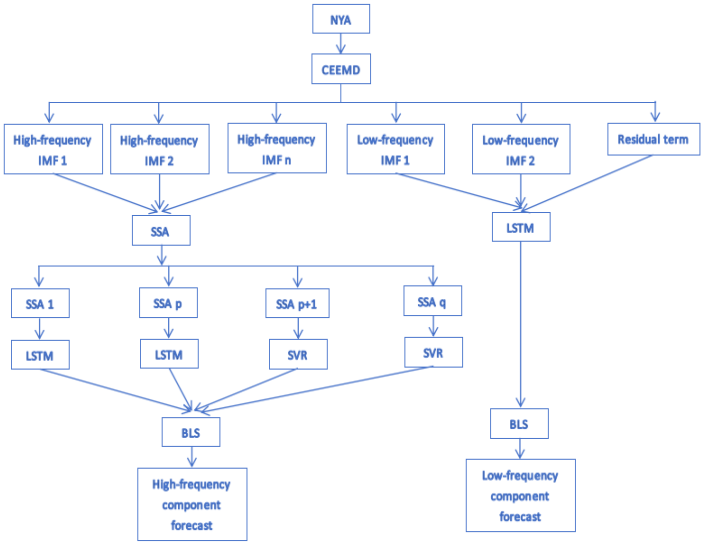

Fig. 2. Methods and Processes

3.2 Results Analysis

To highlight the superiority of our model, the article compared it with six other models: SSA-LSTM, SSA-SVR, EMD-LSTM, EMD-SVR, LSTM, and SVR.

Table 1 Comparison of Predictive Performance Metrics with Other Models .

	RMSE	MAE	R^2_score
CEEMD-SSA-LSTM-SVR-BLS	*0.0634*	*0.0287*	*0.9993*
SSA-LSTM	*0.1715*	*0.1465*	*0.9472*
SSA-SVR	*0.6635*	*0.6048*	*0.9211*
EMD-LSTM	*0.4244*	*0.3248*	*0.9677*
EMD-SVR	*0.4788*	*0.4182*	*0.9589*
LSTM	*0.5100*	*0.3701*	*0.8374*
SVR	*0.4756*	*0.4457*	*0.8503*

As shown in Table 1, the model proposed in this paper exhibits superior performance compared to any existing model. The performance indicators, namely RMSE and MAE, are both below 1, making it the only model among all to achieve this level of precision. Particularly, when compared with EMD-LSTM and EMD-SVR models separately, the advantage of utilizing CEEMD to classify data into low and high-frequency components, subsequently processed by LSTM and SVR respectively, is intuitively demonstrated. This significantly enhances the overall predictive accuracy of the model.

To enhance the contrast, Fig.3, Fig.4 and Fig.5 are inserted here.

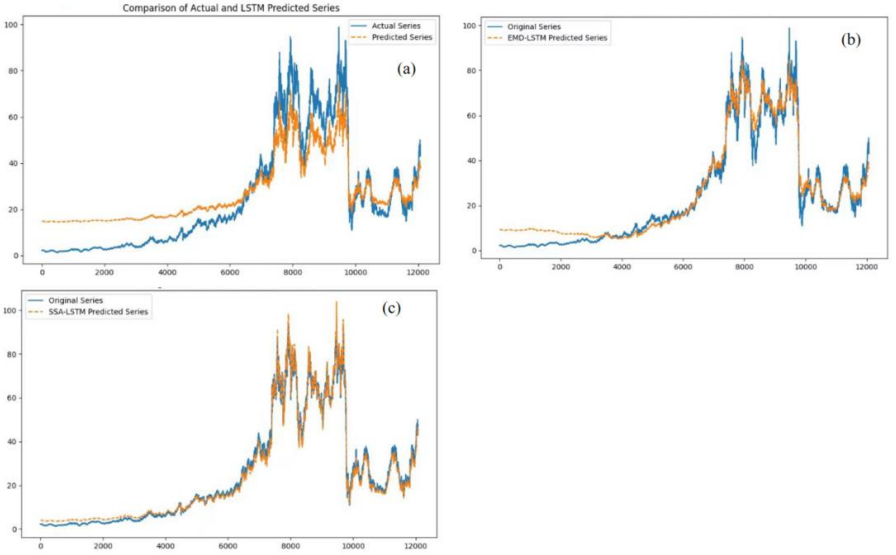

Fig. 3 Predicted and true values of LSTM-based model

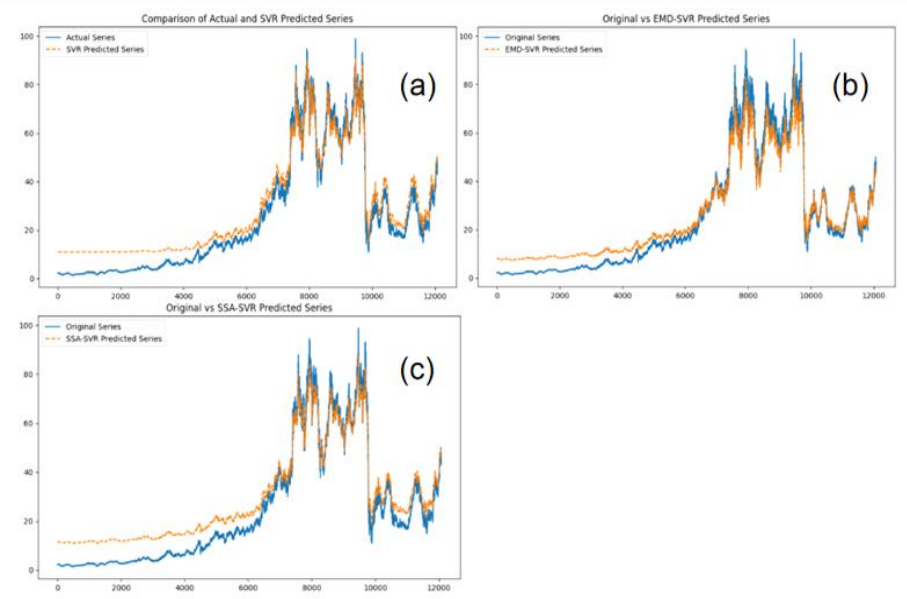

Fig. 4. Predicted and true values of SVR-based model

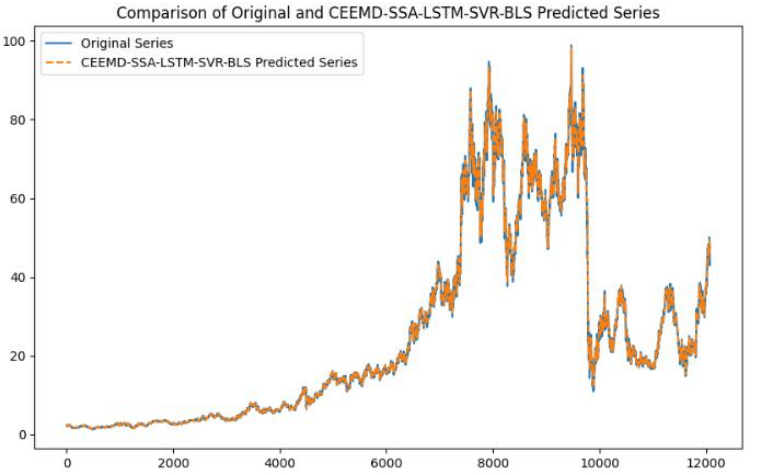

Fig. 5. Predicted and true values of CEEMD-SSA-LSTM-SVR-BLS model

As depicted in Fig. 3, (a), (b), and (c) represent the comparison between predicted values and actual values for LSTM, EMD-LSTM, and SSA-LSTM models respectively. Fig. 4 (a), (b), and (c) depict the comparison between predicted values and actual values for SVR, EMD-SVR, and SSA-SVR models respectively. Fig. 5 illustrates the comparison between predicted values of the model proposed in this paper and the actual values.

Comparing Fig. 3 with Fig. 5 reveals the lagging nature of traditional LSTM models in prediction. Furthermore, due to the long-term dependency of LSTM models, their predictive accuracy tends to decrease during periods of high short-term volatility. However, the use of SVR to handle high-frequency components in this paper compensates for these shortcomings of LSTM, thereby enhancing the model performance.

Comparing Fig. 4 with Fig. 5 indicates significant biases in long-term stable price predictions by traditional SVR models. Utilizing LSTM to process low-frequency components in this paper improves long-term prediction accuracy. Therefore, by ingeniously leveraging the strengths of LSTM and SVR models, the proposed model achieves superior predictions in both long-term stable phases (0 to 6000 time steps) and short-term volatile periods (8000 to 12000 time steps). This improvement addresses the common issues of lagging and poor fitting in existing models. Additionally, the incorporation of BLS enhances the model's generalization ability and code execution speed.

As a result, the predicted values of the CEEMD-SSA-LSTM-SVR-BLS model proposed in this paper almost perfectly fit the actual values, achieving a correlation coefficient(R^2_score) of 0.9993.

4 Conclusion

Stock prices have always been a focal point of investor attention. However, current prediction models suffer from issues such as overfitting and lagging, leading to significant prediction errors. To address this, this paper proposes a CEEMD-SSA-LSTM-SVR-BLS model based on CEEMD, LSTM, and SVR. Firstly, CEEMD is employed to denoise stock data, effectively removing low-frequency noise and providing a clearer data foundation for subsequent predictions. Secondly, SSA is utilized to remove noise from the high-frequency components obtained from CEEMD, further enhancing prediction accuracy. Additionally, the introduction of the BLS method effectively alleviates overfitting risks in the LSTM model, enhancing the stability and generalization ability of the model. Compared to existing models, our model utilizes the strengths of LSTM and SVR more effectively for stock price prediction, providing valuable reference for stock selection and investment for investors.

However, this paper only considers the opening and closing prices for prediction and does not take into account real-time news and government policies. In future work, I will endeavor to utilize the excellent generalization ability of BLS to process textual information effectively and further improve the performance of the model.

References

1. Zhou,Z.K. :Post-World War Ii American Economic Development And Prosperity Reasons. Fortune Today, (12): 14-15 (2020).
2. Theory And The Dow Jones Industrial Average Index." Securities Market Guide, (6): 38-39 (1993).
3. Shiab,A.M.:The Predictability Of The Amman Stock Exchange Using The Univariate Autoregressive Integrated Moving Average (Arima) Model.Journal Of Economic And Administrative Sciences, 22(2):17-35.(2006).
4. Oukhouya,H.,Himdi.E.K.:Comparing Machine Learning Methods—Svr, Xgboost, Lstm, And Mlp— For Forecasting The Moroccan Stock Market† .Computer Sciences &Amp; Mathematics Forum,7(1): (2023).
5. Ye,T . :The Svr Parameters Optimization For Stock'S Closing Price Forecast// Chongqing Global United Institute Of Science And Technology Research.Proceedings Of 2017 2nd Joint International Mechanical,Electronic And Information Technology Conference(Jimet 2017).Wuhan University Of Technology: 4.(2017).
6. Mohit,B.,Archana,S. ,Nand,K.: Forecasting Multistep Daily Stock Prices For Long-Term Investment Decisions: A Study Of Deep Learning Models On Global Indices. Engineering Applications Of Artificial Intelligence,129107617-.(2024).
7. Lin,Y.: Nonlinear Model Analysis Of The Chinese Stock Market. Mathematics In Practice And Understanding, (12):6-12. (2007).
8. Oukhouya,H.,Himdi,E.K.: A Comparative Study Of Arima, Svms, And Lstm Models In Forecasting The Moroccan Stock Market.International Journal Of Simulation And Process Modelling, 20(2):125-143(2023).
9. Lin,C.:High-Frequency Financial Time Series Forecasting Based On Improved Emd-Lstm.Jiangxi University Of Finance And Economics, 000326, (2021).

10. Zhang,Y., Yan,B.,Aasma,M.:A Novel Deep Learning Framework: Prediction And Analysis Of Financial Time Series Using Ceemd And Lstm. Expert Systems With Applications,159: 113609(2020).
11. An,W.,Gao,B.,Liu,J., Et Al.:Predicting Hourly Heating Load In Residential Buildings Using a Hybrid Ssa–Cnn–Svm Approach, Case Studies In Thermal Engineering,59,104516-. (2024).
12. Yang,Y. ,Long,J. ,Yang,L., Et Al.:Correction Control Model Of L-Index Based On Vsc-Opf And Bls Method .Sustainability,16,(9)(2024).
13. Sun,R.Q.: Research On Prediction Model Of Us Stock Index Price Trend Based On Lstm Neural Network. Capital University Of Economics And Business,(2016).
14. Hu,X.C.: Research On Semantic Relationship Classification Based On Lstm. Harbin Institute Of Technology(2015).