Hindawi
Scientific Programming
Volume 2022, Article ID 4758698, 12 pages
https://doi.org/10.1155/2022/4758698

Research Article

Research on Stock Price Time Series Prediction Based on Deep Learning and Autoregressive Integrated Moving Average

Daiyou Xiao[1] **and Jinxia Su**[2]

[1]*School of Finance, Central University of Finance and Economics, Beijing, China*
[2]*School of Business, Central University of Finance and Economics, Beijing, China*

Correspondence should be addressed to Daiyou Xiao; 2019110134@email.cufe.edu.cn

Received 7 December 2021; Revised 24 January 2022; Accepted 21 February 2022; Published 31 March 2022

Academic Editor: Hangjun Che

Different from traditional algorithms and model, machine learning is a systematic and comprehensive application of computer algorithms and statistical models, and it has been widely used in many fields. In the field of finance, machine learning is mainly used to study the future trend of capital market price. In this paper, to predict the time-series data of stock, we applied the traditional models and machine learning models for forecasting the linear and non-linear problem, respectively. First, stock samples that occurred from year 2010 to 2019 at the New York Stock Exchange are collected. Next, the ARIMA (autoregressive integrated moving average model) model and LSTM (long short-term memory) neural network model are applied to train and predict stock price and stock price subcorrelation. Finally, we evaluate the proposed model by several indicators, and the experiment results show that: (1) Stock price and stock price correlation are accurately predicted by the ARIMA model and LSTM model; (2) compared with ARIMA, the LSTM model performance better in prediction; and (3) the ensemble model of ARIMA-LSTM significantly outperforms other benchmark methods. Therefore, our proposed method provides theoretical support and method reference for investors about stock trading in China stock market.

1. Introduction

Stock market forecasting is a behavior to determine the future value of corporate stocks or other financial instruments traded on exchanges. Successful forecast of the future stock price can make considerable profit. According to EMH (efficiency market hypothesis), stock prices reflect all existing information, so any price changes not based on newly released information cannot be forecast. Although other people disagree with the hypothesis, some supporters of the view hold countless methods and techniques that supposedly allow them to access to future price information.

Stock market forecast is especially difficult, given the nonlinearity, volatility, and complexity of the market. Before the emergence of machine learning technology, stock market forecasts were generally realized through fundamental and technical analysis. With the computer technologies, such as machine learning, emerged and developed in business [1], deep learning, especially neural network model, has become the current hot spot of stock prediction model. Meanwhile, stock market forecast has been more convenient and efficient due to these technologies [2]. At present, stock forecasting models usually fall into traditional linear models and models represented by deep learning. However, since the time series data have both linear and nonlinear parts, the forecasting results singly through forecasting models are usually not so reliable. Therefore, many experts and scholars combine various single models to significantly improve the accuracy and stability of the forecasting results.

In addition, the coefficient of association between the stock index and its constituent stocks can reflect the sensitivity of the constituent stocks to the changes of the stock index, that is, the correlation between the constituent stocks and the stock index (also known as "stock character"), which can be referred to by investors to adapt investment strategies. According to the market trend forecast, extraneous income

can be expected to gain by choosing different β coefficient of stocks. Moreover, the stock index and its constituent stock prices often keep trend in sync in the global stock market. Therefore, except for predicting stock index and single stock prices, better portfolio strategies can be worked out by forecasting the correlation coefficients of the expected constituent stock of the stock index for higher returns on investment.

Based on all of this, this paper takes the strong-enough representative S&P 500 stock index and its constituent stocks as the research object to forecast the future trend of the S&P 500 stock index through forecast models and then predict the correlation coefficient between its constituent stocks and the stock index, so as to formulate the optimal investment strategy for investors to refer to at a certain extent.

Over the past few decades, many social science researches have focused on predict social and economic development trends with quantitative methods. Many feasible methods in time-series analysis, both with advantages and disadvantages, can be interpreted as techniques for using past data to build forecasts and strategies on future value.

First, research about linear model: As early as the 1990s, the ARIMA (autoregressive integrated moving average) method has already been used by scholars to forecast in the capital market. Some researchers used the ARIMA and coefficients to predict stock market data [3], and in their experiments, researchers found that the experiment result was better than the prediction of the zero hypothesis of random fluctuations in the base value. The ARIMA model has been used in many fields including temperature prediction, prices prediction for electricity, and wind speed. Some studies adopted the process of ARIMA time in their research [4]. Yang et al. selected the Shanghai Composite Index to structure ARIMA model [5]. Kim and Sayama developed a new method aiming to forecasting the future trend of the S&P 500 index by establishing a complex network of time series of the index-foundation S&P 500 and then linking the network to the interconnected weights [6]. The study showed that adding network measurement results to the ARIMA can improve the prediction accuracy. Khashei and Hajirahimi believe that the time series in the hybrid model is divided into t linear and nonlinear two parts [7]. Therefore, ARIMA and MLP (multiparametric linear programming) are chosen to build hybrid models. They also found that on the whole, the ANN-ARIMA hybrid model can be adopted to achieve more accurate results. Unggara et al. used the Firefly algorithm to optimize the ARIMA (p, d, q) model and determined the best ARIMA model by looking for the smallest AIC (Akaike information criterion) value [8]. As a result, the ARIMA model optimized by the Firefly algorithm has a better forecasting performance.

Second, research about neural network model: The LSTM (long short-time memory) network, which has achieved further success in processing large data sets, is mainly used for deeper learning. Although LSTM model is limited in the number of inputs, Siami-Namini and Namin attempted to use the LSTM in financial data sets [9]. Experiment results indicate that the proposed method performs well in predicting economic and financial time series. Other researchers put forward a stock price prediction method using deep learning models [10]: 14 different DL methods similar to LSTM are comprehensively adopted in S&P stocks; BSE-BANKEX stock index will be capable of forecasting one or even four steps ahead. It is found that the DL methods proposed in their research can obtain a good prediction results for stock price. Joo II and Seung-ho proposed a stock price forecast model of a two-way LSTM recurrent neural network, which adds a hidden layer in the opposite direction of the data flow to deal with the limited network through the previous model based on the RNN [11]. It was found that, compared with the nonbidirectional LSTM recurrent neural network, the stock price prediction model using the bidirectional LSTM recurrent neural network has higher accuracy. To get rid of high noise in stock data, researchers applied the wavelet threshold denoising method to preprocess the initial data sets [12]. In their study, the soft/hard threshold method used for data preprocessing has a significant effect on noise suppression. Based on this research, a new multioptimal combination wavelet transform (MOCWT) was proposed, and the research finally showed that MOCWT is more accurate in forecasting than traditional methods. Researchers also proposed the LSTM model and employed it to intraday stock forecasts [13]. Chen and Ge made an exploration on the forecasting mechanism of stock price movement based on LSTM and found that it significantly improved the forecasting performance [14].

Third, the research on the hybrid model is as follows: Peter and Zhang used ARIMA and ANN hybrid method to study time series estimation [15]. Narendra Babu and Eswara Reddy proposed a linear hybrid model that can simultaneously maintain the prediction accuracy and the trend of the data [16]. Baek and Kim proposed a novel data enhancement method for stock market index prediction based on the ModAugNet framework [17]. The method includes the over-fitting prevention LSTM module and the predictive LSTM module and it is found from analysis that the test performance depends entirely on the latter. An ensemble method LSTM with GARCH is proposed [18]; it has high predictive ability and good applicability. Chen et al. proposed a new ensemble model to problems on portfolio selecting with skewness and kurtosis [19].

Through the analysis of recent literature, it can be found that domestic and foreign forecasting models can be roughly divided into linear, nonlinear, and hybrid models. In general, the current research status at home and abroad can be summarized as follows: The research on linear models mainly focuses on the ARIMA model. For recent researches, many researchers keep more belief in predictive performance of non-linear models than that of linear models. The hybrid model is the best predictive model in all. It can not only process the linear part of time series data, but also has better processing capabilities for its nonlinear part. Therefore, in our study, a single method is first used to predict the trend of stock indexes, and then a hybrid one is adopted to predict the correlation coefficients of stock indexes and their constituent stocks, so that provide investors with guidance to profit to a certain extent.

2. Methods

In this section, we first introduce ARIMA model and LSTM model, respectively, and finally introduce our proposed integration model.

2.1. Autoregressive Integrated Moving Average Model. ARIMA (p, d, and q), where p is the autoregressive term, q is the number of moving average terms, and d is the number of differences made when the time series becomes stationary. The prediction results can be adjusted by adjusting the aforementioned three parameters d, p, and q, so as to draw the optimal model. The model calculation formula is as follows:

$$y_t = \theta_0 + \Phi_1 y_{t-2} + \cdots + \Phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}, \quad (1)$$

where y_t and ε_t are the actual value and random error of the time period t, respectively; Φ_i ($i = 1, 2, \ldots, p$) and θ_j ($j = 1, 2, \ldots, q$) are the model parameters; p and q, the order of the model (p and q are integers), are also the model parameter mentioned earlier; the random error ε_t, whose mean value is 0, is assumed to be independent and obey the same distribution in the model. The variance of constant term is denoted as σ_2. Equation (1) involves several important special cases of ARIMA series models. If $q = 0$, then equation (1) can be simplified to an AR model of order p. When $p = 0$, the model can be simplified to a q-order MA model. Among them, the model order (p, q) is the key link in ARIMA model construction, which determines the accuracy of model prediction. The parameters of the AR and MA operations are defined as (p) and (q), respectively. These two parameters need to be determined by the auto-correlation graph (ACF). ARIMA includes the following steps:

Step 1: Data diagnosis and check: In the first step, it is necessary to check the stationarity of the given time series data, which is essential to improve the accuracy of forecasting. A stationary time series is a time series whose statistical properties such as mean, variance, and covariance are related to time.

Step 2: Model parameter estimation: In order to stabilize the nonstationary time series, a proper degree of difference (d) is performed on it, and the stability test is performed again and this process is continued until a stable series is obtained. (d) is a positive integer that shows the degree of difference. If the difference operation is performed (d) times, the integration parameter of the ARIMA model is set to (d), and then the obtained stationary data are identified. In this process, the model (ACF graph) and partial auto-correlation graph (PACF graph) are determined.

Step 3: Model identification and selection: After ensuring that the input variable is a stationary series, the parameter d has been determined. Next, calculation algorithms are used to estimate the parameters to find the coefficients most suitable for the selected ARIMA model. And then the AIC standard or BIC standard is used to test the model and select the minimum standard value. The essence of the two standards is maximum likelihood estimation or nonlinear least square estimation, and the AIC standard is chosen in this article.

Step 4: Model testing: Model testing is the test on whether the estimated model meets the norms of a stationary univariate process. In particular, the residuals of the model output should be independent of each other, and the mean value and the constant that changes with time should remain unchanged. If the estimate is insufficient, the modeling process must be resumed to build a better model.

Step 5: Data prediction: After the ARIMA model with the minimum AIC standard is obtained, the data can be input into the model to predict its linear part.

2.2. Long Short-Term Memory Model. Many researchers found that different models are good at dealing with different types of prediction problems. This provides a basis for using the ARIMA-LSTM hybrid model, which contains both linear and nonlinear parts, to produce better results than a single method. Figure 1 shows the LSTM neural grid stores the internal structure of cells.

Our study used the standard LSTM including the four interactive neural networks (forgetting gates, input gates, input candidate gates, and output gates).

$$f_t = \sigma\left(W_f \times \left[h_{t-1}, x_t\right] + b_f\right), \quad (2)$$

where σ represents the sigmoid activation function.

$$\sigma(X) = \frac{i}{1 + e^{-x}}. \quad (3)$$

And then, a new unit state C_t is obtained from the input gate, this state will be as an update unit state in the next time step. The input gate employed the σ as the activation function and i_t and $\widetilde{C}_t$ as outputs. i_t is employed to determined the feature in C_t to reflect $\widetilde{C}_t$.

$$\begin{aligned} i_t &= \sigma\left(W_i \times \left[h_{t-1,} x_t\right] + b_i\right), \\ \widetilde{C}_t &= \tanh\left(W_c \times \left[h_{t-1}, x_t\right] + b_c\right), \end{aligned} \quad (4)$$

σ function outputs a value in the range 0 to 1 and the tan$h$ outputs a value in the range −1 to 1.

Next, the value selected by the h_t activates O_t and C_t, which are decided by the output gate.

$$\sigma_t = \sigma\left(W_0 \times \left[h_{t-1,} x_t\right] + b_0\right), \quad (5)$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}_t, \quad (6)$$

$$h_t = o_t \times \tanh\left(C_t\right). \quad (7)$$

Equations (6) and (7) produce the C_t and h_t, and they will be passed to the next time step. The experiment in this article is a regression problem, and the range of output value of the proposed model is −1 to 1; therefore, the last element is activated by the tan$h$ function.

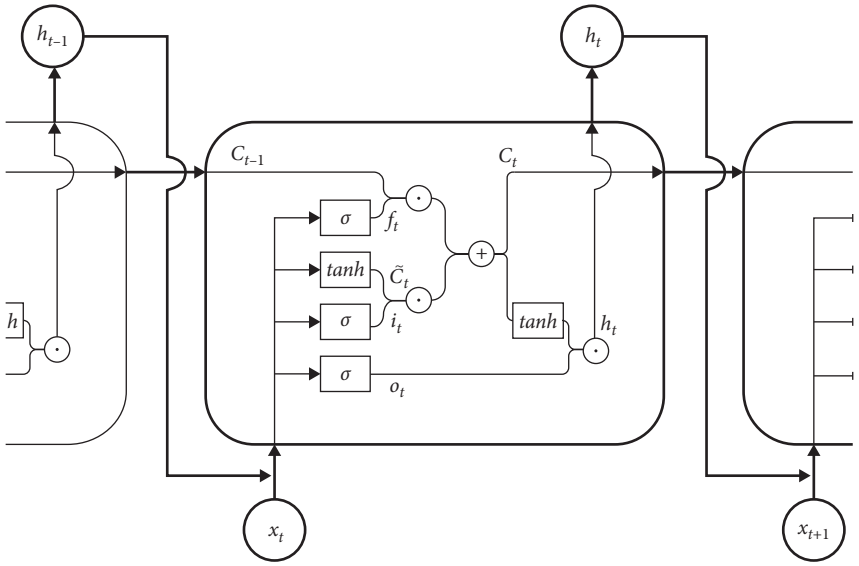


Figure 1: LSTM neural grid stores the internal structure of cells.

2.3. Ensemble Prediction Model. Unlike single algorithms, a combination of multiple methods can obtain higher estimation results [20]. These hybrid models are based on supervised machine learning algorithms, so they can be used for training and prediction purposes. Moreover, the ensemble methods improve the solving problem applicability and will obtain better performance [21, 22]. Traditional econometric models and machine-learning-based models have been widely used in the prediction research of time series. In our study, for the time series in the stock market, due to the existence of a large number of linear and nonlinear relations, the previous single model would be difficult to deal with this type of prediction. Therefore, in our study, we will make a combination prediction based on ARIMA and LSTM, respectively, for different characteristics of stock market data, in order to obtain better prediction effect. Recently, the ensemble method based on ARIMA and LSTM has been applied to some fields like business and energy and achieved great success [23–25].

Even if the results obtained using the mixed model and the results obtained using the model alone are not related to each other, it also demonstrates that it has reduced the prediction error. Therefore, the hybrid model is considered to be the most successful prediction task model [26]. To make predictions, many ensemble methods composed of linear and nonlinear models are employed by different researchers. In our experiment, historical data are used in the time series to predict the future. Figure 2 introduces the structure of the proposed ensemble method.

$$y_t = L_t + N_t. \tag{8}$$

Figure 2 shows the ARIMA-LSTM hybrid model. In our time series data sets, L_t represents the linear part and N_t represents the nonlinear part. In our hybrid model, the linear part L_t is calculated through the ARIMA at first, and then the LSTM is applied to predict the nonlinear part N_t. At last, the sum of the error values of the two models is obtained.

In the mixed model, the linear component L_t is calculated through the ARIMA model at first and then the LSTM model is used to predict the nonlinear component N_t of the time series. At last the sum of the error values of the two models is obtained. The formulas for calculating L_t and N_t are given in formulas (9) and (10):

$$\text{LSTM}_{\text{error}} = \text{LSTM\_mean}[\text{error}], \tag{9}$$

$$\text{ARIMA}_{\text{error}} = \text{ARIMA\_mean}[\text{error}]. \tag{10}$$

Calculate the weight of the model using the error values obtained in equations (11) and (12).

$$\text{ARIMA}_{\text{weight}} = \left(1 - \left(\frac{\text{LSTM}_{\text{error}}}{\text{LSTM}_{\text{error}} + \text{ARIMA}_{\text{weight}}}\right)\right) \times 2, \tag{11}$$

$$\text{ARIMA}_{\text{weight}} = 2 - \text{LSTM}_{\text{error}}. \tag{12}$$

Use the given equation (13) to obtain the weight value of the model and each predicted value of the final mixed model.

$$\text{Hybrid}_{\text{predict}}[i] = \frac{\text{LSTM}_{\text{weight}}[i] \times \text{LSTM}_{\text{error}}[i] + \text{ARIMA}_{\text{weight}}[i] \times \text{ARIMA}_{\text{error}}[i]}{2}. \tag{13}$$

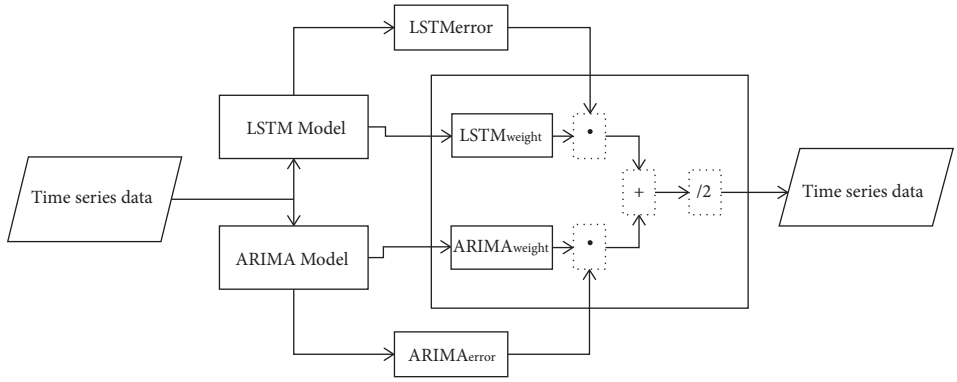


Figure 2: ARIMA-LSTM hybrid model.

3. ARIMA-LSTM Hybrid Model Design and Evaluation

To evaluate the performance of the proposed ensemble model, we employ some commonly used method such as MAE, MSE, and RMSE to compare and evaluate the ensemble method and several benchmark methods.

3.1. Data Set Selection and Preprocessing. In this study, two stock index forecasting models, ARIMA and LSTM, are constructed at first. The S&P 500 stock index is selected in the empirical data selects, and the daily trading data is selected in the data sample interval selects from January 1, 2010, to December 31, 2019, which are 2519 sets of data in total. Among them, the first 90% is used for model training, and the 10% is used for model prediction. The S&P 500 stock index sequence is shown in Figure 3. It can be found from the figure that within the selected time range, the S&P 500 Index generally shows a steady increasing trend.

3.2. Comparative Analysis of Stock Index Forecast Model Results. After obtaining the prediction data set, the aforementioned four test methods are used in this study to test the data of each forecasting method. The following table shows the different loss value obtained on the basis of the prediction of the four foreign exchange median prices and the ARIMA model and the RNN neural network model.

Table 1 shows the fitting results based on the loss values of the prediction results of each model under different loss functions. It can be seen from the table that the loss functions of the LSTM model are all smaller than the ARIMA model, which is because the LSTM model can not only describe the nonlinear relationship of time series data but also has certain processing capabilities for its linear part despite of its instability in comparison with the ARIMA model. However, generally speaking, both models have gained very low loss values, indicating that the two models are both relatively perform well in predicting accuracy. Figure 4 shows the predicted results using LSTM and ARIMA, respectively.

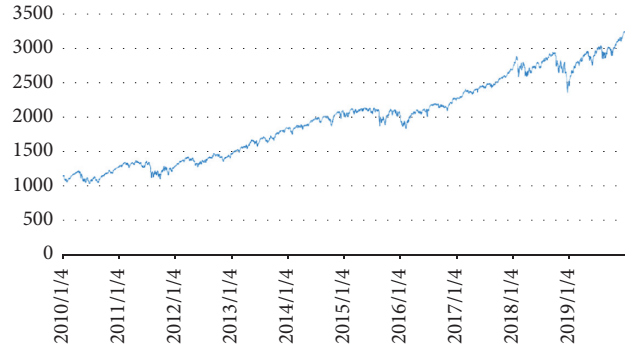


Figure 3: S&P Stock index closing sequence.

Table 1: The experiment results about ARIMA and LSTM model forecasting.

	MSE	MAE	RMSE
ARIMA	0.000101	0.007333	0.043788
LSTM	0.000096	0.007184	0.028828

3.3. The Design of Stock Price Correlation Coefficient Prediction Ensemble Model

3.3.1. The Design of ARIMA for Stock Price Correlation Coefficient Prediction

(1) In the experiment of correlation coefficient prediction, the adjusted closing price of the constituent stocks of the S&P 500 index is selected, and the sample interval is still set from January 1, 2010, to December 31, 2019, on the New York Stock Exchange daily transaction receipts. Data are mainly acquired in the use of Python language's Beautiful Soup function library through crawler technology. The trading data of the constituent stocks originates from the Quandl database, and the industry information of the constituent stocks is from Wikipedia.

After preprocessing the data, the program randomly generates 150 stocks from the remaining 446 assets, and calculates the correlation coefficient of each pair of assets in a 100-day time window. In order to diversify the data, 5 sets of data are set up in this

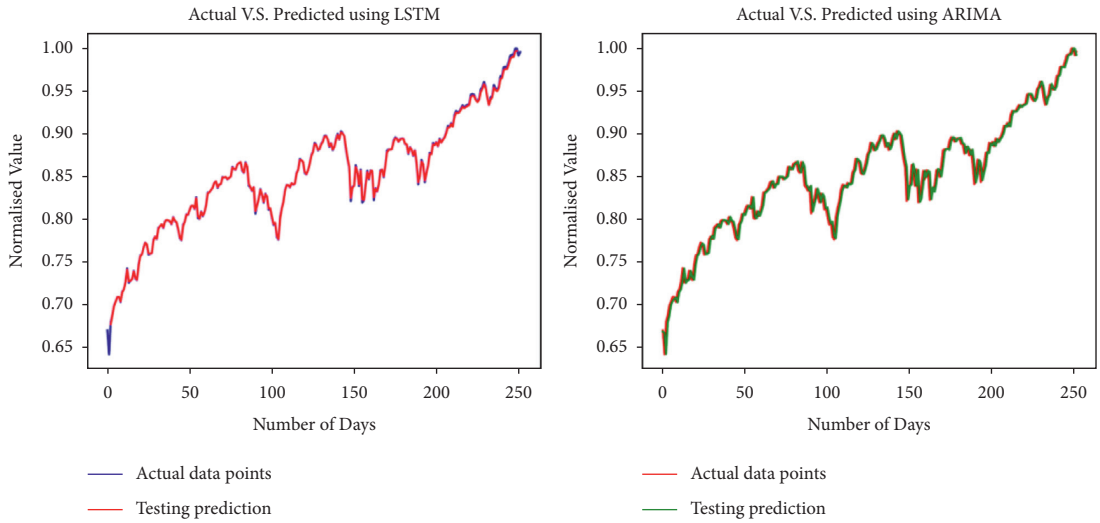


Figure 4: The predicted results using LSTM and ARIMA.

article with a starting value every 20 days: day 1; day 21; day 41; day 61; and day 81. Each value corresponds to a rolling 100-day window, advancing in 100-day time-steps until the end of the data set training. In this process, a total of 55,875 sets of time series data were trained, and each set has 24 time-steps. Development, test1, and test2 are produced using these $55{,}875 \times 24$ data sets. In the model evaluation stage, this paper divides the data as follows to achieve forward optimization.

(2) The parameters of the model should be determined before fitting the ARIMA model. ARIMA (p, d, q), where d is easiest to be determined. Data difference aims to making the last data used is a time series that tends to be stable, which can improve forecasting accuracy. As mentioned in the previous section, the S&P 500 Index and its constituent stocks generally show a steady increasing trend. The data will tend to be stable after a difference, so the parameter d here can be determined as the value 1. The determination of the parameters p and q needs to adopt the ACF and PACF of the data.

The ACF and PACF are set into zero after a certain order is called truncation. The running results show that most data sets show an oscillation trend, as shown in Table 2. There are also notable trends covering rising/falling trends, large drops occasionally when the correlation coefficient is stabilized, and stable periods with mixed oscillations. Although the ACF and PACF images show that most of the data sets are close to white noise, the images show that five groups of parameters can be effectively used in the prediction of the ARIMA model. These five sequences are used in this article to test the ARIMA model, and a total of 55,875 data sets are trained. What is more, for each data set, we will select the smallest AIC-value-based model after training.

AIC (Akaike information criteria) is a commonly used test standard for the prediction performance of ARIMA models. The expression of AIC calculation is as follows:

$$\mathrm{AIC} = -2\mathrm{In}(L) + 2N, \tag{14}$$

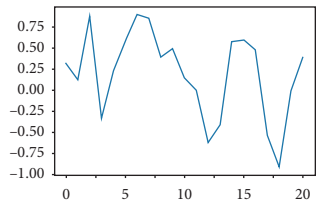
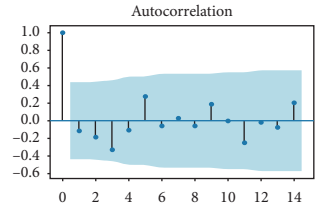
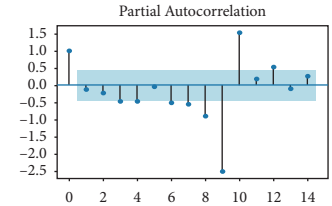
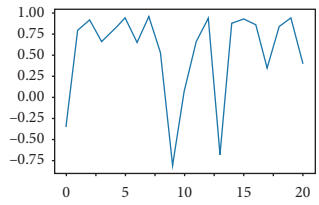
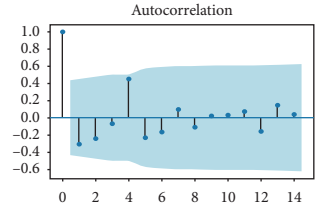
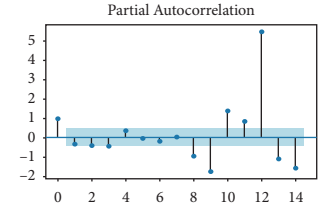
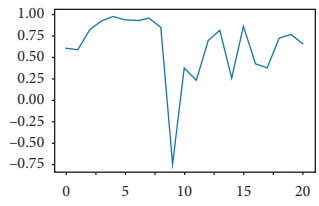
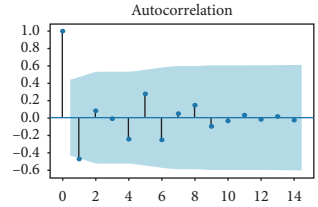
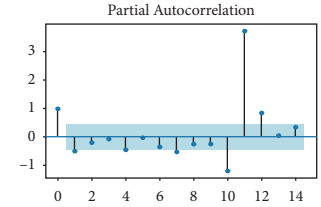
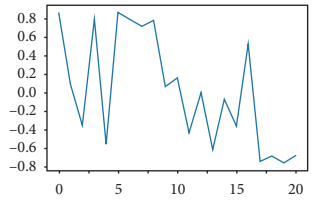
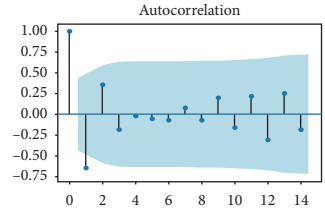
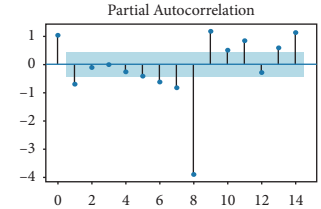
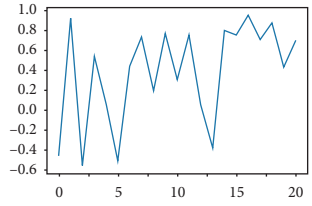
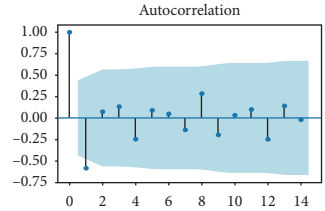
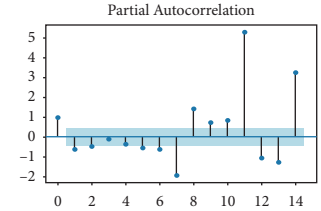
where L represents the maximum likelihood function and N represents the number of parameters.

The AIC standard was proposed by the Japanese statistician Akaike, so it is named directly after initials of his name. To evaluate the performance of the ARIMA model with the application of AIC standard, the maximum likelihood function and the model parameters are used to judge its prediction effect. Specifically, the larger the maximum likelihood function value, the higher the prediction effect; theoretically speaking, the more the number of model parameters is set, the lower the difficulty of fitting the data relationship or the better the fit will be. However, too many parameters will also complicate the model structure, which may lead to more difficulties in parameter estimation, thereby reducing the model prediction accuracy. Therefore, the ideal ARIMA model should be the optimal combination of maximum likelihood function and parameters. The AIC standard comprehensively considers the above two indicators and can perform comprehensively on evaluation of the ARIMA model. Therefore, when optimizing the ARIMA model, the parameter with the smallest AIC value will be selected.

If the ARIMA model is used to predict future data, the generated data are in the ARIMA model. In other words, the underlying process of generating the time series only has a linear correlation structure, but the nonlinear relationship in the experiment data cannot be described. The ARIMA method still has certain limitations in predicting complex real-world problems. In this regard, the NN model can be employed to analyze the nonlinear parts that the ARIMA model cannot deal with.

After fitting the ARIMA model to the linear part of the data, this article generates a new data set to calculate the residual value of the remaining non-linear part at every 21-time steps, as shown in Figure 5. Since the input is the nonlinear partial residuals processed by the ARIMA model, the residual distributions of the X and Y data sets all fall

Table 2: ARIMA model's ACF/PACF.

Type	Data plot	ACF plot	PACF plot
Oscillatory			
Oscillatory and steady			
Dip			
Decreasing			
Increasing			

between 0 and 1. The newly generated X and Y segmentation data set will be used as the input value of the next nonlinear LSTM model for training.

3.3.2. Forecast Design Based on LSTM Stock Price Correlation Coefficient. (*1*) *Data Selection and Acquisition*: After the ARIMA model processes the linear part of 150 pairs of combined assets generated at any time, the remaining nonlinear part is calculated as the residual value and used as the input of the LSTM model, as shown in Figure 5.

The input data set of the LSTM model is also divided into X and Y trains, X and Y developments and two sets of X and Y test set 1 and test set 2. The input data are stored in the X and Y data sets as shown in Figure 6. Each x data set size is a $55{,}874 \times 20$ matrix, and each X time series corresponds to a Y data set.

(2) Training for LSTM Model: The model structure constructed in this paper is an improved LSTM model based on RNN, which contains 25 units. The final output of the cells is combined into a value with a full-connection layer. This value is then output as a final predicted value through a tan$h$ activation function of a two-layer network. The tan$h$ activation function of the two-layer network can be simply understood as the tan$h$ function magnified by two times. Figure 7 shows the simplified architecture of the method.

3.4. Prediction Results Analysis

3.4.1. Forecasting Performance Evaluation. This paper aims to fit the parameters of the model so that the optimal parameters can be used to apply and predict various assets in different time periods. Therefore, only the first window is trained, and the trained model can be applied to the data training of the three time intervals of the validation set and the two test sets. In addition, when the prediction results of

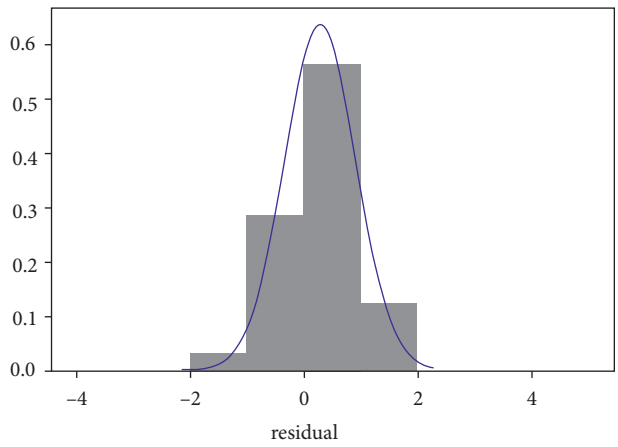


Figure 5: Residual data distribution of training set.

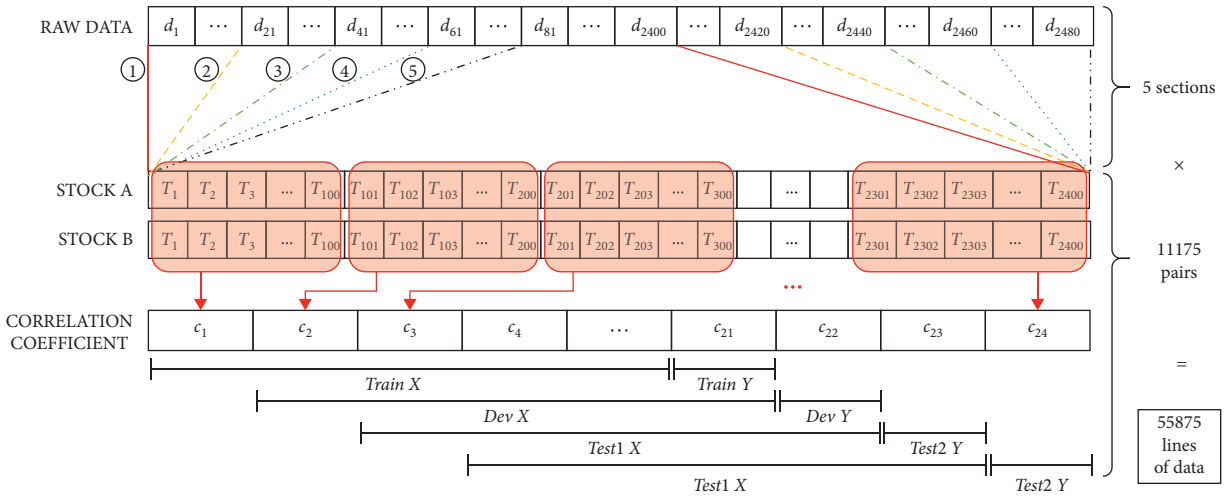


Figure 6: The input data for time series.

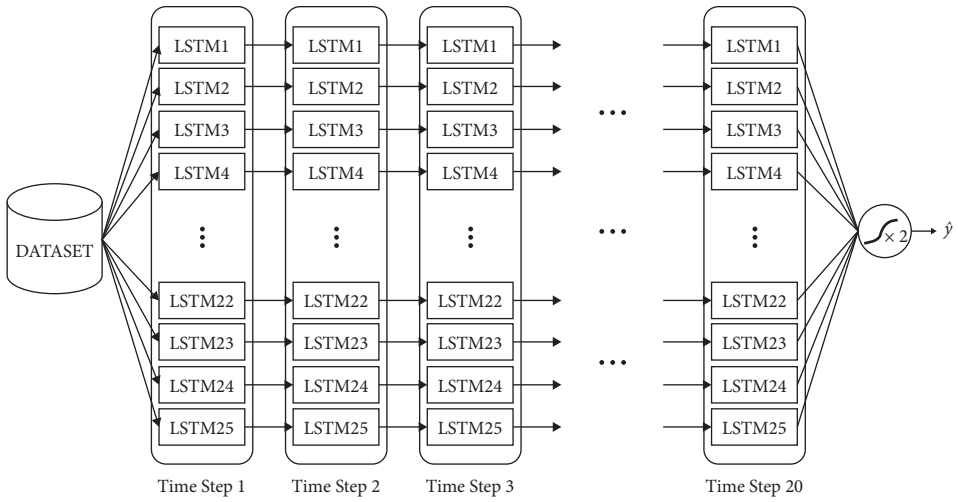


Figure 7: The structure of the LSTM model.

the correlation coefficient of the model in the two time periods are relatively ideal, some classic financial prediction models are selected to analyze the prediction effects of each model to test the model in this article. The MSE and MAE values of four financial models are calculated in this article.

3.4.2. Forecast Results and Analysis. After the data are processed, in the hybrid model prediction experiment, the ARIMA method is first employed in this article to process the S&P 500 index component stocks in the aspect of linear as the first step, and then the nonlinear part of the data residual value processed at the first step is used as the input data of the LSTM model. Finally, model establishment, data training and testing is developed. The final prediction results of the correlation coefficient between the 150 randomly generated asset portfolios and the S&P 500 index in the next 20 time steps are shown in Figure 8.

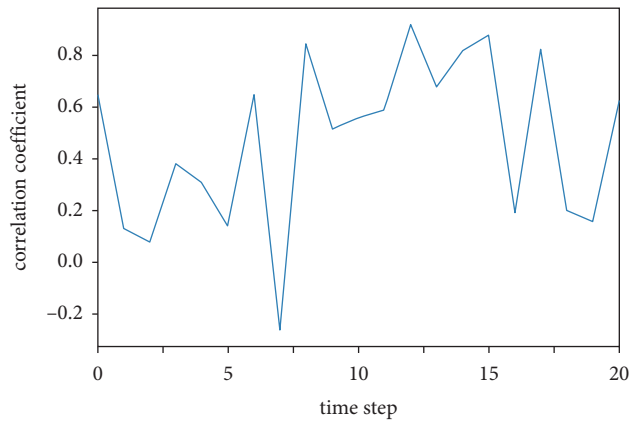


Figure 8: Prediction results of correlation coefficient.

3.5. Control Group Forecasting Model. Predicting the results by the hybrid model alone is not enough to show that the certain advantages of the model in the forecasting performance of research objects such as correlation coefficients. In order to make comparison between the proposed hybrid model proposed and other models for the accuracy of financial sequence forecasting, other commonly used forecasting models are introduced as the reference group. Many studies have shown that the full-sequence model is poor in prediction performance during the period of predicting financial sequences, so three other commonly used prediction models are also discussed, which are compared with the prediction results of hybrid models.

3.5.1. Full-Sequence Model (FS). Adopting the full-sequence algorithm is the easiest way to estimate the portfolio correlation. All the past correlation values are used in the model to predict the future correlation coefficient.

$$\widehat{\rho}_{ij}^{(t)} = \frac{\beta_i \beta_j \sigma^2{}_m}{\sigma_i \sigma_j} \rho_{ij}^{(t)} = \widehat{\rho}_{ij}^{(t-1)}. \tag{15}$$

However, compared with other equivalent models, the prediction quality of this model is relatively poor.

3.5.2. Constant Coefficient Correlation Model (CCC). The CCC model shows that the average value of the correlation coefficients of all asset portfolios can be regarded as the estimated value of the required predicted asset portfolio. Therefore, all assets in the portfolio in this model have the same correlation coefficient.

$$\rho_{ij}^{(t)} = \frac{\sum_{i>j} \rho_{ij}^{(t-1)}}{n(n-1)/2}. \tag{16}$$

3.5.3. Single-Index Model (SI). Adopting the single-index model is a simple way of asset pricing, which is usually used to evaluate the risk and return of stocks. To facilitate calculation and analysis, the single-index model acts with a kind of macro factor, such as the S&P 500 index, to measure the risk and return of stocks. The single-index model assumes that the rate of return on assets and the "single index," that is, the market rate of return changes in the same direction. In order to quantify the volatility of assets and market returns, it is necessary to specify the market returns themselves. This specification is called the "market model."

$$R_{i,t} = \alpha_i + \beta_i R_{m,t} + \varepsilon_{i,t}, \tag{17}$$

where $R_{i,t}$ represents the return of asset i at time t; in the same way, $R_{m,t}$ represents the return of asset m at time t; α_i represents the excess return of asset i after risk adjustment; β_i represents the impact of asset i on the market sensitivity; $\varepsilon_{i,t}$ represents the residual income of asset i at time t, also called the error term. So there is

$$\begin{aligned} E(\varepsilon_i) &= 0, \\ \mathrm{Var}^{(\varepsilon_i)}\mathrm{Var}(\varepsilon_i) &= \sigma_{\varepsilon_i}^2, \\ \mathrm{Cov}(R_i R_j) &= \rho_{ij}\sigma_i\sigma_j = \beta_i\beta_j\sigma_m^2, \end{aligned} \tag{18}$$

where σ_i and σ_j respectively represent the standard deviation of asset i and asset j; σ_m represents the standard deviation of market returns. In a single-index model, the estimated value of the correlation coefficient $\widehat{\rho}_{ij}^{(t)}$ can be expressed as,

$$\widehat{\rho}_{ij}^{(t)} = \frac{\beta_i \beta_j \sigma^2{}_m}{\sigma_i \sigma_j}. \tag{19}$$

3.5.4. Multisequence Model. The industry sector of the asset is considered in the multisequence. The model assumes that assets generally have a trend of volatility in the same direction in the same industry, so it can be considered that the correlation coefficients of the asset portfolio are equal to the average value of the correlation coefficients of the industry. For example, there are company A and B that belong to industry sectors α and β, respectively; then, their correlation coefficients are equal to the average correlation coefficients of all asset portfolios in their respective industry sector combinations (α, β). According to whether the two industrial sectors α and β are the same, the prediction formula is slightly different. The equation is as follows:

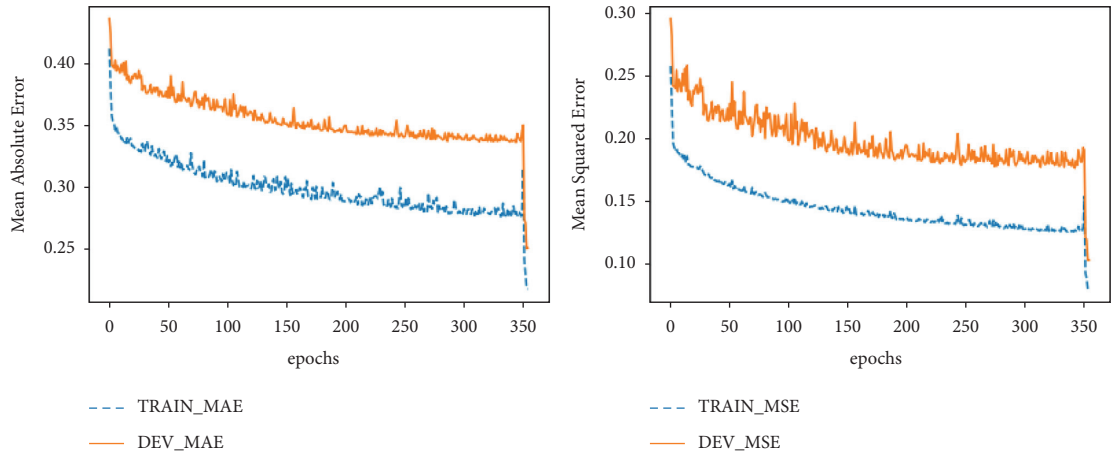


Figure 9: The learning curve of the ensemble model (ARIMA-LSTM).

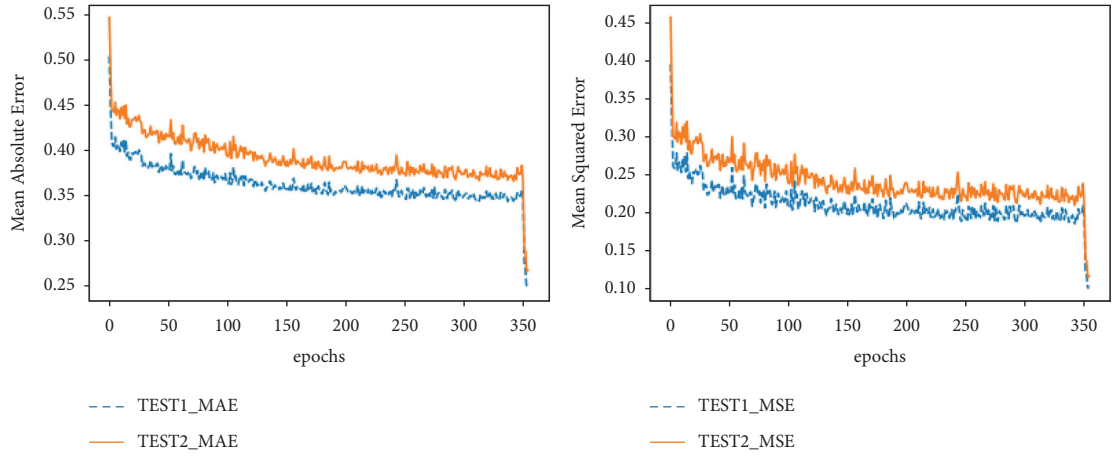


Figure 10: The prediction curve of the ensemble model (ARIMA-LSTM).

Table 3: ARIMA-LSTM mixed-model loss function evaluation table.

	Develop data set			Test1 data set			Test2 data set		
	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE
ARIMA-LSTM	0.103	0.320	0.250	0.101	0.319	0.248	0.116	0.341	0.266
Full history	0.479	0.692	0.558	0.463	0.681	0.554	0.432	0.657	0.523
Constant correlation	0.277	0.526	0.462	0.214	0.463	0.400	0.281	0.530	0.430
Single-index model	0.420	0.624	0.540	0.411	0.641	0.534	0.247	0.497	0.482
Multigroup	0.307	0.554	0.451	0.291	0.539	0.455	0.297	0.536	0.448

$$\begin{cases} \dfrac{\sum_{i\varepsilon\alpha}^{n\alpha}\sum_{j\varepsilon\beta;i\neq j}^{n\beta}\rho_{ij}^{(t-1)}}{n_\alpha\left(n_\beta-1\right)}, & \alpha=\beta, \\ \dfrac{\sum_{i\varepsilon\alpha}^{n\alpha}\sum_{j\varepsilon\beta;i\neq j}^{n\beta}\rho_{ij}^{(t-1)}}{n_\alpha n_\beta}, & a\neq b, \end{cases} \tag{20}$$

where α and β respectively represent different industry sectors in the stock market; n_α and n_β represent the number of companies in the α plate and β plate, respectively.

3.6. Experimental Results and Evaluation. From Figures 9 and 10, it can be found that the learning curve of the train data set and the development data set after a certain period of learning and training (about 350 time steps) begin to converge, and the aforementioned two data sets have obtained smaller MSE and MAE loss function values.

Table 3 shows that the value of the Validation set (develop), test1, and test2 are all smaller than that of the compared models through the ARIMA-LSTM ensemble model designed in our study and calculation of the MSE, RMSE, and MAE for predicted values. Therefore, it could be considered that the

accuracy of the ensemble method has been improved, and the model can be extensively used to other applications of stock market prediction.

4. Conclusion

First, the two single models have good applicability to the data with single dimension. The loss function is used to calculate the prediction results of the proposed model, and we found that both ARIMA and LSTM model have lower loss function values in stock index prediction. By comparing the loss function values of all methods, it can indicate that the three loss function indexes of LSTM model are superior to ARIMA model. Moreover, the prediction accuracy of ARIMA-LSTM hybrid model is better than other financial models. In this paper, we proposed a hybrid model ARIMA-LSTM, linearity is filtered out in ARIMA modeling, and nonlinear trends are predicted in LSTM recursive neural networks. The loss function test results show that the MSE, MAE, and RMSE of ARIMA-LSTM hybrid model are smaller than those of other control models. Therefore, ARIMA-LSTM model is feasible to predict the correlation coefficient of portfolio optimization. Although the prediction results in this paper are basically consistent with the expected results before the experiment, the time series before 2010 is not considered for only the data after 2010 are selected. Therefore, the model's ability to predict the special financial situation before 2010 need to be further tested. What is more, as financial anomalies and noise are common, all special trends cannot be covered by the model. Therefore, in the next step, it is necessary for researchers to further study how to deal with Black Swan Theory in the financial world.

Data Availability

The experimental data of this research are available from the corresponding author upon request.

Conflicts of Interest

All the authors declared that they have no conflicts of interest regarding this study.

References

[1] Z. Bao and C. Wang, "A multi-agent knowledge integration process for enterprise management innovation from the perspective of neural network," *Information Processing & Management*, vol. 59, no. 2, Article ID 102873, 2022.

[2] S. Deng, X. Huang, J. Shen, H. Yu, and C. Wang, "Prediction and trading in crude oil markets using multi-class classification and multi-objective optimization," *IEEE Access*, vol. 7, no. 99, p. 1, 2019.

[3] G. Caginalp and G. Constantine, "Statistical inference and modelling of momentum in stock prices," *Applied Mathematical Finance*, vol. 2, no. 4, 1995.

[4] T. Zheng, J. Farrish, and M. Kitterlin, "Performance trends of hotels and casino hotels through the recession: an ARIMA with intervention analysis of stock indices," *Journal of Hospitality Marketing & Management*, vol. 25, no. 1, pp. 49–68, 2016.

[5] B. Yang, C. Li, D. Wang, and X. He, "Research on the Risk of Shanghai Composite Index Based on VaR and GARCH Model," in *Proceedings of the 2017 3rd International Conference on Economics, Social Science, Arts, Education and Management Engineering (ESSAEME 2017)*, Huhhot, China, January, 2017.

[6] M. Kim and H. Sayama, "Predicting stock market movements using network science: an information theoretic approach," *Applied Network Science*, vol. 2, no. 1, p. 35, 2017.

[7] M. Khashei and Z. Hajirahimi, "A comparative study of series arima/mlp hybrid models for stock price forecasting," *Communications in Statistics-Simulation and Computation*, vol. 48, no. 9, pp. 2625–2640, 2019.

[8] I. Unggara, A. Musdholifah, and K. S. Anny, "Optimization of ARIMA forecasting model using firefly algorithm," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 2, 2019.

[9] S. Siami-Namini and A. S. Namin, "Forecasting Economics and Financial Time Series: ARIMA vs. LSTM," *Papers*, 2018, https://arxiv.org/abs/1803.06386.

[10] A. Jayanth Balaji, D. S. Harish Ram, and B. B. Nair, "Applicability of deep learning models for stock price forecasting an empirical study on BANKEX data," *Procedia Computer Science*, vol. 143, pp. 947–953, 2018.

[11] T. Joo II and C. Seung-Ho, "Stock prediction model based on bidirectional LSTM recurrent neural network," *Journal of Korea Instiute of Information, Electronics, and Communication Technology*, vol. 11, no. 2, pp. 204–208, 2018.

[12] X. Liang, Z. Ge, L. Sun, M. He, and H. Chen, "LSTM with wavelet Transform based data preprocessing for stock price prediction," *Mathematical Problems in Engineering*, vol. 2019, Article ID 1340174, 8 pages, 2019.

[13] S. Borovkova and I. Tsiamas, "An ensemble of LSTM neural networks for high-frequency stock market classification," *SSRN Electronic Journal*, vol. 01, 2018.

[14] S. Chen and L. Ge, "Exploring the attention mechanism in," *LSTM-based Hong Kong Stock price Movement Prediction*, Taylor & Francis Journals, Milton Park, UK, 2019.

[15] G. Peter and Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, 2003.

[16] C. Narendra Babu and B. Eswara Reddy, "Prediction of selected Indian stock using a partitioning-interpolation based ARIMA-GARCH model," *Applied Computing and Informatics*, vol. 11, no. 2, pp. 130–143, 2015.

[17] Y. Baek and H. Y. Kim, "ModAugNet: a new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module," *Expert Systems with Applications*, vol. 113, no. DEC, pp. 457–480, 2018.

[18] H. Y. Kim and C. H. Won, "Forecasting the volatility of stock price index: a hybrid model integrating LSTM with multiple GARCH-type models," *Expert Systems with Applications*, vol. 103, pp. 25–37, 2018.

[19] B. Chen, J. Zhong, and Y. Chen, "A hybrid approach for portfolio selection with higher-order moments: empirical

evidence from Shanghai Stock Exchange," *Expert Systems with Applications*, vol. 145, Article ID 113104, 2019.

[20] D. Opitz and R. Maclin, "Popular ensemble methods: an empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.

[21] J. J. Garcıa Adeva, U. Cervino Beresi, and R. A. Calvo, "Accuracy and diversity in ensembles of text categorisers," *CLEI Electronic Journal*, vol. 8, no. 2, pp. 1–12, 2005.

[22] M. Oliveira and L. Torgo, "Ensembles for time series forecasting," in *Proceedings of the Asian Conference on Machine Learning (ACML'2014)*, pp. 360–370, Nha Trang city, Virtnam, January, 2015.

[23] E. Dave, A. Leonardo, M. Jeanice, and N. Hanafiah, "Forecasting Indonesia exports using a hybrid model ARIMA-LSTM," *Procedia Computer Science*, vol. 179, no. 1, pp. 480–487, 2021.

[24] Y. Deng, H. Fan, and S. Wu, "A hybrid ARIMA-LSTM model optimized by BP in the forecast of outpatient visits," *Journal of Ambient Intelligence and Humanized Computing*, 2020.

[25] Z. Wang, J. Qu, X. Fang, H. Li, T. Zhong, and H. Ren, "Prediction of early stabilization time of electrolytic capacitor based on ARIMA-Bi_LSTM hybrid model," *Neurocomputing*, vol. 403, 2020.

[26] M. Khashei and M. Bijari, "Improving forecasting performance of financial variables by integrating linear and nonlinear ARIMA and artificial," *QJER*, vol. 8, no. 2, pp. 83–100, 2008.