IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 26, 2020, accepted June 18, 2020, date of publication June 22, 2020, date of current version July 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3004284

Forecasting Stock Prices Using a Hybrid Deep Learning Model Integrating Attention Mechanism, Multi-Layer Perceptron, and Bidirectional Long-Short Term Memory Neural Network

QIAN CHEN, WENYU ZHANG, AND YU LOU
School of Information Management and Artificial Intelligence, Zhejiang University of Finance and Economics, Hangzhou 310018, China

Corresponding author: Wenyu Zhang (wyzhang@e.ntu.edu.sg)

This work was supported in part by the National Natural Science Foundation of China under Grant 51875503, and in part by the Zhejiang Natural Science Foundation of China under Grant LZ20E050001.

ABSTRACT Stock prices forecasting is a topic research in the fields of investment and national policy, which has been a challenging problem owing to the multi-noise, nonlinearity, high-frequency, and chaos of stocks. These characteristics of stocks impede most forecasting models from extracting valuable information from stocks data. Herein, a novel hybrid deep learning model integrating attention mechanism, multi-layer perceptron, and bidirectional long-short term memory neural network is proposed. First, the raw data including four types of datasets (historical prices of stocks, technical indicators of stocks closing prices, natural resources prices, and historical data of the Google index) are transformed into a knowledge base with reduced dimensions using principal component analysis. Subsequently, multi-layer perceptron is used for the fast transformation of feature space and rapid gradient descent, bidirectional long-short term memory neural network for extracting temporal features of stock time series data, and attention mechanism for making the neural network focus more on crucial temporal information by assigning higher weights. Finally, a comprehensive model evaluation method is used to compare the proposed model with seven related baseline models. After extensive experiments, the proposed model demonstrated its good forecasting performance.

INDEX TERMS Stock prices forecasting, attention mechanism, bidirectional long-short term memory neural network, multi-layer perceptron, deep learning.

I. INTRODUCTION

As is known to all, stock prices reflect the quality of national economy. However, the multi-noise, nonlinearity, high-frequency, and chaos of stock render stock forecasting one of the most difficult problems in time series [1]. Nonetheless, stock forecasting still attracts increasing attention from scholars and investors owing to its importance.

The stock price forecasting has been a highly controversial topic in the past decades. Some researchers have proposed the random work theory, according to which, stock prices could not be forecasted using historical stock data [2]–[4]. In other words, the stock prices are completely random time series and the forecasting accuracy of stock prices could not exceed 50%. The main reason behind this theory was that the researchers believed that stock prices were primarily affected by news and policy. Nevertheless, a number of researchers have not followed the random work theory owing to their experimental results; they believed that historical data as features could forecast stock prices to some extent. For example, [5]–[7] rejected the theory and elucidated their reasons in their articles, respectively.

Many types of models have been used to forecast stock prices, including statistical models, machine learning models, deep learning models, and hybrid learning models. Statistical models include autoregressive moving average, autoregressive integrated moving average (ARIMA) [8], autoregressive conditional heteroscedasticity [9], generalized autoregressive conditional heteroskedasticity [10] and so on. Although statistical models have the ability to forecast stock prices, their drawbacks are evident. For example, the ARIMA model assumes the time series to be stationary sequences,

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Tucci.

or non-stationary sequences transformed by differencing the time series [11]. The generalized autoregressive conditional heteroskedasticity model could not explain the asymmetry caused by the sign of residuals. Worst of all, statistical models could only deal with the linear relationships in nature but not the nonlinear relationships in stock time series.

Considering the nonlinearity of stock prices, machine learning models, which are good at processing nonlinear data, have been widely used in the field of stock forecasting. Many scholars used support vector machine (SVM) [12], decision tree (DT) [13], and logistic regression for stock forecasting. Cao and Tay [14] discovered that SVM with adaptive parameters achieved better forecasting results than back-propagation neural network and regularized radial basis function neural network. Pai and Lin [15] proposed a hybrid methodology that combined ARIMA with SVM model for forecasting stock prices and obtained improved results compared with any single models. Li and Tam [16] used various machine learning techniques including DT to investigate the momentum and reversal effects occurring in the stock market. Chen *et al.* [17] compared various machine learning models with a sample dimension engineering method for Bitcoin price prediction, and proved the superiority of machine learning models.

Recently, owing to the enhancement of computer processing power and the ability of deep learning to learn complex nonlinear mapping and self-adaptation for various statistical distributions [18], some deep learning models, such as artificial neural network (ANN) [19], multi-layer perceptron (MLP) [20], the group method of data handling (GMDH), convolutional neural network (CNN) [21], recurrent neural network (RNN), and long-short term memory (LSTM) neural network [22] have become a focus of recent studies in various fields including the financial field. For example, Selvamuthu *et al.* [23] used ANNs based on Levenberg-Marquardt, Scaled Conjugate Gradient and Bayesian Regularization for stock market forecasting, and verified the superiority of the method. Zhong and Enke [24] used DNN and ANN to forecast the daily return direction of the S&P 500 index, showing that its classification accuracy was higher than that of other machine learning algorithms. Shaverdi *et al.* [25] used a GMDH type-neural network based on Genetic algorithm to forecast the stock price and achieved high accuracy. Ge *et al.* [26] combined LSTM with Quantum Genetics to improve the prediction accuracy of equipments' remaining useful life. Bidirectional long-short term memory (BiLSTM) neural network is an extended variant of LSTM by taking into account the effect of subsequent inputs on the initial inputs. For example, Bin *et al.* [27] proposed a framework integrating BiLSTM to generate better global representations for videos, which enhances the recognition of lasting motions in videos, and the experimental results demonstrated the superiority of the proposed approach compared to other state-of-the-art methods. Many scholars used CNN, ANN, and MLP to forecast intraday stocks directions or stocks prices effectively [28].

Different models exhibit their respective advantages. For example, ARIMA is good at processing linear data, CNN is good at processing data with spatial dimension, and RNN have been proven effective for time series data. However, in reality, predictive issues are generally complex with multiple characteristics; therefore, the emergence of hybrid models was inevitable. Ullah *et al.* [29] proposed an action recognition method for processing the video data by combining CNN and LSTM. The method proposed by Ullah *et al.* [29] was capable of learning long-term sequences and could process lengthy videos by analyzing features for a certain time interval. Yu and Yan [30] proposed a DNN model based on phase-space reconstruction method and LSTMs for forecasting stock prices. In our previous work [31], CNN and LSTM were combined to forecast the wind speed, which performed well with the real world data.

The construction and selection of input variables are critical in deep learning. Many researchers have used technical indicators, refactored technical indicators, online data (the Google or Wikipedia indexes), or sentiment analysis of text as input variables to forecast stock prices. For example, Weng [32] built a dataset including historical stock prices, several technical indicators, counts and sentiment analysis of published news articles, the Google index, and the number of unique visitors to pertinent Wikipedia pages to forecast short-term stock prices. Cambria [33] claimed the key role of affective computing and sentiment analysis for the advancement of artificial intelligence. Xing *et al.* [34] clarified the scope of natural language based financial forecasting (NLFF), after recognizing that natural language processing techniques have been widely leveraged to predict financial markets. Picasso *et al.* [35] used indicators of technical analysis and the sentiment of news articles to forecast the stock prices, which achieved good forecasting results. However, the current study mainly focuses on forecasting stocks prices using a hybrid deep learning model with numerical data (historical prices of stocks, technical indicators of stocks closing prices, natural resources prices, and historical data of the Google index), because there is a more intuitive relation between numerical data and stock prices.

Meanwhile, principal component analysis (PCA) [36], least absolute shrinkage and selection operator (LASSO) [37], and other dimension reduction methods have been widely used to reduce the computational complexity of time and space to enhance the learning effectiveness. For example, by comparing PCA and LASSO with the DT, Zhang *et al.* [38] demonstrated that the DT had a role in dimension reduction. Zhong and Enke [39] proved that combining the ANN with PCA yielded higher classification accuracy in forecasting the daily direction of stock market returns. Tian *et al.* [40] proposed a model using LASSO to select variables, which demonstrated superior out-of-sample predictive power in bankruptcy forecasting.

LSTM was used to forecast the stock prices with the features of a fixed length specified by researchers. With the continuous growth of the time series sequence, information loss will be inevitable and the forecasting effect will become increasingly worse. Attention mechanism (AM) [41] could be used to mitigate this problem. Recently, AM has been used widely and successfully in natural language processing, image recognition, and time series forecasting. For example, Cinar *et al.* [42] used improved AM to forecast energy consumption and weather. Xu *et al.* [43] utilized a model based on AM to deal with the image captioning tasks. Choi *et al.* [44] proposed a predictive model based on AM for healthcare.

The abovementioned studies have demonstrated good progress in their respective fields, but also have respective disadvantages. For example, the relationships between the natural resources prices and stocks prices have seldom been explored. The recent popular deep learning models are usually stacked by means of some neural network layers, whose selection are not justified according to the characteristics of different types of neural networks. To address the above disadvantages, the proposed model is designed by constructing a knowledge base involving the comprehensive relationships between the natural resources prices, stocks prices and other data, and taking advantage of different neural networks to process different types of data. Compared with ARIMA, CNN, GMDH and other approaches, MLP can achieve faster transformation of feature space and faster gradient descent because of its good capability in parallel data processing. Through the gate control structure and bidirectional transmission mechanism, BiLSTM can determine whether the input should be remembered or forgotten, and integrate the initial and subsequent temporal features. Therefore, a novel hybrid deep learning model integrating AM, MLP, and BiLSTM is proposed in this study to overcome the shortcomings of existing deep learning structures. Specifically, MLP is used to deal with the fast transformation of feature space so as to achieve the rapid algorithmic convergence. BiLSTM is used to extract temporal features of stock time series data. AM is used to increase the weight of crucial information in the neural network for more efficient reasoning.

The remainder of this paper is organized as follows. Section 2 describes the construction of the knowledge base. Section 3 provides a brief overview of the deep learning modules used and the proposed model. Section 4 presents the experimental results with some discussions. Our conclusions and suggestions for future research are presented in Section 5.

II. CONSTRUCTION OF KNOWLEDGE BASE

In the study, a knowledge base was constructed as the prerequisite to forecast the closing prices of stocks so that most salient predictors could be included in the proposed model. The knowledge base integrates the following: (a) historical prices of stocks; (b) technical indicators of stocks closing prices; (c) natural resources (gold, silver, and oil) prices and (d) historical data of the Google index. The historical prices of stocks were derived from the daily opening, highest, lowest, and closing prices of four stock indexes: S&P 500, NASDAQ, Russell 2000, and Dow Jones. The knowledge base used to forecast each stock index has combined all four types of dataset, so that the salient features can be considered comprehensively to improve the forecasting performance of the developed model. The S&P 500 index was used as the main illustrative research object. The remaining three stock indexes were used for testing and comparison to evaluate the robustness and generalization capability of the model.

TABLE 1. Descriptive statistics of closing prices of the S&P 500 index.

Statistics	S&P 500
Mean	1792.441
Min	676.530
Median	1837.880
Max	3013.770
Standard deviation	605.983
Skewness	0.209
Kurtosis	-1.102
Normality	158.214***
Stationarity	0.495

Note: *** denotes statistical significance at 1%.

The historical closing prices of each stock index, obtained from Yahoo Finance [45], comprise 2734 records from September 2, 2008 to July 12, 2019. Descriptive statistics such as minimum, maximum, mean, median, standard deviation, skewness, kurtosis, normality, and stability are listed in Table 1. Among them, the Jarque-Bera test [46] was used to analyze the normality of the data, and the stability was analyzed using the augmented Dickey-Fuller (ADF) test [47]. The null hypothesis of the Jarque-Bera test is that the data conform to the normal distribution. However, the result of normality under 1% significance level is 158.214, which rejects the null hypothesis. Therefore, the closing prices of the S&P 500 index do not conform to the normal distribution. Meanwhile, the value of the ADF test statistic is 0.495, which is much higher than −3.432, i.e., the p-value under 1% significance level. It indicates that the null hypothesis of the existence of unit root cannot be rejected, that is, the data are nonstationary. The abnormal distribution and nonstationary nature of the closing prices of the S&P 500 index illustrate the difficulty of stock prices forecasting.

The second type of dataset in the knowledge base are technical indicators of stocks closing prices. For each stock index, 27 technical indicators including simple moving average (SMA), relative strength index (RSI), etc. were constructed. The technical indicators reflect the market behavior through the rise, fall, and trend of stock prices; however, not all technical indicators could improve the performance of stock prices forecasting. Therefore, the Pearson correlation coefficient [48] was used to select the salient technical indicators in this study. Table 2 shows that the correlations between the technical indicators 1-8 (or 22-24) and closing prices are very close, but it does not mean the different technical indicators 1-8 (or 22-24) are highly correlated with each other. In the stock market, for example, it is generally believed that the prices of stocks will change towards different directions when Bollinger higher band, Bollinger middle band and Bollinger lower band have different trends. Under the circumstances, 14 technical indicators whose correlation coefficient with

TABLE 2. Selecting technical indicators of S&P 500 index closing prices through the Pearson correlation coefficient.

	Technical indicators	Pearson correlation coefficient	Selected (1: Yes, 0: No)
1	5-day simple moving average (SMA) of closing prices	0.999522	1
2	10-day SMA of closing prices	0.998941	1
3	15-day SMA of closing prices	0.998402	1
4	20-day SMA of closing prices	0.997899	1
5	5-day exponential moving average (EMA) of closing prices	0.999688	1
6	10-day EMA of closing prices	0.999262	1
7	15-day EMA of closing prices	0.998861	1
8	20-day EMA of closing prices	0.998476	1
9	Relative strength index (RSI)	0.150152	0
10	Chande momentum oscillator (CMO)	0.150152	0
11	Commodity channel index (CCI)	0.080084	0
12	Moving average convergence divergence (MACD)	0.203426	0
13	9-period EMA of MACD	0.210372	0
14	Rate of change	0.061459	0
15	Percentage price oscillator (PPO)	0.094942	0
16	Triangular moving average (TMA)	0.996418	1
17	Slow stochastic %K	0.079836	0
18	Slow stochastic %D	0.084849	0
19	Fast stochastic %K	0.023293	0
20	Fast stochastic %D	0.027772	0
21	Chaikin A/D oscillator	-0.033183	0
22	Bollinger higher band	0.998881	1
23	Bollinger middle band	0.999522	1
24	Bollinger lower band	0.999252	1
25	Highest high	0.996121	1
26	Lowest low	0.996121	1
27	William's %R	0.112234	0

closing prices is greater than 0.7 were selected as candidate features for forecasting.

The third type of dataset in the knowledge base includes the daily opening, highest, lowest, and closing prices of natural resources (gold, silver, and oil) in dollars [49]. They are affected by global factors such as supply and demand, United States dollar exchange rate, political factors, and inflation. Because the stock prices in the same environment are a form of global measurement index that are affected by these global factors, the prices of the natural resources exhibit a strong homogeneity with the stock prices and can be used as important features to forecast stock prices.

The fourth type of dataset in the knowledge base is the daily search flow of relevant stocks on the Google index [50], which reflects the daily popularity of stocks.

By integrating the four types of datasets above, the knowledge base comprises 31 salient features that are used as promising predictors in the proposed model for forecasting each stock index. The construction of knowledge base reduces the risk of model overfitting and increases the generalization capability of model. Owing to the large dimension of data in the knowledge base, PCA [51], one of the most widely used data dimension reduction methods, was chosen to reduce the dimension of data to smaller set of salient features in the proposed model, because PCA had been proven capable of reducing the computational complexity and improving the accuracy of stock forecasting [52]. Otherwise, the algorithmic convergence of the model will be slow, and even the problem of model degradation will occur.

Compared to variables with low variances, PCA tends to give more emphasis on the variables with high variances. In order to reduce the effect of variables with larger magnitude, data must be normalized according to Equation (1) before using PCA:

$$X_i^* = (X_i - mean)/\sigma \tag{1}$$

where i indicates the index of the variables. σ and mean indicate the standard deviation and the mean of the variables, respectively.

The main purpose of PCA is to map n-dimensional features to k-dimensional features, where k is smaller than n. PCA could also reduce the possibility of model overfitting. The k-dimensional features based on n-dimensional raw data are brand new orthogonal features, known as principal components (PCs). The detailed calculation for PCA can be found in Qiu *et al.* [53]. The obtained different values of k for the proposed model will be discussed in Section 4.

III. SOLUTION METHODOLOGY

This section explores the proposed deep learning model, before which, the underlying deep learning structures including MLP, bidirectional long-short term memory neural network (BiLSTM), and attention mechanism (AM) as well as some model evaluation indicators are described.

A. MULTI-LAYER PERCEPTRON

The simplest MLP [54] comprises three layers: an input layer, a hidden layer, and an output layer. Each layer contains some neurons, and the input of each neuron of the latter layer corresponds to the output of the former layer multiplied by the weight matrix followed by adding the bias matrix. Nonlinear activation functions are used to transmit the calculation results of each layer. The output of MLP with a three-layer structure can be expressed by Equation (2) [55]:

$$O_{MLP}=f^{o}\left[\sum_{j=1}^{J} W_{pj}\times f^{h}\left(\sum_{i=1}^{I} W_{ji}X_i+\xi_j^h\right)+\xi_p^o\right] \quad (2)$$

where, i, j, and p are the indexes of input, hidden, and output layers neurons, respectively. f^o and f^h are the activation functions in output and hidden layers, respectively. I and J are the numbers of input layer and hidden layer neurons, respectively. X_i is the input matrix. W_{pj} is the weight matrix connecting the jth neuron of the hidden layer and the pth neuron of the output layer. W_{ji} is the weight matrix connecting the ith neuron of the input layer and the jth neuron of the hidden layer. ξ_j^h is the bias matrix of the jth neuron in the hidden layer. ξ_p^o is the bias matrix of the pth neuron in the output layer.

B. BIDIRECTIONAL LONG-SHORT TERM MEMORY

RNN [56] is a deep learning model that is good at processing time series data because the internal state of RNN could display dynamic temporal features. However, as the information interval (the fixed length specified) increases, the possibility of gradient disappearance arises, which is caused by the successive multiplication with the reciprocal of the tanh (between 0 and 1) function and the weight matrix [57]. LSTM [58], as an extended variant of RNN, could alleviate the phenomenon of gradient disappearance in conventional RNN effectively. LSTM determines whether the input should be remembered or forgotten through gate control structure, and could make use of the information of long-time sequence to some extent.

LSTM replaces the neurons of RNN with blocks of memory and three types of gates (input, forget, and output gates). The information calculation in blocks of memory of LSTM can be expressed by Equations (3-8) [59].

$$g_t^f = \sigma\left(A^f X_t + B^f h_{t-1} + \xi^f\right) \quad (3)$$
$$g_t^i = \sigma\left(A^i X_t + B^i h_{t-1} + \xi^i\right) \quad (4)$$
$$m_t^c = tanh\left(A^m X_t + B^m h_{t-1} + \xi^m\right) \quad (5)$$
$$c_t = g_t^f * c_{t-1} + g_t^i * m_t^c \quad (6)$$
$$g_t^o = \sigma\left(A^o X_t + B^o h_{t-1} + \xi^o\right) \quad (7)$$
$$h_t = g_t^o * tanh\left(c_t\right) \quad (8)$$

where g_t^f, g_t^i, and g_t^o are the forget, input, and output gates at time t, respectively. m_t^c, c_t, and h_t are the candidates of input to be stored, memory cells, and hidden state at time t, respectively. X_t represents the input vectors at time t. ξ^f, ξ^i, ξ^o, and ξ^m are bias vectors of the forget, input, and output gates and the candidates of the input, respectively. A^f, A^i, A^o, A^m, B^f, B^i, B^o, and B^m are the corresponding weight matrices. $*$ denotes the Hadamard product [60] between two matrices. σ and *tanh* are activation functions.

The forecasting results are influenced by not only the initial inputs, but also the subsequent inputs in some regression problems. BiLSTM can improve the forecasting accuracy by integrating the initial and subsequent inputs. BiLSTM is divided into forward LSTM and backward LSTM. The final output of BiLSTM is determined by the results of both forward and backward calculations, whose structures are consistent with the structure of blocks of memory of LSTM.

C. ATTENTION MECHANISM

Recently, AM has become a significant topic in deep learning. It is applied to make the neural network focus more on the crucial temporal information by assigning higher weights.

The implementation of AM can be expressed by Equations (9-11) [61]:

$$S^t = F\left(X^t, C^{t-1}, \gamma^{t-1}\right) \quad (9)$$
$$\gamma_i^t = e^{S_i^t} / \sum_{i=1}^{I} e^{S_i^t} \quad (10)$$
$$F_i^t = \gamma_i^t X_i^t \quad (11)$$

where S^t is the attention score at time t, i and I mean the index and number of attention scores at time t. Equation (9) implies that S^t is determined by the input variable (X^t), previous state (C^{t-1}), and previous attention weight (γ^{t-1}). S_i^t, γ_i^t, and X_i^t are the ith instances of S^t, γ^t, and X^t at time t, respectively. F_i^t is the ith weighted feature at time t, which is determined by both the attention weight and input variable.

D. MODEL EVALUATION

Suitable evaluation indicators are required for verifying the deep learning model. After comparison, the following six indicators are selected in this study: mean absolute error (MSE), explained variance score (EVS), mean absolute error (MAE), mean squared log error (MSLE), median absolute error (MedAE), and R^2 score (R^2). The prediction value would be more accurate with lower values of MSE, MAE, MSLE, and MedAE, or higher values of EVS and R^2. The model evaluation method utilizing the six indicators is comprehensive and considers the respective emphasis of different indicators. The specific calculation of each indicator is shown in Equations (12-17) [62]:

$$MSE = \frac{1}{n} \times \left[\sum_{i=1}^{I} (y_i - p_i)^2\right] \quad (12)$$
$$EVS = 1 - [\mathrm{var}(y-p)/\mathrm{var}(y)] \quad (13)$$
$$MAE = \frac{1}{n} \times \left(\sum_{i=1}^{I} |y_i - p_i|\right) \quad (14)$$
$$MSLE = \frac{1}{n} \times \sum_{i=1}^{I} \left[\log(y_i+1) - \log(p_i+1)\right]^2 \quad (15)$$
$$MedAE = median\left(|y_i - p_i|\right) \quad (16)$$
$$R^2 = 1 - \left[\sum_{i=1}^{I} (y_i - p_i)^2 / \sum_{i=1}^{I} (y_i - \overline{y_i})^2\right] \quad (17)$$

where i and I indicate the index and number of forecasting days, respectively, and y and p indicate the true values and forecasting values, respectively. $\bar{y}$ indicates the mean value of y. *var*, *log*, and *median* mean the functions of calculating the variance, logarithm, and median of the data, respectively.

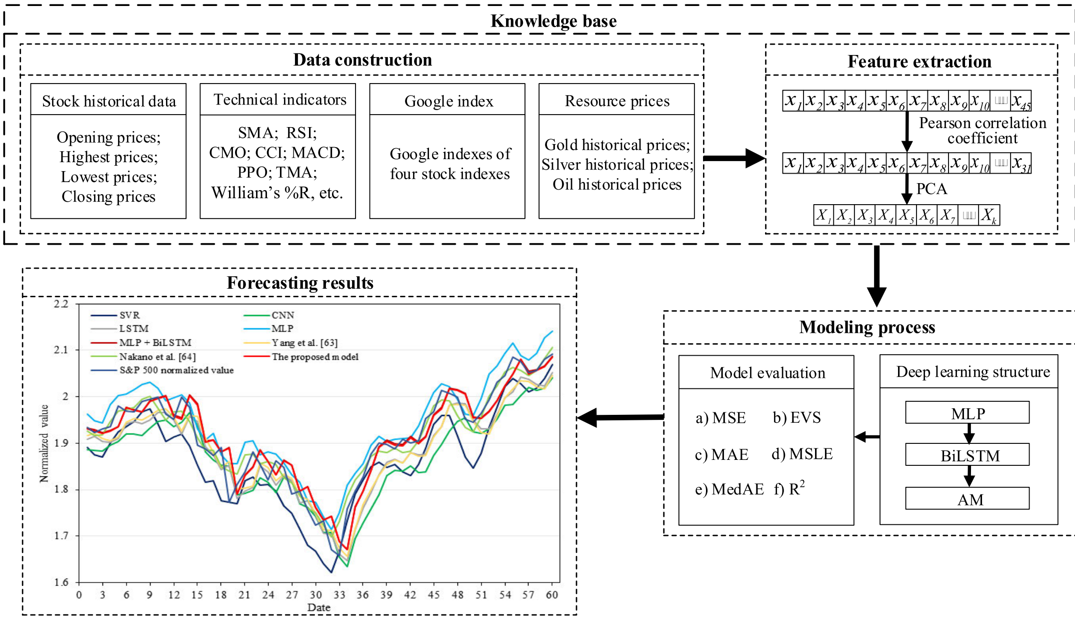

FIGURE 1. An overview of the proposed model.

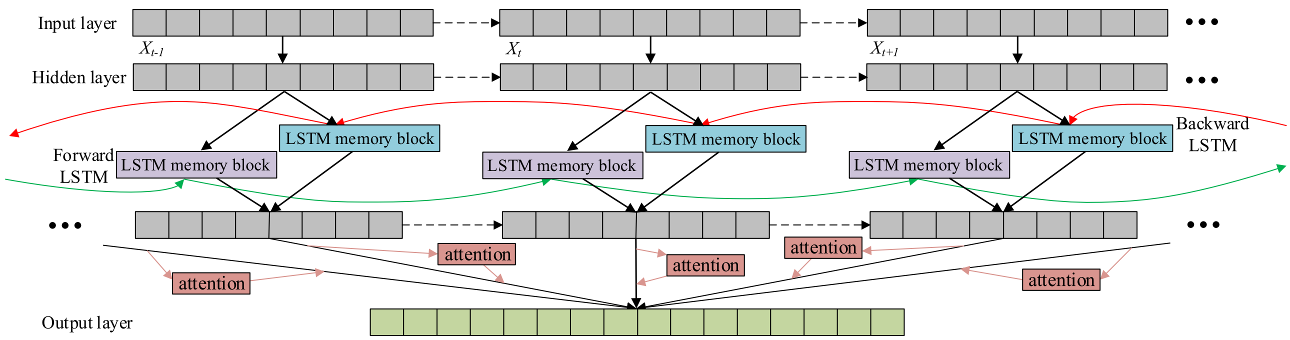

FIGURE 2. Deep learning structure of the proposed model.

E. THE PROPOSED MODEL

Considering that the traditional models could not extract the valuable information from stocks data effectively, the current study proposes a hybrid deep learning model integrating AM, MLP, and BiLSTM, by utilizing their respective advantages and avoiding their respective disadvantages. For example, if BiLSTM is used to process the long-time sequence input directly, the convergence of the model will be slow. So, MLP and BiLSTM are combined in this study to expedite the transformation of feature space and algorithmic convergence. A knowledge base was constructed for the proposed model to utilize most of the salient predictors. Figure 1 shows an overview framework of the proposed model, which is explained as follows.

Four types of datasets (historical prices of stocks, technical indicators of stocks closing prices, natural resources prices, and historical data of the Google index) were integrated into a knowledge base that includes 45 candidate features. After screening the technical indicators through Person correlation coefficient analysis, 31 salient features were left. Subsequently, PCA was used to reduce the dimension of the data to k salient features so that the computational complexity of time and space of the proposed model can be further reduced. The value of k will be discussed in Section 4, and the details of the knowledge base can be found in Section 2.

A deep learning structure combining MLP and BiLSTM was constructed in the study, which is shown in Figure 2. MLP expedites the transformation of feature space and gradient descent so as to achieve the rapid algorithmic convergence. BiLSTM is good at extracting the temporal features of stocks time series data. By combining MLP and BiLSTM, the features, in particular, the temporal features of stock data can be extracted effectively with rapid gradient descent. The effective extraction of temporal features plays an important role in improving the forecasting accuracy. In addition, AM is integrated with MLP and BiLSTM to make the neural networks focus more on the crucial temporal information by assigning higher weights. Therefore, different weights can

TABLE 3. Comparison of the forecasting results corresponding to different *k* values.

Evaluation indicator	No. of principal components (k)							
	1	2	5	10	15	20	25	No PCA
EVS	0.37832	0.31261	0.85277	**0.87062**	0.84354	0.86696	0.85149	0.84692
MAE	0.04650	0.04770	0.03076	**0.02565**	0.02870	0.02855	0.03002	0.03082
MSE	0.00319	0.00338	0.00157	**0.00124**	0.00156	0.00145	0.00154	0.00156
MSLE	0.00039	0.00041	0.00019	**0.00015**	0.00019	0.00017	0.00019	0.00019
MedAE	0.03809	0.04054	0.02691	**0.01933**	0.02184	0.02325	0.02387	0.02451
R^2	0.37777	0.31248	0.84922	**0.86777**	0.83316	0.85594	0.83620	0.83976

Note: Significant values are boldfaced.

be assigned to different input sequences, thus improving the forecasting accuracy further. Additionally, a comprehensive model evaluation method was constructed. Six evaluation indexes, seven baseline models, and four different stock indexes were used to evaluate the superiority and robustness of the proposed model.

As shown in Figure 1 and Figure 2, an input layer with 180 neurons was used to input the salient data outputted from the knowledge base. A hidden layer with 25 neurons followed the input layer. The temporal features of the time series data were subsequently captured by BiLSTM. The BiLSTM layer was connected to the hidden layer of MLP with 750 neurons. Next, AM was used to enhance the forecasting accuracy by assigning higher weights to the crucial temporal information. Finally, the forecasting results were outputted. The learning rate was set to 0.0001, and *relu* was used as the activation function of the neural network. MSE was used as the loss function. The batch size and epoch were set to 1024 and 2400, respectively. The setting of the parameters in the proposed model was based on empiricism, and the parameters were determined through trial-run experiments. The model is implemented in Python 3.7 environment. A computer with an Intel Core i7-8700k CPU at 3.7 GHz and the GPU of GeForce RTX 2080 is used to train the model.

IV. RESULTS OF THE EXPERIMENTS AND DISCUSSION

The results of the experiments are presented and discussed in this section. The knowledge base includes 2734 records in units of days. The last 60 records from April 17, 2019 to July 12, 2019 were used as the test set through six evaluation indicators; 10% of the data (267 records from March 26, 2018 to April 16, 2019) were used as the verification set, while the remaining of the data (2407 records from September 2, 2008 to March 25, 2018) were used as the training set for the deep learning model. All the results were obtained by averaging the running results of the model in 20 times.

A. COMPARISON WITH DIFFERENT NUMBERS OF PRINCIPAL COMPONENTS

As described in Section 2, PCA was used to reduce the data dimension to k, that is, to select k principal components. The forecasting results of the proposed model corresponding to different k values were compared through six evaluation indicators, as shown in Table 3. The forecasting results were obtained based on the S&P 500 index with the time horizon set to one day. The values of the evaluation indicators with a certain k value are boldfaced if they outperform that of evaluation indicators with other k values.

As shown in Table 3, the forecasting accuracy does not increase or decrease linearly with the increase or decrease of k. The model performs the best when $k = 10$ and the worst when $k = 2$. The normalized EVS, MAE, MSE, MSLE, MedAE, and R^2 of the forecasting results when $k = 10$ are improved by 178.5%, 46.2%, 63.3%, 63.4%, 52.3%, and 177.7%, respectively, than when $k = 2$. Table 3 shows not only that the forecasting accuracy when $k = 10$ performs better than that when no PCA is applied on the model, but also indicates that the forecasting accuracy when $k = 2$ performs much worse than that when no PCA is applied. It can be concluded that using PCA with appropriate k values improves the forecasting accuracy effectively, while using PCA with inappropriate k values reduces the forecasting accuracy. Therefore, $k = 10$ was used in the proposed deep learning model.

B. EFFECT OF ATTENTION MECHANISM

The comparison of the forecasting results corresponding to different time horizons (1, 2, 3, 4, and 5 days) of stock trading days in a week are shown in Table 4 through six evaluation indicators. The values of evaluation indicators are boldfaced if the proposed model with attention mechanism (AM) applied performs better than or at least the same as when AM is not applied. The S&P 500 index was validated in this experiment. It was discovered that the model with AM applied generally demonstrated superior performance compared with the model without AM applied, especially when the time horizons were 1 and 3 days. In most time horizons, the performance comparison between models with and without AM applied indicated a subtle difference. The experiment verifies that AM could improve the forecasting accuracy obviously by focusing more on crucial temporal information. In other words, the effect of AM is prominent in the field of stock prices forecasting. The following experiments are implemented with the time horizon set to one day.

C. COMPARISON WITH DIFFERENT STOCK INDEXES

Table 5 shows the comparison of the forecasting accuracies corresponding to different stock indexes through six

TABLE 4. Comparison of the forecasting results corresponding to different time horizons and the use of attention mechanism.

Time horizons (days)	AM	Evaluation indicator					
		EVS	MAE	MSE	MSLE	MedAE	R^2
1	with AM	**0.869727**	**0.025393**	**0.001204**	**0.000148**	**0.017290**	**0.869326**
	without AM	0.869217	0.025827	0.001230	0.000151	0.018968	0.867302
2	with AM	0.856970	**0.027326**	**0.001308**	**0.000160**	0.020724	**0.855778**
	without AM	0.857234	0.027704	0.001341	0.000164	0.020646	0.854065
3	with AM	**0.858830**	**0.028790**	**0.001372**	**0.000167**	**0.023246**	**0.855398**
	without AM	0.852666	0.029319	0.001424	0.000173	0.025072	0.845693
4	with AM	**0.859917**	**0.029337**	**0.001427**	**0.000174**	0.026230	**0.843090**
	without AM	0.848510	0.030261	0.001490	0.000181	0.025965	0.835682
5	with AM	0.847211	**0.028297**	0.001429	**0.000174**	**0.020735**	0.845298
	without AM	0.851152	0.028723	0.001427	0.000174	0.021061	0.846216

Note: Significant values are boldfaced.

TABLE 5. Comparison of the forecasting results corresponding to different stock indexes.

Stock indexes	AM	Evaluation indicator					
		EVS	MAE	MSE	MSLE	MedAE	R^2
S&P 500	with AM	**0.869727**	**0.025393**	**0.001204**	**0.000148**	**0.017290**	**0.869326**
	without AM	0.869217	0.025827	0.001230	0.000151	0.018968	0.867302
Dow Jones	with AM	**0.880160**	**0.027087**	**0.001368**	**0.000158**	0.019647	**0.879714**
	without AM	0.878234	0.027153	0.001375	0.000158	0.019325	0.877560
NASDAQ	with AM	0.859265	0.032853	**0.001951**	**0.000214**	0.023811	**0.855379**
	without AM	0.859442	0.032829	0.001974	0.000216	0.023097	0.855227
Russell 2000	with AM	**0.789047**	**0.034149**	**0.001965**	**0.000314**	**0.028010**	**0.788701**
	without AM	0.786465	0.034596	0.001993	0.000318	0.028203	0.785430

Note: Significant values are boldfaced.

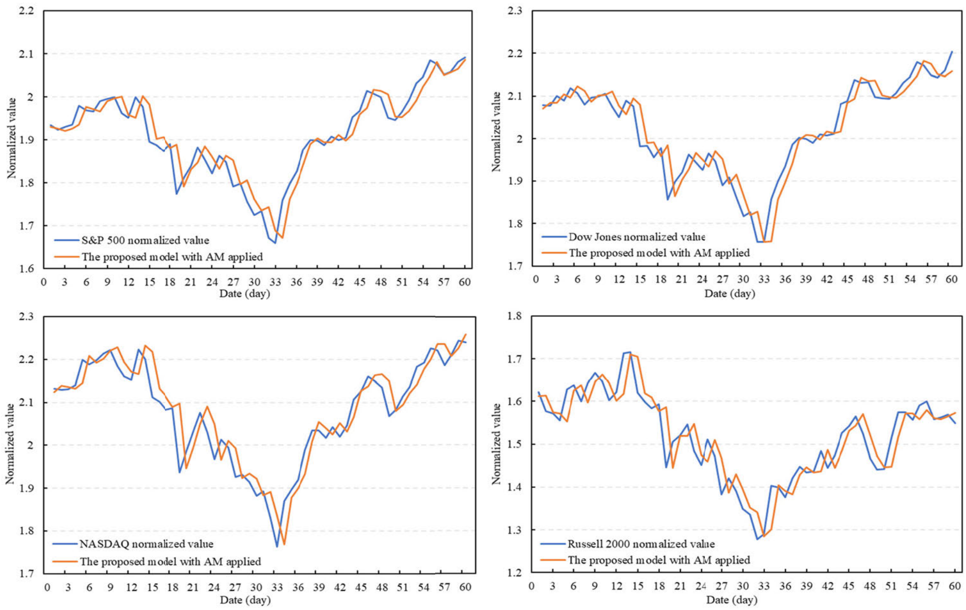

FIGURE 3. Forecasting results of the proposed model with AM applied and actual normalized values of different stock indexes over time.

evaluation indicators. The values of evaluation indicators are boldfaced if the proposed model performs better than or at least the same as when AM is not applied. In terms of the S&P 500 and Russell 2000 indexes, the proposed model demonstrated obvious improvement on forecasting accuracy when AM was applied, with all the evaluation indicators outperforming those when AM was not applied. Generally, for all four stock indexes, the forecasting accuracies of the proposed model with AM applied outperformed those without AM applied.

Figure 3 and Figure 4 show the comparison among the corresponding forecasting results of the proposed model with and without AM applied and the actual normalized values of four stock indexes. The experimental results indicate that the forecasting results of the proposed model with AM applied are closer to the actual values. The proposed model has achieved good forecasting results on the test sets of four different stock indexes, also indicating that the proposed model's possibility of overfitting is very low.

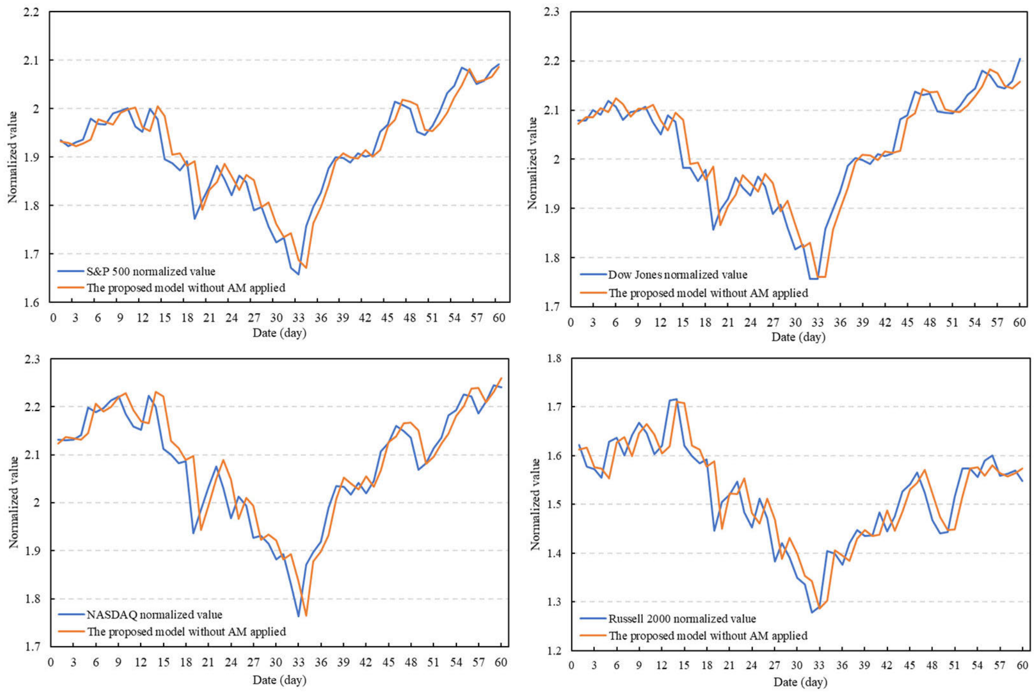

FIGURE 4. Forecasting results of the proposed model without AM applied and actual normalized values of different stock indexes over time.

TABLE 6. Performance of different models with six evaluation indicators.

Methods	EVS	MAE	MSE	MSLE	MedAE	R^2
SVR	**0.943496**	0.049373	0.003009	0.000359	0.046726	0.716351
LSTM	0.853031	0.039465	0.002207	0.000266	0.033488	0.760591
CNN	0.703118	0.072582	0.006661	0.000792	0.073571	0.006610
MLP	0.901269	0.041324	0.002882	0.000338	0.038708	0.719088
MLP + BiLSTM	0.869217	0.025827	0.001230	0.000151	0.018968	0.867302
Yang et al. [63]	0.828717	0.039052	0.002147	0.000257	0.034421	0.752297
Nakano et al. [64]	0.903007	0.028608	0.001320	0.000161	0.024855	0.845055
The proposed model	0.869727	**0.025393**	**0.001204**	**0.000148**	**0.017290**	**0.869326**

Note: Significant values are boldfaced.

TABLE 7. DM test results of the performance comparison between the proposed model and different baseline models.

Models	SVR	LSTM	CNN	MLP	Yang et al. [63]	Nakano et al. [64]
DM test results	3.7241***	3.1240***	4.8363***	0.3170	3.3048***	-2.6383

Note: *** represents 1% significance level.

D. COMPARISON WITH THE BASELINE PREDICTIVE MODELS

The proposed model was compared with support vector regression (SVR), LSTM, CNN, MLP, multi-layer perceptron and bidirectional long-short term memory neural network (MLP + BiLSTM), and other baseline deep learning models proposed by Yang *et al.* [63] and Nakano *et al.* [64], and the comparison results are shown in Table 6. The values of evaluation indicators are boldfaced if the corresponding model performs better than other models. The experimental results were obtained based on the S&P 500 index. SVR performs the best in terms of the EVS among all models. The proposed model performs the best in terms of the MAE, MSE, MSLE, MedAE, and R^2 among all the eight models, while subtly worse than SVR in terms of the EVS. It was discovered that MLP + BiLSTM could improve the forecasting results of MLP effectively, indicating that BiLSTM is good at processing time series data. However, the proposed model still performs better than MLP + BiLSTM. The forecasting results of all models are shown in Figure 5.

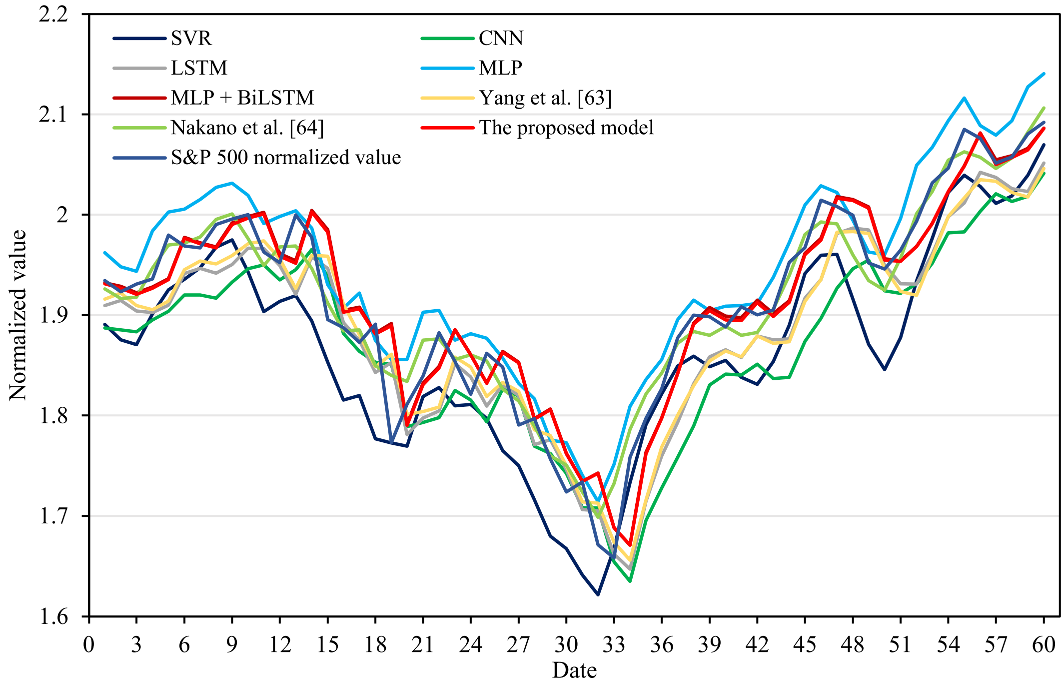

FIGURE 5. Forecasting results of different models and actual normalized values on the S&P 500 index over time.

In addition, the Diebold-Mariano [65] test (DM test) is used to test the statistical significance of the performance comparison between the proposed model and different baseline models, as shown in Table 7. It is found that the proposed model significantly outperforms the baseline models except MLP and Nakano *et al.*'s model. However, Table 6 has shown that the proposed model performs better than MLP and Nakano *et al.*'s model in terms of five of six evaluation indicators.

V. CONCLUSION AND FUTURE WORKS

Stock forecasting is a topic that has attracted significant attention for a long time. Multi-noise, nonlinearity, high-frequency, and chaos of stocks impede most forecasting models from extracting valuable information from stocks data. Hitherto, a system that could forecast stock prices perfectly has not been developed. Therefore, a novel hybrid deep learning model integrating attention mechanism (AM), MLP, and bidirectional long-short term memory neural network (BiLSTM) was proposed to forecast the closing prices of four stock indexes due to their respective advantages mentioned in Section 3. A new knowledge base was constructed to include the historical prices of stocks, technical indicators of stocks closing prices, natural resources prices, and historical data of the Google index. PCA was applied to reduce the 31 candidate features to 10, which implied that the computational complexity of time and space was reduced to more than a half but a better predictive accuracy was achieved.

The robustness of the proposed model was proven through testing on four stock indexes (S&P 500, Dow Jones, NASDAQ, and Russell 2000). Based on the results, the proposed model performed the best in terms of the MAE, MSE, MSLE, MedAE, and R^2. The results demonstrated the good forecasting performance of the proposed model and that AM was especially prominent in the field of stock prices forecasting.

Although the proposed model herein could obtain good forecasting accuracy, some shortcomings still exist. First, the parameters of the deep learning model were set through trial-run experiences. In future studies, evolutionary algorithms should be used (GA, PSO, etc.) to select the salient features of the model self-adaptively and obtain the optimal parameters for the deep learning model. Next, only the basic AM (soft attention) was used in the study and the internal structure of AM was not improved, which would have otherwise reduced the computing time of the deep learning model. Hence, in future studies, stacked latent attention should be applied to investigate the internal structure of AM. The future studies will also be considered to take advantage of both numerical data and textual data, and employ natural language processing methods including sentiment analysis, sentiment embeddings, and natural language analysis to further improve the forecasting performance of the model. Besides, we will explore the combination of the proposed approaches with other neural networks such as GMDH.

SUPPORTING INFORMATION

The raw data has been uploaded to Figshare (https://doi.org/10.6084/m9.figshare.12045675). The raw data are divided into four parts: historical prices of stocks, technical indicators of stocks closing prices, natural resources (gold, silver, and oil) prices, and historical data of the Google index. These raw data were downloaded from Yahoo Finance.

REFERENCES

[1] Y. S. Abu-Mostafa and A. F. Atiya, "Introduction to financial forecasting," *Appl. Intell.*, vol. 6, no. 3, pp. 205–213, 1996.
[2] P. H. Cootner, *The Random Character of Stock Market Prices*. Cambridge, MA, USA: MIT Press, 1967.
[3] E. F. Fama, L. Fisher, M. C. Jensen, and R. W. Roll, "The adjustment of stock prices to new information," *Int. Econ. Rev.*, vol. 10, no. 1, pp. 1–21, 1969.
[4] E. F. Fama, "Random walks in stock market prices," *Financial Anal. J.*, vol. 51, no. 1, pp. 75–80, Jan. 1995.
[5] B. G. Malkiel, "The efficient market hypothesis and its critics," *J. Econ. Perspect.*, vol. 17, no. 1, pp. 59–82, Feb. 2003.
[6] J. R. Nofsinger, "Social mood and financial economics," *J. Behav. Finance*, vol. 6, no. 3, pp. 144–160, Sep. 2005.
[7] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, Mar. 2011.
[8] G. E. P. Box, G. M. Jenkins, and J. F. MacGregor, "Some recent advances in forecasting and control," *J. Roy. Stat. Soc. C, Appl. Stat.*, vol. 23, no. 2, pp. 158–179, 1974.
[9] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1007, 1982.
[10] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *J. Econometrics*, vol. 31, no. 3, pp. 307–327, Apr. 1986.
[11] R. Y. M. Li, S. Fong, and K. W. S. Chong, "Forecasting the REITs and stock indices: Group method of data handling neural network approach," *Pacific Rim Property Res. J.*, vol. 23, no. 2, pp. 123–160, May 2017.
[12] L. J. Cao and F. E. H. Tay, "Support vector machine with adaptive parameters in financial time series forecasting," *IEEE Trans. Neural Netw.*, vol. 14, no. 6, pp. 1506–1518, Nov. 2003.
[13] R. J. Quinlan and J. M. Cherrett, "The role of fungus in the diet of the leaf-cutting ant *Atta cephalotes (L.)*," *Ecol. Entomol.*, vol. 4, no. 2, pp. 151–160, May 1979.
[14] L. Cao and F. E. H. Tay, "Financial forecasting using support vector machines," *Neural Comput. Appl.*, vol. 10, no. 2, pp. 184–192, May 2001.
[15] P.-F. Pai and C.-S. Lin, "A hybrid ARIMA and support vector machines model in stock price forecasting," *Omega*, vol. 33, no. 6, pp. 497–505, Dec. 2005.
[16] Z. Li and V. Tam, "A machine learning view on momentum and reversal trading," *Algorithms*, vol. 11, no. 11, p. 170, Oct. 2018.
[17] Z. Chen, C. Li, and W. Sun, "Bitcoin price prediction using machine learning: An approach to sample dimension engineering," *J. Comput. Appl. Math.*, vol. 365, Feb. 2020, Art. no. 112395.
[18] S. C. Nayak, B. B. Misra, and H. S. Behera, "Efficient financial time series prediction with evolutionary virtual data position exploration," *Neural Comput. Appl.*, vol. 31, no. S2, pp. 1053–1074, Feb. 2019.
[19] S. C. Nayak and B. B. Misra, "Estimating stock closing indices using a GA-weighted condensed polynomial neural network," *Financial Innov.*, vol. 4, no. 1, p. 21, Dec. 2018.
[20] S. C. Nayak, B. B. Misra, and H. S. Behera, "ACFLN: Artificial chemical functional link network for prediction of stock market index," *Evol. Syst.*, vol. 10, no. 4, pp. 567–592, Dec. 2019.
[21] H. Gunduz, Y. Yaslan, and Z. Cataltepe, "Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations," *Knowl.-Based Syst.*, vol. 137, pp. 138–148, Dec. 2017.
[22] K. Chen, Y. Zhou, and F. Dai, "A LSTM-based method for stock returns prediction: A case study of China stock market," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Santa Clara, CA, USA, vol. 1, Oct./Nov. 2015, pp. 2823–2824.
[23] D. Selvamuthu, V. Kumar, and A. Mishra, "Indian stock market prediction using artificial neural networks on tick data," *Financial Innov.*, vol. 5, no. 1, p. 16, Dec. 2019.
[24] X. Zhong and D. Enke, "Predicting the daily return direction of the stock market using hybrid machine learning algorithms," *Financial Innov.*, vol. 5, no. 1, p. 4, Dec. 2019.
[25] M. Shaverdi, S. Fallahi, and V. Bashiri, "Prediction of stock price of Iranian petrochemical industry using GMDH-type neural network and genetic algorithm," *Appl. Math. Sci.*, vol. 6, no. 7, pp. 319–332, 2012.
[26] Y. Ge, L. Sun, and J. Ma, "An improved PF remaining useful life prediction method based on quantum genetics and LSTM," *IEEE Access*, vol. 7, pp. 160241–160247, 2019.
[27] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, "Describing video with attention-based bidirectional LSTM," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2631–2641, Jul. 2019.
[28] E. Hadavandi, H. Shavandi, and A. Ghanbari, "Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting," *Knowl.-Based Syst.*, vol. 23, no. 8, pp. 800–808, Dec. 2010.
[29] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2018.
[30] P. Yu and X. Yan, "Stock price prediction based on deep neural networks," *Neural Comput. Appl.*, vol. 32, no. 6, pp. 1609–1628, Mar. 2020.
[31] Y. Chen, S. Zhang, W. Zhang, J. Peng, and Y. Cai, "Multifactor spatio-temporal correlation model based on a combination of convolutional neural network and long short-term memory neural network for wind speed forecasting," *Energy Convers. Manage.*, vol. 185, pp. 783–799, Apr. 2019.
[32] B. Weng, L. Lu, X. Wang, F. M. Megahed, and W. Martinez, "Predicting short-term stock prices using ensemble methods and online data sources," *Expert Syst. Appl.*, vol. 112, pp. 258–273, Dec. 2018.
[33] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.
[34] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: A survey," *Artif. Intell. Rev.*, vol. 50, no. 1, pp. 49–73, Jun. 2018.
[35] A. Picasso, S. Merello, Y. Ma, L. Oneto, and E. Cambria, "Technical analysis and sentiment embeddings for market trend prediction," *Expert Syst. Appl.*, vol. 135, pp. 60–70, Nov. 2019.
[36] S. Thawornwong and D. Enke, "The adaptive selection of financial and economic variables for use with artificial neural networks," *Neurocomputing*, vol. 56, pp. 205–232, Jan. 2004.
[37] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., B, Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.
[38] X. Zhang, Y. Hu, K. Xie, S. Wang, E. W. T. Ngai, and M. Liu, "A causal feature selection algorithm for stock prediction modeling," *Neurocomputing*, vol. 142, pp. 48–59, Oct. 2014.
[39] X. Zhong and D. Enke, "Forecasting daily stock market return using dimensionality reduction," *Expert Syst. Appl.*, vol. 67, pp. 126–139, Jan. 2017.
[40] S. Tian, Y. Yu, and H. Guo, "Variable selection and corporate bankruptcy forecasts," *J. Banking Finance*, vol. 52, pp. 89–100, Mar. 2015.
[41] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. 21st Neural Inf. Process. Syst.*, Kuching, Malaysia, Nov. 2014, pp. 2204–2212.
[42] Y. G. Cinar, H. Mirisaee, P. Goswami, E. Gaussier, A. Aït-Bachir, and V. Strijov, "Position-based content attention for time series forecasting with sequence-to-sequence RNNs," in *Proc. 24th Neural Inf. Process. Syst.*, Guangzhou, China, Nov. 2017, pp. 533–544.
[43] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, and R. Salakhudinov, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, Jul. 2015, pp. 2048–2057.
[44] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Proc. 23rd Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 3504–3512.
[45] *Stock Index Data*. Accessed: Jul. 12, 2019. [Online]. Available: https://finance.yahoo.com
[46] C. M. Jarque and A. K. Bera, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Econ. Lett.*, vol. 6, no. 3, pp. 255–259, Jan. 1980.
[47] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *J. Amer. Stat. Assoc.*, vol. 74, no. 366, pp. 427–431, 1979.
[48] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer, 2009.
[49] *Natural Resources Data*. Accessed: Jul. 12, 2019. [Online]. Available: https://cn.investing.com
[50] *Google Index Data*. Accessed: Jul. 12, 2019. [Online]. Available: https://www.google.com
[51] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.

[52] C.-F. Tsai and Y.-C. Hsiao, "Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches," *Decis. Support Syst.*, vol. 50, no. 1, pp. 258–269, Dec. 2010.
[53] G. Qiu, Y. Gu, and Q. Cai, "A deep convolutional neural networks model for intelligent fault diagnosis of a gearbox under different operational conditions," *Measurement*, vol. 145, pp. 94–107, Oct. 2019.
[54] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.
[55] M. A. Ghorbani, R. C. Deo, V. Karimi, M. H. Kashani, and S. Ghorbani, "Design and implementation of a hybrid MLP-GSA model with multi-layer perceptron-gravitational search algorithm for monthly lake water level forecasting," *Stochastic Environ. Res. Risk Assessment*, vol. 33, no. 1, pp. 125–147, Jan. 2019.
[56] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 694–707, Apr. 2016.
[57] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 2625–2634.
[58] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[59] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: A deep learning approach for short-term traffic forecast," *IET Intell. Transp. Syst.*, vol. 11, no. 2, pp. 68–75, Mar. 2017.
[60] C. Davis, "The norm of the schur product operation," *Numerische Math.*, vol. 4, no. 1, pp. 343–344, Dec. 1962.
[61] S. Wang, X. Wang, S. Wang, and D. Wang, "Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting," *Int. J. Elect. Power Energy Syst.*, vol. 109, pp. 470–479, Jul. 2019.
[62] A.-M. Fuertes, M. Izzeldin, and E. Kalotychou, "On forecasting daily stock volatility: The role of intraday information and market conditions," *Int. J. Forecasting*, vol. 25, no. 2, pp. 259–281, Apr. 2009.
[63] H. Yang, Z. Pan, and Q. Tao, "Robust and adaptive online time series prediction with long short-term memory," *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–9, Dec. 2017.
[64] M. Nakano, A. Takahashi, and S. Takahashi, "Bitcoin technical trading with artificial neural network," *Phys. A, Stat. Mech. Appl.*, vol. 510, pp. 587–609, Nov. 2018.
[65] F. Diebold and R. Mariano, "Comparing predictive accuracy," *J. Bus. Econ. Statist.*, vol. 20, no. 1, pp. 134–144, Jan. 2002.

QIAN CHEN is currently pursuing the M.S. degree with the School of Information Management and Artificial Intelligence, Zhejiang University of Finance and Economics, China. His current research interests are deep learning and machine learning.

WENYU ZHANG received the B.S. degree from Zhejiang University, China, in 1989, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2002. He is currently a full-time Professor with the School of Information Management and Artificial Intelligence, Zhejiang University of Finance and Economics, China. He has published more than 40 articles in international journals and more than 20 papers in international conference proceedings in the recent ten years, covering supply chain management, digital library, bibliometrics, concurrent engineering, distributed manufacturing, business intelligence, business analytics, data mining, multi-agent technology, and semantic Web.

YU LOU is currently pursuing the bachelor's degree with the School of Information Management and Artificial Intelligence, Zhejiang University of Finance and Economics, China.