Exploring the Performance of the CNN-LSTM Model in Stock Prediction

Xiuze Zhao*

College of Information Science, Beijing Language & Culture University, Beijing, 100083, China

*Corresponding author: 202311580456@stu.blcu.edu.cn

Abstract. Stock prediction in the financial market helps investors formulate investment strategies. However, the stock prediction has the characteristics of high dimension and high noise, and traditional linear models have difficulty meeting this demand. Therefore, it is necessary to use deep learning algorithms with stronger processing power for prediction. This paper combines the advantages of convolutional neural network (CNN) in feature extraction and the advantages of long short-term memory network (LSTM) in large-time scale relationship prediction to construct a CNN-LSTM hybrid network. In order to evaluate the performance of the model, this paper selects the closing price of the Shanghai Stock Exchange (SSE) 50 index in the past five years for experimental prediction. This hybrid model was compared with the ARIMA model and a single neural network, and it was found that the prediction results were more accurate, only lower than the LSTM model, but the interpretability of the prediction results was lower, only higher than the CNN model. This paper can provide ideas for the improvement of CNN-LSTM in the field of prediction. It also provides a reference for investors to decide on investment strategies and create a healthier stock market.

Keywords: Deep learning, Convolutional neural network (CNN), Long short-term memory (LSTM), stock prediction.

1. Introduction

The volatility of the stock market directly affects the stability of investment returns and policy formulation. Stock market forecasts help to respond to future fluctuations in advance, but the market is affected by multiple factors such as the economic environment, policy adjustments, and investor sentiment. Its complexity and uncertainty have become the main difficulties in forecasting. In order to make the forecast more accurate, predecessors have proposed the Fama-French three-factor model and the autoregressive integrated moving average (ARIMA) model [1,2]. Although these forecasting models have a certain degree of accuracy, with further research and analysis, the interactions between factors are complex and nonlinear, making it difficult for traditional models to capture these implicit relationships [3]. At the same time, the increasing noise in the market makes traditional models too sensitive to data, which in turn affects the stability of forecasts. In the face of these problems, deep learning has become one of the solutions to improve the effectiveness of forecasts.

The development of deep learning has been widely used in image recognition, image classification, medicine, and other fields [4]. In recent years, it has also attracted much attention from scholars in terms of financial market forecasting, and its applications include models such as convolutional neural networks (CNN) and recurrent neural networks (RNN) [5,6]. Compared with traditional time series models, deep learning models have many advantages. First, they do not require complex correlation principle models and knowledge structures and can make future predictions only through training on past experience. Second, these models have strong data processing capabilities and can more effectively identify the relationship between factors. In addition, deep learning can better adapt to the high-dimensional and nonlinear characteristics of financial market data [7]. Zhang predicted the price of cryptocurrency by building a CNN-LSTM network, and Livieris predicted the price of gold by using it, both of which achieved good results [8,9].

This paper aims to explore whether the CNN-LSTM model can more accurately predict the direction of my country's two stock markets compared with the traditional ARIMA model and a single neural network. This will provide a reference for creating a healthier market and investors' investment strategies. Therefore, this paper constructs a CNN-LSTM model and describes the market's trading behavior with five factors: opening price, highest price, lowest price, closing price, and trading volume. At the same time, an ARIMA time series prediction model is constructed, and predictions are made based on multiple stocks. Finally, the prediction error results are evaluated and the possible reasons are explained.

2. Data and Methods

2.1. Data Source

The data in this article comes from the SSE 50 on Eastmoney.com. The daily closing prices for five consecutive years from September 10, 2019, to September 10, 2019 (excluding holidays) were selected as the research sample. After deleting some missing values, 1,217 sets of stock closing price data were included. The variables included the opening price, highest price, lowest price, trading volume, closing price, and time variables of each set of data.

2.2. Methods

2.2.1 CNN Model

CNN is a multi-layer neural network that is sensitive to data features. Each layer includes (1) a convolution layer for extracting features after the convolution kernel and activation function operation. (2) a pooling layer for amplifying features and avoiding overfitting. Each layer is connected through a fully connected layer and the results are reflected in the output layer. Using it for time series prediction can better extract the deep features of the data.

2.2.2 LSTM Model

LSTM is a variant of RNN. Its improvement mainly lies in the introduction of new internal states and gating mechanisms [10]. LSTM performs excellently in processing time series data. Especially when facing long-distance dependencies and nonlinear patterns in time series, LSTM can effectively capture the temporal characteristics and potential laws in time series, thereby achieving high-precision prediction.

2.2.3 ARIMA Model

The ARIMA model is one of the time series forecasting and analysis methods, and its parameters include (p, d, q). p is the number of autoregressive terms; q is the number of moving average terms, and d is the number of differences performed to make the series stationary. The ARIMA model can be expressed as:

$$(1-\sum_{i=1}^{p}\alpha_i L^i)(1-L)^d y_t = \alpha_0 + (1+\sum_{i=1}^{q}\beta_i L^i)\varepsilon_t \tag{1}$$

Where α is the coefficient of the AR model, β is the coefficient of the MA model, and L is the lag operator. Its disadvantage is that it can only predict the next few periods and cannot make long-term predictions in the future.

2.2.4 CNN-LSTM Model

The CNN-LSTM model combines the advantages of two single models, including LSTM's ability to efficiently utilize historical data and CNN's ability to reduce and extract features. This hybrid model usually uses CNN to process image features and LSTM to process time information, achieving a comprehensive analysis of multiple data types. In this article, the CNN part of the CNN-LSTM combined model is mainly responsible for the implicit features in the time series, and the LSTM part is mainly responsible for price prediction.

The prediction steps of the hybrid model in this article are as follows: First, get the raw data from the Choice financial terminal. Then, extract the five variables of the opening price, highest price, lowest price, closing price, and trading volume as factors. Then, standardize the data. Next, Utilize the initial 80% of the data collection for training purposes, with the subsequent 20% designated as a validation set to assess the model's efficacy. After that, the data will be flattened into one dimension and a hybrid neural network constructed. Finally, substitute the data to draw the image and calculate the root mean square error (RMSE) and determination coefficient (R square)

2.3. Evaluation Metrics

The evaluation indicators of this paper include RMSE and R-squared.

RMSE is expressed as:

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_{pre,i}-X_{r,i})^2} \tag{2}$$

R-squared is expressed as:

$$\frac{\sum\left(X_{pre}-\overline{X}\right)^2}{\sum\left(X_r-\overline{X}\right)^2} \tag{3}$$

Both indicators can reflect the error size of the predicted value, but the root mean square error is easily affected by outliers, and R-squared may be affected by the complexity of the model. Hence, the present study assesses the predictive outcomes utilizing these two metrics. The R-squared value spans from 0 to 1, with values nearer to 1 indicating superior prediction accuracy. As for the root mean square error, it is influenced by the data's dimensionality; a lower value signifies a more favorable prediction performance.

1. Result analysis

2.4. Result presentation

Table 1. Comparison of ARIMA model and CNN-LSTM model test set

	ARIMA Model	CNN-LSTM Model
R-Squared	0.993	0.930
RMSE	0.0840	0.0798

From Table 1, in the prediction of Shanghai Stock Exchange 50 by CNN-LSTM, its RMSE is 0.0798, which is less than 0.0840 in the ARIMA model. As can be seen from Table 2, the R-square of the ARIMA model is 0.993, which is close to 1, and the overall interpretability is high, but the R-square of CNN-LSTM is 0.930, which is lower than that of the ARIMA model.

Table 2. Comparison of neural networks

Model/Sets	CNN		LSTM		CNN-LSTM	
Metrics	Training	Test	Training	Test	Training	Test
R-Squared	0.9781	0.9330	0.9959	0.9958	0.9958	0.9303
RMSE	0.0782	0.1217	0.0626	0.0614	0.0532	0.0798

Table 2 shows the RMSE and R-squared of the training and test sets of the hybrid neural network and the single neural network. The CNN-LSTM model exhibits the lowest RMSE with 0.0532 on the training dataset. On the training and test sets of SSE 50, the RMSE of CNN is 0.0782 and 0.1217. The error of the test set is higher than that of the training set, indicating that the model may have a certain degree of overfitting. The RMSE of the LSTM training set is 0.0626, and the RMSE of the test set is 0.0614. The errors of the training and test sets are very close, indicating that the model has good generalization ability and can maintain relatively consistent performance on training data and test data. The hybrid model demonstrated an RMSE of 0.0532 on the training set and 0.0798 on the test set, outperforming both standalone CNN and LSTM models. Notably, it achieved the smallest error on the training set, suggesting superior capability in capturing complex features. However, the error on the test set is still slightly higher than that of LSTM, and further optimization may be required.

On the training set of SSE 50, the R-squared of LSTM is 0.9959, and the R-squared of CNN-LSTM is 0.9958. Both are close to 1, which indicates that the fitting effect of these two models is good. The R-square of CNN is 0.9781, which is slightly lower than the other two models, indicating that the fitting effect of CNN is slightly worse, and there may be some overfitting. The R-square of the LSTM test set is 0.9958, which is very close to the training set, indicating that it has good generalization ability and high prediction accuracy. The R-squares of CNN and CNN-LSTM are 0.9330 and 0.9303 respectively. The R-square of this hybrid model is low, which may be due to problems such as information attenuation and gradient disappearance.

In summary, the CNN-LSTM hybrid model exhibits the minimum error on the training set; however, its performance on the test set is less stable compared to the LSTM model used independently. Further adjustments to the parameters or additional regularization may be necessary to prevent overfitting.

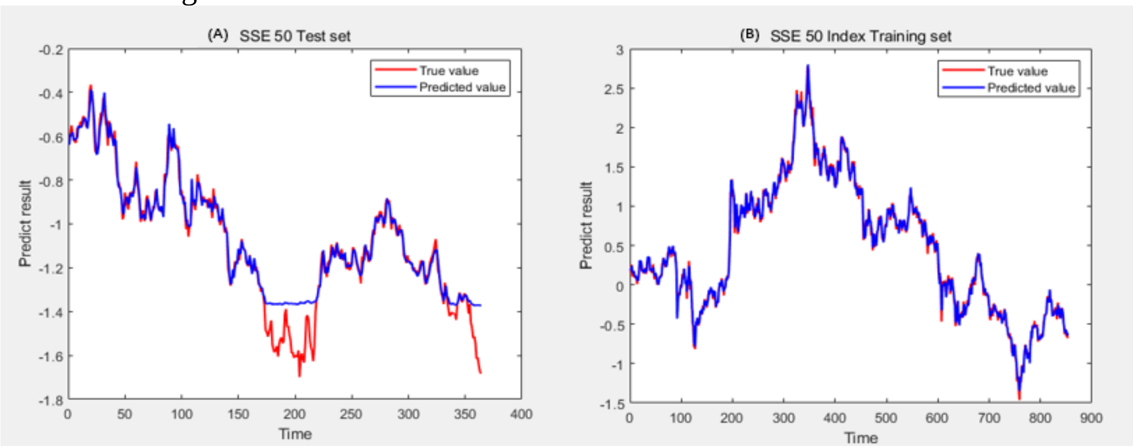

Fig. 1 CNN-LSTM test set and training set

Fig. 1(A) is the CNN-LSTM training set, most of the data fit well, but a few extreme values have a poor fit. Fig.1(B) is the CNN-LSTM test set, the overall prediction data is good, but the accuracy is extremely low near samples 160 to 220, and the prediction accuracy decreases over time.

3. Conclusion

This paper uses the ARIMA model and a single neural network to compare with CNN-LSTM. The daily data of the five indices of the Shanghai Stock Exchange 50 from September 10, 2019, to September 10, 2024, are selected to predict the stock closing price in time series. Through research, it is found that compared with the ARIMA model, CNN-LSTM has higher accuracy, but lower interpretability. Compared with a single neural network, the fitting effect of the training set of CNN-LSTM is better than the other three models, but its test set accuracy is not as good as LSTM. The above results that are inconsistent with expectations may be caused by the low robustness of the CNN model, the unclear features used, and the inappropriate number of convolution kernels or the number of LSTM nodes.

This experiment has limitations, including the difficulty in interpreting the deep learning model, the inadequate representation of the stock market through daily data, and the subpar prediction performance of certain data in the test set. In the future, the prediction accuracy can be increased by combining with the ATTENTION mechanism or improving the number of neural network layers and the size of convolution kernels.

References

[1] Badea L., Ionescu V., Guzun A.-A. What is the causal relationship between STOXX Europe 600 sectors? But between large firms and small firms? Economic Computation and Economic Cybernetics Studies and Research, 2019, 53(3): 5–20.

[2] Sousa Junior W. T. D., Montevechi J. A. B., Miranda R. D. C., Rocha F., Vilela F. F. Economic lot-size using machine learning, parallelism, metaheuristic, and simulation. International Journal of Simulation Modelling, 2019, 18(2): 205–216.

[3] Harvey C. R., Liu Y. A census of the factor zoo. Available at SSRN 3341728, 2019.

[4] Shorten C., Khoshgoftaar T. M., Furht B. Deep Learning applications for COVID-19. Journal of Big Data, 2021, 8(1): 1-54.

[5] Durairaj D. M., Mohan B. K. A convolutional neural network based approach to financial time series prediction. Neural Computing and Applications, 2022, 34(16): 13319-13337.

[6] Wang J., Hong S., Dong Y., Li Z., Hu J. Predicting stock market trends using LSTM networks: overcoming RNN limitations for improved financial forecasting. Journal of Computer Science and Software Applications, 2024, 4(3): 1-7.

[7] Xu J, Zhu Y, Xing C. A review of research on the application of machine learning in financial asset pricing. Computer Science, 2022, (06): 276-286.

[8] Zhang X. Analyzing Financial Market Trends in Cryptocurrency and Stock Prices Using CNN-LSTM Models, 2024.

[9] Livieris I. E., Pintelas E., Pintelas P. A CNN–LSTM model for gold price time-series forecasting. Neural Computing and Applications, 2020, 32: 17351-17360.

[10] Zhan X., Ling Z., Xu Z., Guo L., Zhuang S. Driving Efficiency and Risk Management in Finance through AI and RPA. Unique Endeavor in Business & Social Sciences, 2024, 3(1): 189-197.