Journal of Trends in Financial and Economics
Print ISSN: 3007-6951
Online ISSN: 3007-696X
DOI: https://doi.org/10.61784/jtfe3063

FORECASTING STOCK PRICES WITH DEEP LEARNING MODELS: A COMPARISON OF LONG-SHORT TERM MEMORY (LSTM), GATED RECURRENT UNIT (GRU), ATTENTION MECHANISM, AND TRANSFORMER MODE

ZeTong Li[1], JiuRu Lyu[2], ZiHan Wang[3], Liu Yang[3*]
[1]*Department of Electronic Science and Technology, Xi'an Jiaotong-Liverloop University, Suzhou 215000, Jiangsu, China.*
[2]*Department of Mathematics, Emory College of Art and Science Emory University, Atlanta, United States.*
[3]*School of Mathematics and Physics, Xi'an Jiaotong-Liverloop University, Suzhou 215000, Jiangsu, China.*
Corresponding Author: Liu Yang, Email: liu.y73612@gmail.com

Abstract: With the expansion of the stock market, more and more people have started to use deep learning models to predict the stock market and facilitate their trading decisions. This paper compares four mainstream deep learning models for stock price prediction: Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Attention Mechanism, and Transformer Model. Using MSE and RMSE as the evaluation metrics, we found LSTM performs the best in stock price prediction of the four companies of selection: Boeing Co, General Electric Co, Coca-Cola Co, and Johnson & Johnson. With a deeper analysis of the result, we found several limitations of LSTM, such as inconsistency of accuracy when forecasting the stock price of different firms. Hence, we suggested corresponding ways of improvement: adding more training data, introducing external factors, and integrating LSTM with other models.
Keywords: Deep learning; LSTM; GRU; Attention; Transformer; Stock; Forecasting; BA; GE; KO; JNJ

1 INTRODUCTION

With the stock market's expansion, more people are involved in stock trading. In 2019, about 600 million people worldwide bought stocks, and global stock transactions reached 60.36 trillion US dollars. Whether the stock market can be predicted has attracted more and more attention because effective prediction of the volatility of stock prices can not only strengthen financial risk management but also increase investors' enthusiasm in decision-making. In recent years, artificial intelligence, such as deep learning technology, has been combined with the financial industry to build some models to predict and analyze the stock market's volatility, improving the accuracy of stock price volatility prediction. However, because the financial market is a nonlinear and ever-changing complex dynamic system, we still can not accurately predict the changes in the stock market.

In the past, economists have been devoted to explaining economic phenomena, hoping to find the laws of economic operation existing in economic phenomena. Therefore, the previous methods of predicting stock prices, which usually need professional financial knowledge and only an understanding of uncomplicated time sequences in finance, need help to make perfect predictions [1]. Instead, deep learning technology has apparent benefits in dealing with complex and ever-changing problems, so more researchers intend to use it to make predictions. Recently, some models such as LSTM, Attention, and Transformer have been designed to predict stock prices and made some achievements. In 2017, Nelson et al. first used the LSTM model to make predictions [2]. However, following their work, in 2019, Li et al. found that LSTM could not obtain long-term dependence in long-term time series because it is limited by distinguishable position [3]. In 2019, Qiu et al. used wavelet transform to process stock data and LSTM neural network based on attention to forecasting the opening price of stocks and achieved good results. In addition, some researchers have also improved the LSTM model [4]. Li et al. 2018 proposed a multi-input LSTM element that can differentiate the mainstream factors from the auxiliary ones and perform better than the traditional LSTM model [3]. Furthermore, in 2017, Vaswani et al. created a sequence-to-sequence model called 'transformer' that adopts a multi-head self-attention mechanism to improve its ability to learn long-term dependence [5]. Ding et al. 2020 improved the ability of the original Transformer model to seize the short-term, long-term, and hierarchical dependence of financial time sequence [1].

With the continuous improvement of basic models such as LSTM, more and more advanced models have been developed and applied to stock price forecasting. However, each model has advantages and disadvantages, so finding the most suitable model is essential. In this paper, we used LSTM, GRU, Attention, and Transformer models to predict the stocks of General Electric Company, Johnson & Johnson, Coca-Cola, and Boeing. We calculated the values of the mean square error (MSE) and root mean square error (RMSE) to find out which model is the most accurate.

The rest of our paper is organized as follows: In Section II, we collect the literature related to this field. Section III walks through the related theories behind each model. Section IV introduces how to collect data, conduct experiments and analyze the experimental results. In section V, We draw our experimental conclusions and our ideas for future research fields.

2 LITERATURE REVIEW

Long Short-Term Memory (LSTM), a recursive network structure with an appropriate gradient-based learning algorithm, was first proposed by Hochreiter and Schmidhuber [6]. The advantage of LSTM is that it can handle noises, distributed representations, and continuous values [6]. Nevertheless, it also has limitations, such as the difficulty of solving problems like the strongly delayed XOR problem [3].

Several alternative models were proposed to improve the limitations of LSTM. For example, Li et al. proposed a new MI-LSTM model that enabled the mainstream to determine the use of other factors and to use a dual-stage attention mechanism for hidden states with different memory cell inputs and different time steps to improve accuracy [3].

Continuing to improve the performance of deep learning models to predict the stock market, Gupta proposed a new data expansion method in the GRU-based StockNet model, which consists of an injection module to prohibit overfitting and a survey module for stock index forecasting [7]. Compared with other models, this model has significantly lower RMSE, MAE, and MAPE [7]. Meanwhile, the GRU-based in-network data augmentation method is the unique feature of this study [7].

Apart from LSTM and GRU, the Transformer model is also a mainstream model for stock price prediction. Wang et al. proved that the Transformer model is better than traditional deep learning models in forecasting accuracy and net worth analysis [8]. Because the Transformer has a more vital ability to collect critical features and gets better prediction performance, Wang et al. inferred that financial time series prediction is a promising application field of transformer architecture [8]. In practice, by predicting transformers, investors can obtain higher excess returns [8].

To improve on the Transformer model, Ding et al. proposed some improvements to the Transformer model, including Multi-Scale Gaussian Prior, Orthogonal Regularization, and Trading Gap Splitter [1]. Their proposed Transformer-based method is superior to several advanced baselines in two fundamental trading markets compared with models such as CNN, LSTM, and ALSTM [1].

Inspired by the attention mechanism in biological phenomena, studies also reveal that attention mechanisms can be successfully integrated with other deep learning models. For instance, Zhang and Zhang proposed to optimize the LSTM model using the attention mechanism to improve its accuracy in predicting stocks [9]. The three models were also evaluated using K-fold cross-validation, and the LSTM-Attention model was more accurate and effective than the LSTM and Transformer models [9]. However, only the factor considered in this paper is time: if other factors are considered when training the model, the accuracy might be higher [9].

It is also common to see an integration of GRU with the attention model. Take Lee's work in 2022 as an example. Lee proposed a GRU-Attention deep neural network as a strategy reference for stock trading, and this study showed a significant improvement in prediction accuracy compared to other deep networks [10].

Despite using LSTM, GRU, Attention, and Transformer models, several other variations of LSTm models exist. For example, Qiu et al. proposed to predict stock price by using the WLSTM+Attention model [4]. The data is firstly processed by a wavelet transformer to make it more precise [4]. The prediction results were evaluated using S&P 500, DJIA, and HSI datasets. It was found that the WLSTM+Attention model outperformed several other models [4]. Another work by Kumar et al. proposed a method of forecasting the closing price of the stock market by using LSTM-TLBO [11]. Compared with the traditional LSTM model, TLBO focuses on execution speed, error frequency, and accuracy of results [11]. Research showed that TLBO outperforms other methods in optimizing stock price forecasts [11]. For large-scale processing of high-dimensional problems, TLBO is more effective in calculation [11]. Finally, Rajanand et al. proposed a DWCNN-SLSTM model, and they checked the performance on several baseline data sets by simply switching models while keeping all other network and training parameter constants [12]. As a result, they found that the proposed model is superior to the Transformer model in data sets in all performance indicators [12].

Although various deep learning models are available, we are still curious about which model can yield the most accurate predictions of different firms' stock prices. To achieve this goal, we will use the original LSTM, GRU, Attention Mechanism, and Transformer Models to predict the stock prices of different firms.

3 RELATED THEORIES

3.1 Long Short-Term Memory (LSTM)

The original LSTM was created in order to solve the problem of "long-term dependencies" and was proposed by Hochreiter and Schmidhuber in 1997 [6], which improves the memory capacity of standard circulating cells by bringing a "gate" into the cell. Then, the forget gate was introduced by Gers et al. in 2000 [13]. The following Figure 1 presents the inner connections of an LSTM with forget gates.

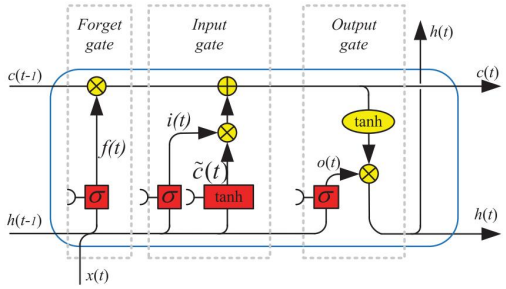

Figure 1 Inner Connections of an LSTM Cell [14]

Mathematically, we can present the inner structure of an LSTM unit with the following expressions:

$$f_t = \sigma(W_{fh}h_{t-1} + W_{fx}x_t + b_f) \quad (1)$$

$$i_t = \sigma(W_{ih}h_{t-1} + W_{ix}x_t + b_i) \quad (2)$$

$$\tilde{c}_t = \tanh\left(W_{\tilde{c}h}h_{t-1} + W_{\tilde{c}x}x_t + b_{\tilde{c}}\right) \quad (3)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \quad (4)$$

$$o_t = \sigma(W_{oh}h_{t-1} + W_{ox}x_t + b_o) \quad (5)$$

$$h_t = o_t \bullet \tanh(c_t) \quad (6)$$

In Eq. (1)-(6), x_t, h_t, and c_t denote the input, the recurrent information, and the cell state, respectively. W_f, W_c, W_o, and W_i are the weights of forget gate, input gate, cell state, and output gate, and b is the bias. Further, f_t, i_t, and o_t are the activation functions used for output. The operator "·" is the pointwise multiplication of two vectors. When the value of a forget gate (f_t) is 1, it keeps the information. Alternatively, if the value of f_t is 0, it will delete the information.

3.2 Gated Recurrent Unit (GRU)

Although the LSTM cell is better than other standard recurrent cells, the additional parameters add computational burden. Hence, Cho et al. introduced the gated recurrent unit (GRU) in 2014 [15]. The Figure 2 below shows the details of the architecture and connections of a GRU cell:

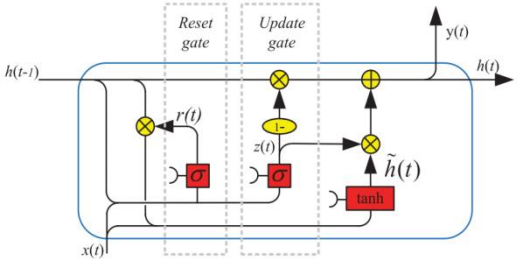

Figure 2 Architecture of a GRU Cell [14]

The following mathematical expressions are used to construct the GRU cells:

$$r_t = \sigma(W_{rh}h_{t-1} + W_{rx}x_t + b_r) \quad (7)$$

$$z_t = \sigma(W_{zh}h_{t-1} + W_{zx}x_t + b_z) \quad (8)$$

$$\tilde{h}_t = \tanh\left(W_{\tilde{h}h}(r_t \cdot h_{t-1}) + W_{\tilde{h}x}x_t + b_z\right) \quad (9)$$

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t \quad (10)$$

In this model, Cho et al. (2014) integrate the forget gate and the input gate of the LSTM cell as an update gate to reduce the data needed to be computed [15]. However, since one gate is reduced in GRU, individual GRU cells are less potent than the original LSTM cells.

3.3 Attention Mechanism

Biological phenomena are essential in inspiring people to develop different powerful algorithms for deep learning models, and the attention mechanism is no exception. It is inspired by the study of human vision, which highlights the allocation of enough attention to information that is more important than others. Integrating the idea in stock price prediction, the attention mechanism is mainly used to predict stock prices by extracting news information. The attention mechanism was first implemented by Dzmitry as a soft research structure for French-English machine translation tasks [16]. With the expansion of the stock market and the demand for predicting stock prices, a recurrent neural network based on an attention mechanism is proposed to train financial news to predict stock prices [17]. The following Figure 3 represents the general structure of the attention mechanism:

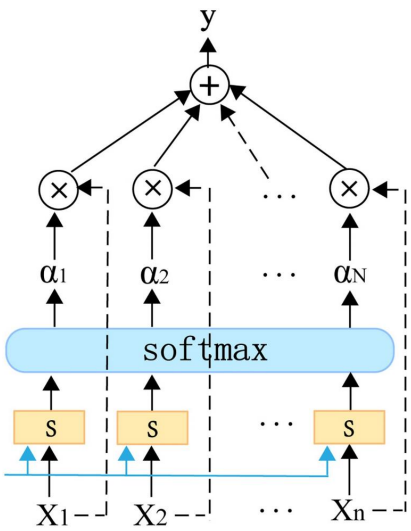

Figure 3 Architecture of the Attention Mechanism [4]

Dzmitry's study reveals the general mathematical expressions used to build an attention mechanism [16]:

$$e_{ij} = a(Q_i, K_j) \tag{11}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{12}$$

$$\text{Attention}(Q_i, K, V) = \sum_{j=1}^{T_x} \alpha_{ij} * V_j \tag{13}$$

In the equations above, Q_i is the query value corresponding to the i^{th} output element in the target. K denotes the key of all elements in the source, and, more specifically, K_j is the key of the j^{th} constituent element in the source. Moreover, V represents the value of all elements in the source, and similarly, V_j is the value of the j^{th} constituent element in the source. Lastly, T_x is the length of the source, and α is the calculation function of the correlation between Q and K_j.

3.4 Transformer Model

The Transformer model is a new generation of network architecture after Convolutional Neural Network (CNN) model proposed by Google [5]. It was initially used for natural language processing (NLP). However, due to its exact performance in downstream tasks, it is now widely used in computer vision to do tasks such as image classification, object detection, and image segmentation. The Transformer is developed based on the attention mechanism and thus is simpler and more violent than RNN. To be more specific, RNN obtains global information by recursion, whereas the transformer model based on the attention mechanism can obtain global information in only one step. The illustration below gives the architecture of a transformer model (Figure 4).

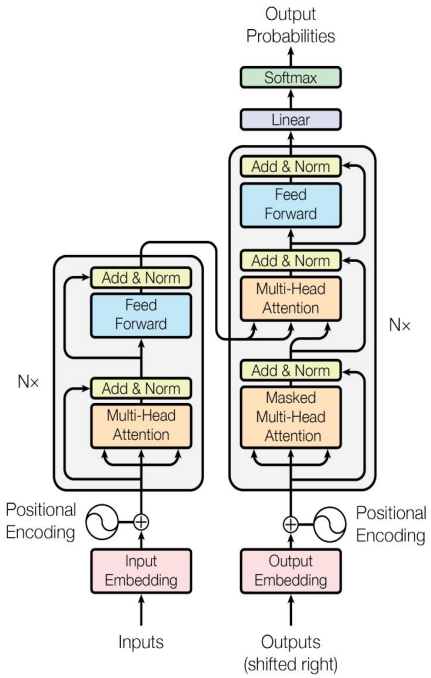

Figure 4 Architecture of the Transformer Model [5]

The following mathematical expressions give the essential idea of implementing a transformer model.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{14}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \cdots, \text{head}_h)W^O,$$
$$\text{where head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{15}$$

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2, \tag{16}$$

$$\text{PE}_{\text{pos},2i} = \sin\left(\frac{\text{pos}}{10000^{2i/d\text{model}}}\right) \tag{17}$$

$$\text{PE}_{\text{pos},2i+1} = \cos\left(\frac{\text{pos}}{10000^{2i/d\text{model}}}\right) \tag{18}$$

In Eq. (14)-(18), Q is the query vector, K denotes the key vector, V is the value vector, and QK^T is a dot product operation that calculates the weight of attention for Q on V. The purpose of scaling the result by the square root of d_k is to avoid significantly large valuesin computation. Further, W^Q, W^K, and W^V are the three matrices computed during training. To introduce nonlinearity (ReLU activation function), FFN was added to increase the model's performance. Moreover, pos represents the position of the word, and *d*model is the dimension of the position vector, which equals to the dimension of the word encoding. Lastly, $i \in [0, d\text{model})$ represents the i^{th} dimension of the position vector. The formula above gives us the *d*model vector at its corresponding pos position.

4 EXPERIMENTAL ANALYSIS

4.1 Data Collection

To compare different neural network models for stock price prediction, we collected stock prices of four firms from Yahoo Finance. Firms of selection include Boeing Co (NYSE: BA), General Electric Co (NYSE: GE), Coca-Cola Co (NYSE: KO), and Johnson & Johnson (NYSE: JNJ). The data collected includes the stock information of those enterprises on those trading days from Jan 2nd, 1962 to Nov 11th, 2017. The following table gives an example of the data collected:

Table 1 Sample Data For GE Stock Information Collected

Date	Close
1962-01-02	4.675709
1962-01-03	4.628796
1962-01-04	4.574063
1962-01-05	4.456780
1962-01-08	4.448961

In Table 1, Close refers to the closing price or the stock's final price on the corresponding trading day. We use the value of close to represent the price of all four stocks.

4.2 Data Processing

To better understand the stock price and for the sake of easy computation, we standardized the stock prices to a scale from –1 to 1. The following Figure 5 shows an example of normalized data of GE. Similar procedures are applied to the other three firms.

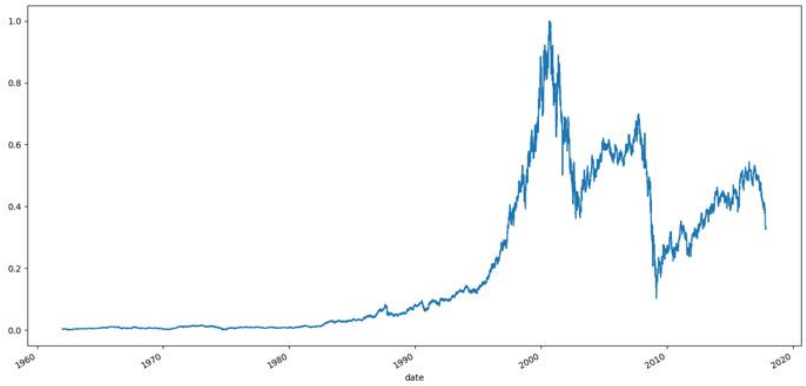

Figure 5 The General Trend of Normalzied GE Stock Price

Then, we move a lag window on the data set and classify the data into training sets, validation sets,s and testing sets. We got 9830 training data, 2810 validating data, and 1404 testing data in each data set.

4.3 Evaluation Metrics

Following the research done by Li et al., our study uses Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) to evaluate and compare the models [3]. The evaluation metrics can be calculated using the following formula:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y)^2 \tag{19}$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{20}$$

where N is the total number of samples, y_i hat is the value predicted by the model, and y_i is the expected value. The model is better and can yield more reliable outcomes with lower MSE and RMSE.

4.4 Discussion

In our study, we set the epoch of all the models to 100, and Table 2-5 gives the evaluation metrics of all the models with different stock prices.

Table 2 Evaluation Metrics for Predicting BA Prices

Model	MSE	RMSE
LSTM	0.0007	0.0266
GRU	0.0039	0.0625
Attention	0.0039	0.0625
Transformer	0.0020	0.0447

Table 3 Evaluation Metrics for Predicting GE Prices

Model	MSE	RMSE
LSTM	0.000051	0.0071
GRU	0.0009	0.0307
Attention	0.0002	0.0151
Transformer	0.0006	0.0251

Table 4 Evaluation Metrics for Predicting KO Prices

Model	MSE	RMSE
LSTM	0.0002	0. 0150
GRU	0.0003	0.0167
Attention	0.0007	0.0263

Transformer	0.0007	0.0263

Table 5 Evaluation Metrics for Predicting JNJ Prices

Model	MSE	RMSE
LSTM	0.0001	0.0139
GRU	0.0469	0.2165
Attention	0.0469	0.2165
Transformer	0.0015	0.0390

All the trials agree that LSTM gives the most accurate predictions for stock prices because the MSE and RMSE of LSTM are the lowest among the four models in each trial. We are indecisive about which model is the second most accurate in predicting stock prices because the situation varies. As for predicting BA and JNJ prices, the transformer model is the second most accurate. However, when predicting GE prices, the attention mechanism becomes the second most accurate, whereas the GRU model gives the second most reliable predictions in the case of KO price prediction.

To further our understanding of price prediction using LSTM, we plot the predicted stock price of LSTM with the actual value on the same graph, respectively, with different firms. Figure 6-9 shows the outcomes of LSTM according to the four firms of selection:

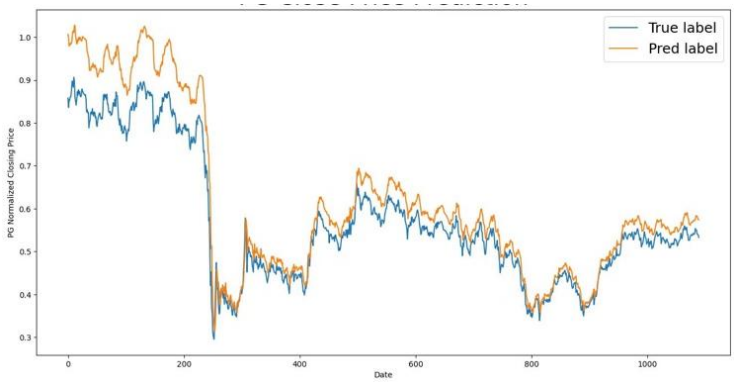

Figure 6 LSTM Prediction of BA Stock Price

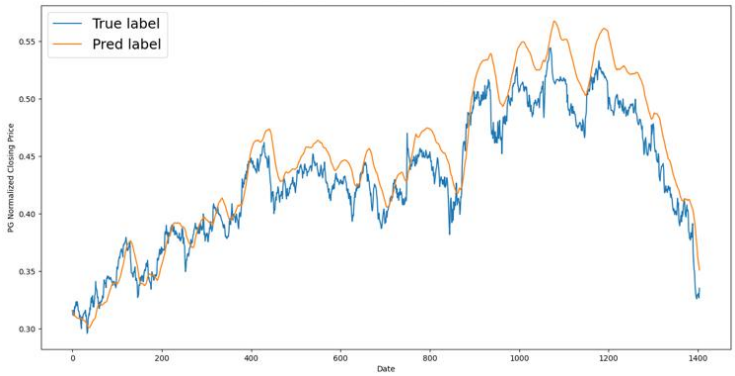

Figure 7 LSTM Prediction of GE Stock Price

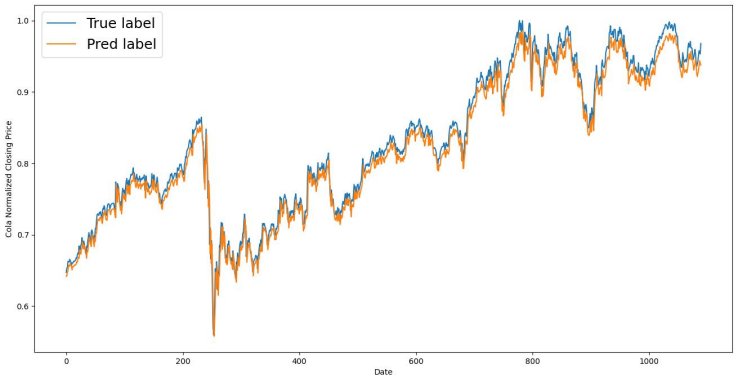

Figure 8 LSTM Prediction of KO Stock Price

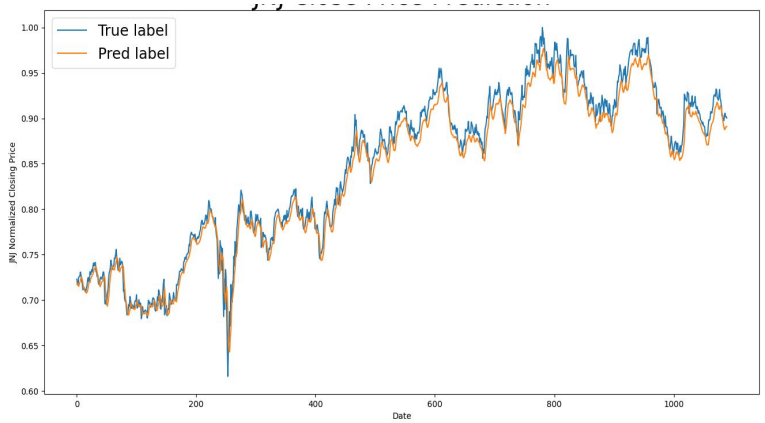

Figure 9 LSTM Prediction of JNJ Stock Price

All the plots suggest that LSTM is suitable for predicting the general trends of the stock price and can signal sudden but significant increases or decreases. In other words, when the stock price increases, the model returns an increasing trend as predicted. Moreover, when the actual data tends to decrease, the model also agrees with the tendency of the real data.
However, when predicting the BA price, the model shows relatively poor fitness at the beginning of the prediction, but as time goes on, the prediction becomes better and better. As for the GE price prediction, the model can only predict the general trend of the price but needs more precision.
To improve the consistency of accuracy and precision of prediction, we could consider adding more training data for the model to predict the stock price with higher accuracy and precision even at the very beginning of prediction. Moreover, we can integrate other factors into our model, such as stock news or fluctuations of other stock prices, so that the model can better understand the stock market and make better predictions based on that additional information. Lastly, we could integrate different models, and thus not only can the model yield more accurate results, but also we could save time and the amount of original data needed.

5 CONCLUSION

This paper explored the history of using deep learning models to predict the stock market. Then we compared the accuracy of stock predictions of four mainstream deep learning models: LSTM, GRU, Attention, and Transformer. We used MSE and RMSE as our evaluation metrics and found that LSTM provides the most accurate prediction. However, we also found that the model lacks consistency in predicting the stock prices of different firms. To improve, we suggest adding more training data, introducing additional factors, and integrating different models so that the model can understand the stock market better and yield more accurate and precise predictions.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

ACKNOWLEDGMENT

L. L. W. Y. sincerely thanks the anonymous author for his help and dedication.

REFERENCES

[1] Q Ding, S Wu, H Sun, et al. Hierarchical Multi-Scale Gaussian Transformer for Stock Movement Prediction. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, 2020: 4640–4646. DOI: 10.24963/ijcai.2020/640.
[2] D Nelson, A Pereira, R de Oliveira. Stock market's price movement prediction with LSTM neural networks. 2017: 1419–1426. DOI: 10.1109/IJCNN.2017.7966019.
[3] H Li, Y Shen, Y Zhu. Stock price prediction using attention-based multi-input LSTM. Journal of Financial Data Science. 2022, 4(3): 45-60. DOI: 10.1016/j.jfds.2022.05.002.
[4] J Qiu, B Wang, C Zhou. Forecasting stock prices with long-short term memory neural network based on attention mechanism. PLoS One, 2020, 15(1): e0227222. DOI: 10.1371/journal.pone.0227222.
[5] A Vaswani, N Shazeer, N Parmar, et al. Attention Is All You Need. 2017. http://arxiv.org/abs/1706.03762.
[6] S Hochreiter, J Schmidhuber. Long Short-Term Memory. Neural Computation, 1997, 9(8): 1735–1780.
[7] U Gupta, V Bhattacharjee, P Bishnu. StockNet - GRU based Stock Index Prediction. Expert Systems with Applications, 2022, 207: 117986. DOI: 10.1016/j.eswa.2022.117986.
[8] C Wang, Y Chen, S Zhang, et al. Stock market index prediction using deep Transformer model. Expert Systems with Applications, 2022, 208: 118128. DOI: 10.1016/j.eswa.2022.118128.

[9] S Zhang, H Zhang. Prediction of Stock Closing Prices Based on Attention Mechanism. 2020 16th Dahe Fortune China Forum and Chinese High-educational Management Annual Academic Conference (DFHMC), Zhengzhou, China: IEEE, 2020. DOI: 10.1109/DFHMC52214.2020.00053.
[10] M C Lee. Research on the Feasibility of Applying GRU and Attention Mechanism Combined with Technical Indicators in Stock Trading Strategies. Applied Sciences, 2022, 12(3). DOI: 10.3390/app12031007.
[11] K Kumar M, V R C, I C Kumari P, et al. Stock Price Prediction using LSTM and TLBO. 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), 2023: 1–5. DOI: 10.1109/ICICACS57338.2023.10100074.
[12] A Rajanand, P Singh. Stock Price Prediction using Depthwise Pointwise CNN with Sequential LSTM. 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2023: 82–86. DOI: 10.1109/ICAAIC56838.2023.10140728.
[13] F A Gers, J Schmidhuber, F Cummins. Learning to forget: continual prediction with LSTM. Neural Comput, 2000, 12(10): 2451–2471. DOI: 10.1162/089976600300015015.
[14] Y Yu, X Si, C Hu, J Zhang. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," Neural Computation, 2019, 31(7): 1235–1270. DOI: 10.1162/neco_a_01199.
[15] K Cho, B van Merrienboer, C Gulcehre, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. 2014. DOI: 10.48550/arXiv.1406.1078.
[16] D Bahdanau, K Cho, Y Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv, 2016. http://arxiv.org/abs/1409.0473.
[17] H Liu. Leveraging Financial News for Stock Trend Prediction with Attention-Based Recurrent Neural Network. 2018. DOI: 10.48550/arXiv.1811.06173.