2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT) | 979-8-3503-9902-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCECT57938.2023.10140896

IBM Stock Forecast Using LSTM, GRU, Attention and Transformer Models

Sihan Fu*
College of Science and Technology
Wenzhou-Kean University
Wenzhou, Zhejiang, China
1194226@wku.edu.cn

Zining Tang
College of Science and Technology
Wenzhou-Kean University
Wenzhou, Zhejiang, China
1194240@wku.edu.cn

Jialin Li
College of Science and Technology
Wenzhou-Kean University
Wenzhou, Zhejiang, China
1195015@wku.edu.cn

***Abstract*—In the stock market, the change of stock price has always been the most concerned thing of shareholders. However, due to the uncertainty of the stock market, it is also very difficult to predict the trend of stock price. In this paper, first, we collected IBM's stock price data from January 2, 1962 to August 21, 2017, and then we calculated the middle price of the stock. Next, we present four model approach to stock prediction. The method used in our study is known as LSTM, LSTM+GRU, Attention, and Transformer. We carried out several sets of experiments to test the validity of the analysis. The test results show that the LSTM and GRU model superposition method is more effective than other methods in predicting the stock price trend of IBM.**

***Keywords*—LSTM, GRU, Attention, Transformer, Stock, IBM, stock, forecast**

I. Introduction

By the conclusion of 2020, the aggregate market valuation of all publicly traded securities across the globe witnessed a surge from US$2.5 trillion to US$93.7 trillion [1]. A stock exchange constitutes a platform where stockbrokers and traders engage in the purchase and sale of financial instruments such as stocks, bonds, and other securities. Numerous prominent corporations have their stocks listed on stock exchanges, resulting in enhanced liquidity and rendering them more appealing to a wider array of investors. Additionally, the exchange may function as a guarantor for settlements. Various other stocks can be traded through over-the-counter (OTC) markets, involving dealer-mediated transactions. To entice international investors, some large corporations may be listed on multiple exchanges across different nations [2]. Nonetheless, the complexity of stock market data renders the prediction of stock price fluctuations challenging.

International Business Machines Corporation (IBM), a multinational technology firm headquartered in the United States, primarily focuses on computer hardware, software, and middleware, while also providing hosting and consulting services in areas such as nanotechnology and mainframe computing. Boasting 19 research facilities dispersed across 12 countries, IBM holds the distinction of being the world's largest industrial research organization. The company has consistently set records for the highest number of annual U.S. patents generated by an enterprise, a feat it has achieved for 29 consecutive years [3-4].

This paper analyzes and predicts the stock of IBM,we divided the data into three groups and used four different method to analyze IBM's public stock price data over 14,000 days to make stock predictions and compared the results with known actual data to evaluate which method is better.Although the act of stock prediction is very difficult, this predictive modeling can help investors to a certain extent.

The paper will describe this experiment into literature review part and experimental analysis part.In the review of literature,We read the papers of our predecessors in the related field and determined the data model for this experiment by analyzing and comparing the advantages and disadvantages of different methods to analyze and estimate the stock of IBM Corporation.In the analysis of experimental,we documented the experimental process and results in textual and graphical form, and summarized the predictions of stock movements,selected the most appropriate method.

II. Literature Review

In recent years, the swift advancements in time series neural networks have garnered substantial attention from investors in the stock market. The ability to comprehend the dynamic nature of the stock market and predict its trends has consistently been a topic of interest for investors and investment firms. Consequently, the prediction of stock movements has emerged as a prominent and lucrative area of research. Deep learning methodologies, such as convolutional neural networks (CNN), long short-term memory (LSTM), deep neural networks (DNN), and recursive neural networks (RNN), have been extensively investigated.

A considerable body of literature has introduced four contemporary models: LSTM, LSTM+GRU, Attention, and Transformer. Zhang et al. employed long short-term memory networks (LSTM) for the prediction of stock price trends [5]. In comparison to other artificial neural networks (ANNs), LSTM is better suited for handling nonlinear, non-stationary, and intricate financial time series, given the inherent characteristics of stock price data as time series. Lu et al. proposed a stock price prediction method based on CNN-LSTM, leveraging the long and short-term memory (LSTM) capabilities of machine learning to analyze the relationships between time series data via memory functions [6].Giuliari et. al. proposed a new method of trajectory prediction using Transformer Networks [7]. Achieves a radical shift from sequential processing in LSTM to attention-only memory mechanisms in Transformer. Chen et. al. proposed the use of genetic algorithm (GA) for feature selection, and developed an optimized long and short term memory (LSTM) neural network stock prediction model [8]. Ding et al. proposed a deep recursive neural network model of multi-input multi-output correlation based on long and short term memory network [9]. Since deep neural networks are good at dealing with prediction problems with large amount of data and complex nonlinear mapping relationship, an attention-guided deep neural network stock prediction algorithm is proposed by Zhao [10]. Reza et. al. compared the comprehensive performance of GRUs and LSTMS [11]. It is suggested that Transformer is more suitable for obtaining long range features than GRU or LSTM. Wang et. al. proposes R-Transformer, which has the advantages of RNN and multi-attention mechanism [12]. The proposed model can effectively capture the local structure and global long-term dependence in the sequence without using any positional embedding.Yang et. al. proposed Complex Transformer, which takes transformer model as the backbone of sequence modeling [13]. Attention and encoder - decoder networks for complex inputs are developed.At the same time, some neural networks lack the ability to retrieve distinguishing features from chaotic signals in the stock information flow. Therefore Li et al. developed a new hybrid neural network for price prediction, the frequency decomposition induced gate cycle Unit (GRU) transformer, which can extract discriminative insights from cluttered signals [14-15]. A novel RL-based GRUs structure is designed by Xu et. al. to filter out some irrelevant news grade representations (i.e., news grade noise) and capture a large number of long-term dependencies. Lee proposed a GRU-Attention deep neural network as a strategic reference for stock trading. It can effectively predict important stock price movements [16].

III. Method

In this paper, we use four models, respectively LSTM, LSTM+GRU, Attention, and Transformer models, to analyze and predict the price trend of IBM stock, and compare the similarity of the results to the actual data to find the most suitable model.

A. Long Short-Term Memory (LSTM)

According to Graves, LSTM is a recurrent network structure based on the original neural network RNN with the addition of long-term state memory, combined with gradient-based learning algorithms, and its main application scenarios are speech recognition, image description, and chat bots, etc [17]. In general, LSTM belongs to an upgrade of RNN, which avoids the problem of long-term dependency by a special design. The LSTM interacts specially and its structure is divided into four main structures, namely the forgetting gate, the input gate, and the output gate, as well as the important cell states as shown in Fig. 1. LSTM models are good at handling prediction problems with time series data as input, so in this paper, we use LSTM models to predict the stock movement of IBM Corporation.

As shown in Fig. 1, the four yellow blocks are called neural network layers, or activation layers, and are also referred to as "gates". From left to right, the first, second, and fourth yellow blocks belong to Sigma's activation layers but the third gate belongs to the special tan h activation layer, which ultimately generates a value. The core of the LSTM model is a line called the cell state, where all the interacting information flows in a certain order, the information is filtered by various gate structures, and the final output is the filtered information. The following is a description of the LSTM model structure and formulas.

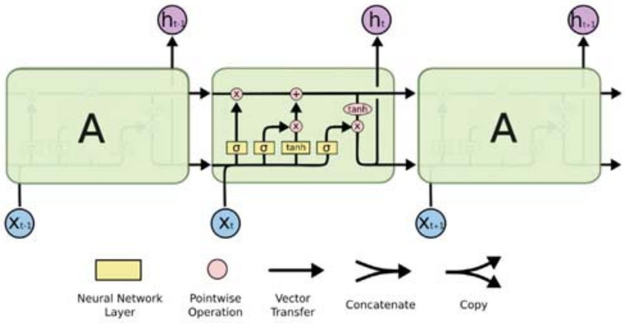

Fig. 1. LSTM structure

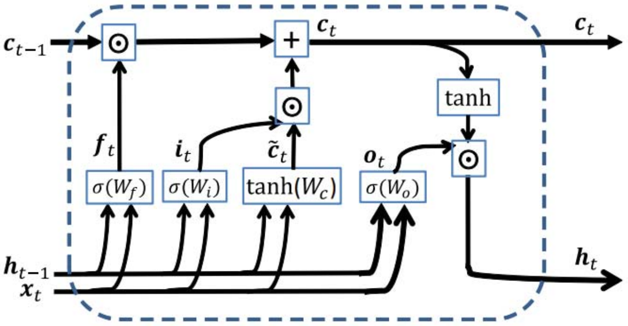

Fig. 2. [19] Long Short-Term Memory (LSTM)

According to Fig.2, the first part of the LSTM model also called the "forgetting gate", is to determine what information is discarded from the cell state. First, the value of the vector at moment h_{t-1} is read, then the value of the input at moment x_t is read, and the activation value in the middle of 0-1 is output through the σ activation layer, then it is dotted with the weights and finally output again to the cell state. In (1), σ indicates the network activation layer, W_f indicates the weight matrix, the two values inside the brackets h_{t-1} and x_t represent the input values, and finally b_f represents the bias vector.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \qquad (1)$$

The second part is called the "input gate", which determines what type of new information is present in the cell state, this part is divided into two steps as in (2) and (3), the first step is that h_{t-1} and x_t decide which values need to be updated through the σ activation layer, the second step is that the input is activated through the tan h layer and finally generates $\widetilde{C}_t$.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \qquad (2)$$

$$\widetilde{C}_t = \sigma(W_c \cdot [h_{t-1}, x_t] + b_c) \qquad (3)$$

The third part is to update the cell state, f_t which is the part to be retained, the dot product C_{t-1}, plus the part of the dot product of i_t and $\widetilde{C}_t$ in the (4).

$$C_t = f_t \otimes C_{t\text{-}1} + i_t \otimes \widetilde{C}_t \qquad (4)$$

The fourth part is called the “output gate”, and its role is to restrict the value of the final output, the process contains (5) and (6), here W_o denote the weight matrix, b_o the bias vector, and h_t is the part of the final output.

$$O_t = \sigma(W_o[h_{t-1,}x_t] + b_o) \qquad (5)$$

$$h_t = O_t \otimes \tanh(C_t) \qquad (6)$$

B. Gated Recurrent Unit (GRU)

Cho et al. proposed the GRU model [18]. The GRU model is a variant of the traditional RNN, which has similar effects to the LSTM and a simpler structure than the LSTM, so the computation is also smaller than the LSTM and it can be said that GRU combines the strengths of RNN and LSTM. Fig. 3 below shows the two main structures of the GRU, which are also two gates, the "update gate" and the "reset gate", respectively.

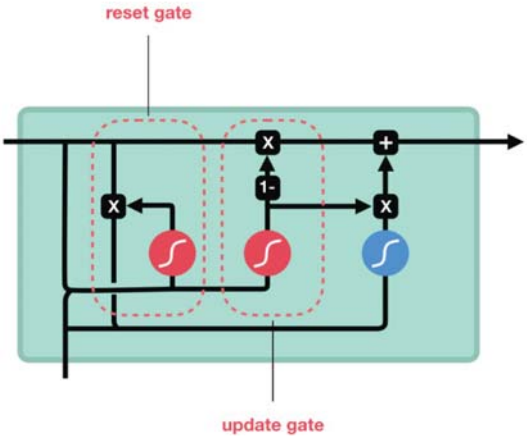

Fig. 3. [23] GRU structure

Fig. 4 is a schematic diagram of the internal structure of GRU, by comparing it with Fig. 2 LSTM structure diagram, we can see that the gating of GRU is one less than LSTM, and the parameters are also less than the LSTM model, and the equations of this model are (11)(12)(13) and (14).

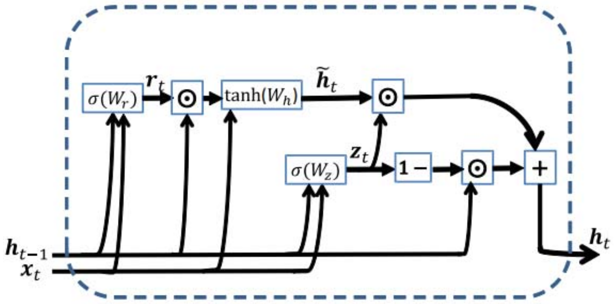

Fig. 4. [19] a diagram of the internal structure of GRU

First, (7) and (8) are used to calculate the values of the update gate z_t and reset gate r_t. These two steps represent how much of the information coming from the previous time step can be used to control.

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z) \qquad (7)$$

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r) \qquad (8)$$

Then (9) is the basic RNN calculation of the updated h_{t-1}, that is, the final activation of the tan h layer to obtain the new $\widetilde{h}_t$.

$$\widetilde{h}_t = \tanh(W_h[r_t \otimes h_{t-1}, x_t] + b_h) \qquad (9)$$

Finally, the two are summed up as shown in (10), and this process is controlled by whether the value of Z_t tends to 1 or 0 to retain the amount of information in the previously hidden layer h_{t-1} and the amount of information in the current $\widetilde{h}_t$.

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \widetilde{h}_t \qquad (10)$$

LSTM+GRU

In this paper, we used the LSTM model and GRU model superimposed to predict the stock movement trend of IBM. As mentioned in Zhou's research, the researcher prefers an RNN system with fewer gates, but still wants to keep the advantage of LSTM accuracy [19]. Because GRU has fewer parameters and stable performance, we combined LSTM and GRU to test our experiments.

C. Attention mechanism

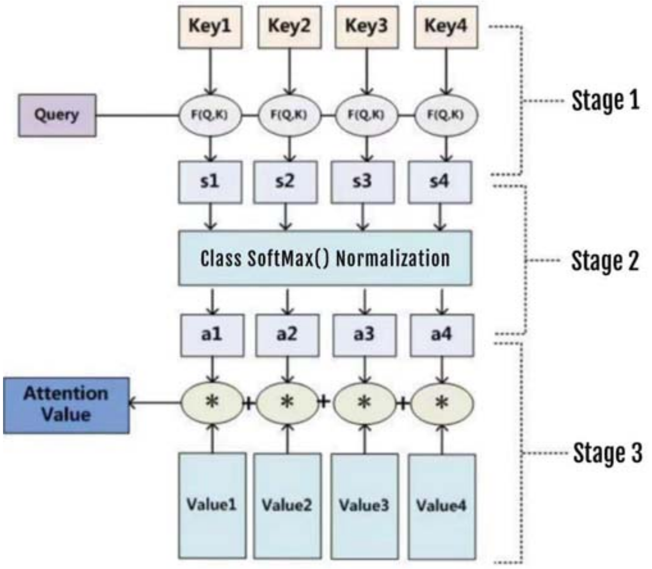

Fig. 5. Schematic diagram of the computational process of the attention mechanism

According to research, the attention mechanism was first born in the field of computer vision, and it has also been applied to natural language processing, and in addition, it has been used in areas such as LSTM models [20]. Since LSTM models cannot process information like human beings, the attention mechanism compensates for this deficiency by extracting the important information from the complex input information. The attention mechanism consists of three main parts, namely query (Q), key (K), and value (V), where K and V always appear in groups.

According to Fig. 5, the first step is to calculate the similarity between the query and key to get the weights, the second step is to normalize the weights to get the weights that can be used directly, and the third step is to sum the weights and values [18]. In other words, the process is to find the most important information to pay attention to by calculating the correlation between Q and K and getting the output value using V. The above is the principle of the general attention mechanism, and the following (11) is a calculation rule for the attention mechanism. Here d_k represents the dimension of k, so that the gradient can be kept stable during training by dividing by $\sqrt{d_k}$. QK^T represents the degree of similarity of the two messages, and then normalized by softmax, to obtain a weight matrix, and finally weighted with V.

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \qquad (11)$$

There are also many evolutionary forms of attention mechanisms, such as the self-attention mechanism and multi-headed attention mechanism, which are constantly enhancing the effect of attention and help us to better develop experiments in various fields.

D. Transformer model

Ding et al.'s research highlights that Transformer-based models are more effective in addressing financial time series forecasting challenges, as the Transformer model is adept at capturing the long-term and intricate structures inherent in financial time series data [20]. In comparison to RNN models, the Transformer model enhances its capacity to learn long-term dependencies through the utilization of a multi-head self-attention mechanism. Consequently, we also employ the Transformer model for experimentation. The overall architecture of the Transformer model can be categorized into four components: the input, output, encoder, and decoder sections, as depicted in Fig. 6.

First of all, it is clear from Fig. 6 that the output part is the two pink rectangles below the image, which contains two contents, the first one is the left part of the source text embedding layer and its position encoder, and the second one is the right part of the target text embedding layer and its position encoder. The second output part, located at the top of the image, also contains two parts, the first is the line layer, whose purpose is to get the output size, and the second is the softmax layer, through which the value with the highest probability will be extracted. Then the third part is the encoder, which is the content of the box on the left side of the figure, it is composed of N encoder layers stacked, and each encoder layer has two sub-layers connected to the structure. Finally, the fourth part is the content in the box on the right side of the figure, which is called the decoder, and it is made up of N decoder layers stacked together, and each decoder layer consists of three sub-layers connected to the structure, one more than the encoder layer sub-layers. In addition, the encoder and the decoder are connected by an output from the encoder into the decoder to form a connection. Moreover, the blue dashed box is marked as "Encoder-decoder attention", and the red solid box is "Decoder self-attention".

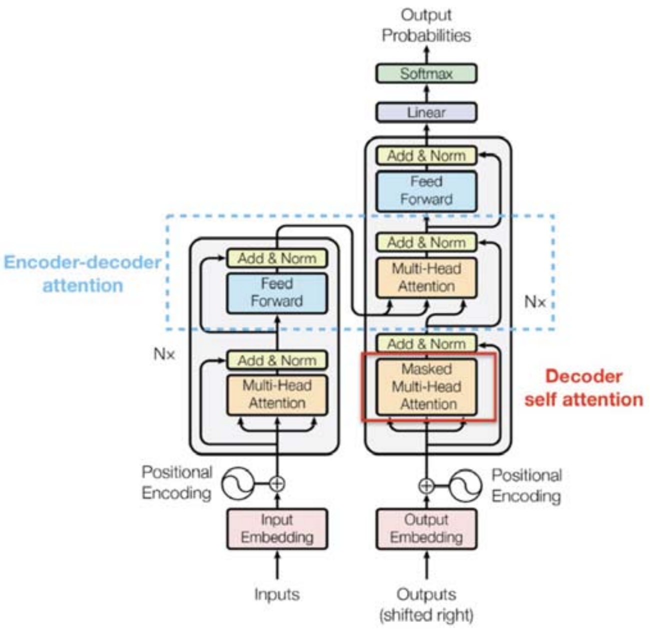

Fig. 6.[24] The Transformer-model architecture

IV. Experiments

We extracted IBM's stock price data from January 2, 1962, to August 21, 2017, including open prices, high prices, low prices, closed prices and volume data. Next, we use the sum of the high price and the price and divide by two to calculate the mid price, and the visualization results are as follows.

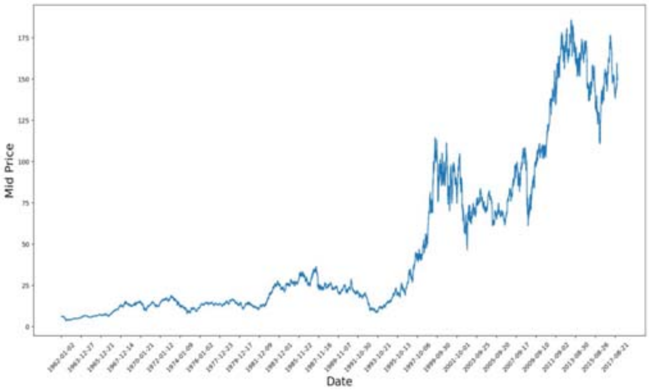

Fig. 7. Actual mid price data of IBM from January 2, 1962 to August 21, 2017

According to Fig. 7, it can be seen that from the initial data to December 1981, the mid-price was below 25. From December 1981 to October 1991, the mid-price fluctuated slightly, first increasing, but then decreasing to its original state. However, from October 1993 onwards, the mid-price rose steeply to reach a peak in October 1997. Because in the 1990s, IBM achieved the greatest transformation in the history of business, the emergence of "e-commerce" made IBM again at the forefront of global technology and business. As you can see in the chart, it reached its highest point for the first time in October 1997. And then the mid price of the stock fluctuated between 70 and 120 until the end of 1999, when it fell sharply once again, to a low of about 50 in mid-2001. In Fig.7, the mid-price maximum is near 2012, it was in an erratic upward trend from 2002 and rose sharply from September 2007, with the mid-price rising from 60 to over 175 in just five years. During this period, IBM's acquisition of PricewaterhouseCoopers Consulting and Rational Software in 2002 helped IBM attract many customers and transform itself into an all-in-one service company, and its "Smart Planet" strategy has made IBM so rich that its mid-stock price has been above 100 since 2012.

Next, we performed a normalized transformation of the training and test sets into three groups 70%, 20%, and 10%. As shown in the fig. 8, the blue line is the train target, the gray line is the valid target and the black line is the test target.

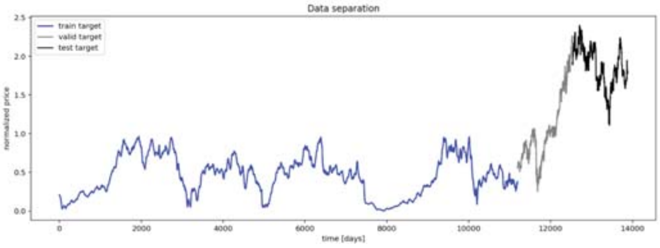

Fig. 8. Normalized price data is divided into train, valid and test.

A. Long Short-Term Memory (LSTM)

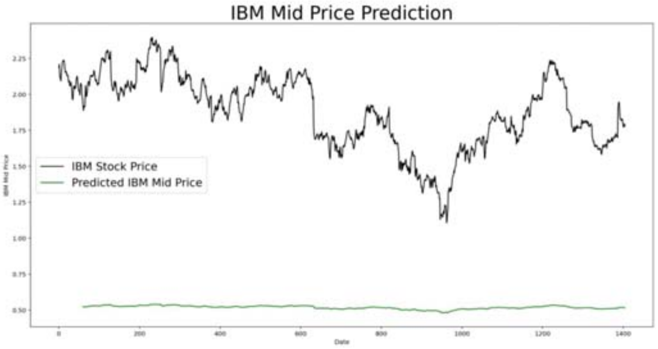

Fig. 9. IBM mid price prediction using the LSTM model

The green line in Fig. 9 is the result of 10 training sessions using the LSTM model, and the black is the original data. From the graph, it can be observed that the green line is a line that tends to be straight on 0.5, with no dramatic fluctuations. So this training result is not satisfactory, which is very poor, probably also due to the low number of training.

Training a bit further on the baseline

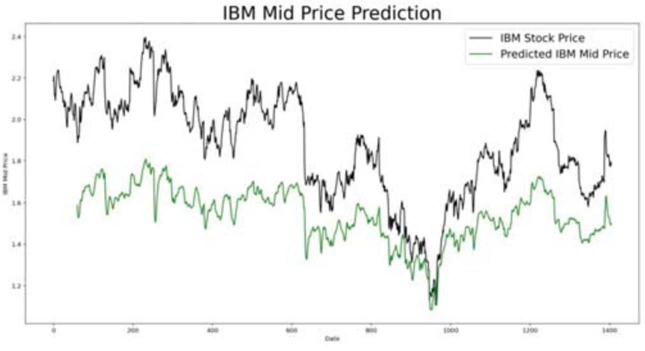

Fig. 10. Further training using LSTM model for IBM mid price prediction

We then further trained the LSTM model, this time setting the number of training runs to 100, and the results are shown in the figure above. By comparison, we can observe that the green part of the trend has great similarity with the black part, although the green part has been under the black part, the green prediction result is very close to the actual data near the position of date 900. Comparing Fig. 9 and Fig. 10, we can see that the more times the LSTM model is trained, the closer the predicted value is to the true value, and if we want to make the prediction more accurate, we can increase the training number again, but this is still more troublesome. So we need to find a better model to make it less trained and achieve higher similarity.

B. LSTM+GRU

GRU and LSTM are two variants of RNN, which can solve the problem of RNN gradient disappearance and obtain learning ability with long-term dependence [10]. The following figure is the effect after 10 times of training using the LSTM model and GRU model superimposed.

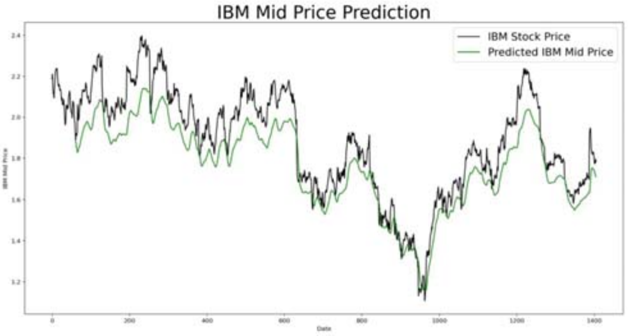

Fig. 11. IBM mid price prediction using the LSTM and GRU models

In this chart, we can see that this result is shown to be well done, the predicted IBM mid-price is much more similar to the actual IBM stock price data, with the green and black lines more closely aligned, and the green line moving closer to the black line, even though the green predicted value is still below the true value. By comparing Fig.10 and Fig.11, we found that the effect of training only 10 times after superimposing the two models is better than that of training 100 times with the LSTM model. This leads to the conclusion that the overlay model of LSTM and GRU is more accurate and suitable for predicting the stock trend.

C. Attention mechanism

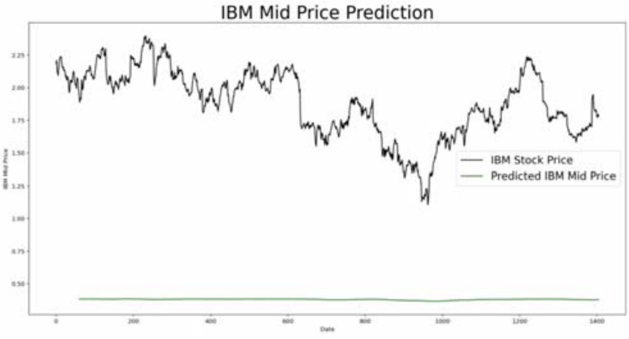

Fig. 12. IBM mid price prediction using the Attention model

We try to use the attention mechanism model to predict the IBM stock price trend, as shown in Fig. 12. This is the result of ten training sessions, although the attention mechanism model can accept ultra-long data, from the results given a super-low green predicted price can be seen, the model is still not effective after ten training sessions, so next, we try Transformer's self-attention mechanism.

D. Transformer model

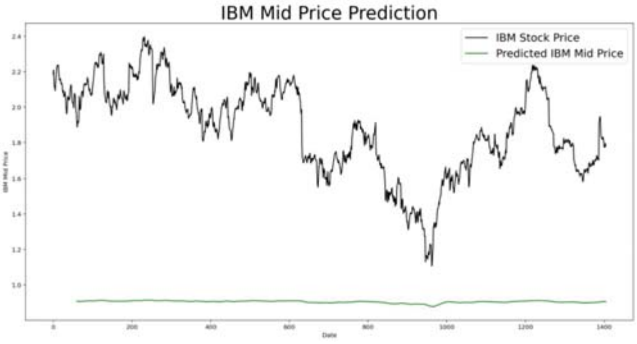

Fig. 13. IBM mid price prediction using the Transformer model

Fig. 13. is the Transformer model trained ten times, the picture shows us that there is still no difference between the results of the Attention model, only the green predicted price changes a little, but still at a very low pair of prices, roughly smooth. Although in principle the Transformer model is better than the Attention model, from the effect of training ten times, these two are not too useful for IBM stock prediction, perhaps we can improve the effect by increasing the number of times, but that is also too much effort.

By comparing the above four models, it is easy to find that the superposition model of LSTM and GRU is the most effective for predicting the price change of IBM stock.

V. Conclusion

The main goal of the current study was to determine the development tendency of IBM's stock, and the insights gained from this study may be of assistance to the investor. This investigation's finding complement those of earlier studies. Before this experiment, evidence of IBM stock tendency was purely anecdotal. A limitation of the study is that we did not try the superposition of multiple models, such as Attention+LSTM, or the superposition of three models. Moreover, fewer training sessions make these findings less generalizable. As aforementioned several questions remain to be answered. This study has identified that LSTM and GRU model superposition method is more effective than other methods in predicting the stock price trend of IBM. The present study was designed to determine the effect of changing the stock market.

Acknowledgment

F. T. L. sincerely thanks the anonymous author for his help.

References

[1] Hu Y., Tao Z., Xing D., Pan Z., Zhao J., & Chen X. (2020, August). Research on stock returns forecast of the four major banks based on ARMA and GARCH model. In Journal of Physics: Conference Series (Vol. 1616, No. 1, p. 012075). IOP Publishing.

[2] Tao Z., & Gupta G. (2022). Stock Investment Strategies and Portfolio Analysis. In Proceedings of Academia-Industry Consortium for Data Science: AICDS 2020 (pp. 397-406). Singapore: Springer Nature Singapore.

[3] Bajpai Prableen (January 29, 2021). "Top Patent Holders of 2020". nasdaq.com. Nasdaq. Archived from the original on January 30, 2021. Retrieved February 2, 2021.

[4] "2021 Top 50 US Patent Assignees". IFI CLAIMS Patent Services. January 5, 2022. Retrieved August 22, 2022.

[5] Yongjie Zhang; Gang Chu; Dehua Shen; "The Role of Investor Attention in Predicting Stock Prices: The Long Short-term Memory Networks Perspective", FINANCE RESEARCH LETTERS, 2020.

[6] Wenjie Lu; Jiazheng Li; Yifan Li; Aijun Sun; Jingyang Wang; "A CNN-LSTM-Based Model to Forecast Stock Prices", COMPLEX., 2020.

[7] Francesco Giuliari; Irtiza Hasan; Marco Cristani; Fabio Galasso;" Transformer Networks For Trajectory Forecasting", ARXIV, 2020.

[8] Shile Chen; Changjun Zhou; "Stock Prediction Based on Genetic Algorithm Feature Selection and Long Short-Term Memory Neural Network", IEEE ACCESS, 2021.

[9] Guangyu Ding; Liangxi Qin; "Study on The Prediction of Stock Price Based on The Associated Network Model of LSTM", INTERNATIONAL JOURNAL OF MACHINE LEARNING AND CYBERNETICS, 2020.

[10] Yangzi Zhao; "A Novel Stock Index Intelligent Prediction Algorithm Based on Attention-Guided Deep Neural Network", WIRELESS COMMUNICATIONS AND MOBILE COMPUTING, 2021.

[11] Reza S., Ferreira M. C., Machado J. J. M., & Tavares J. M. R. (2022). A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. Expert Systems with Applications, 202, 117275.

[12] Wang Z., Ma Y., Liu Z., & Tang J. (2019). R-transformer: Recurrent neural network enhanced transformer. arXiv preprint arXiv:1907.05572.

[13] Yang M., Ma M. Q., Li D., Tsai Y. H. H., & Salakhutdinov, R. (2020, May). Complex transformer: A framework for modeling complex-valued sequence. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4232-4236). IEEE.

[14] Li C., & Qian G. (2023). Stock Price Prediction Using a Frequency Decomposition Based GRU Transformer Neural Network. Applied Sciences, 13(1), 222.

[15] Xu H., Chai L., Luo, Z., & Li S. (2022). Stock movement prediction via gated recurrent unit network based on reinforcement learning with incorporated attention mechanisms. Neurocomputing, 467, 214-228.

[16] Lee M. C. (2022). Research on the Feasibility of Applying GRU and Attention Mechanism Combined with Technical Indicators in Stock Trading Strategies. Applied Sciences, 12(3), 1007.

[17] Graves A. (2012). Long short-term memory. Supervised sequence labeling with recurrent neural networks, 37-45.

[18] K. Cho B. van Merrienboer C. Gulcehre F. Bougares H. Schwenk D. Bahdanau, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv: 1406.1078, 2014.

[19] Zhou G. (2016, March 31). Minimal Gated Unit for Recurrent Neural Networks. arXiv.org. https://arxiv.org/abs/1603.09420

[20] Fukui H., Hirakawa T., Yamashita T., & Fujiyoshi H. (2019). Attention Branch Network: Learning of Attention Mechanism for Visual Explanation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2019.01096