Comparison and Analysis of Prediction for High Volatility Stocks using Different Machine Learning Models

Minglong Chen[1, †], Mohan Ren[2, *, †], Yujie Zhou[3, †]
1950598@tongji.edu.cn, *Corresponding author: mohan.ren@student.manchester.ac.uk,
Sgyzho50@liverpool.ac.uk

[1]School of Economics and Management Tongji University Shanghai, China, 201804

[2]Department of Computer Science The University of Manchester Manchester, UK, M13 9PL

[3]Department of Mathematical Sciences The University of Liverpool Liverpool, UK, L69 7ZX

[†]These authors contributed equally.

Abstract—As the global situation is facing many unpredictable risks, like the covid pandemic and the wars between some countries, a high level of volatility is disturbing the stock market in a world range. Besides, due to the development of machine learning techniques, more and more of these techniques have been used in predicting the price of the stock to get profits, and some of them like Long Short-Term Memory (LSTM) have also shown an extraordinary performance. However, these applications should be reevaluated because of nowadays market volatility while in the past, most experiments were built based on generic volatility. In this experiment, therefore, this study chose two most volatile and two least volatile stocks in the past two years and utilized some popular machine learning models, such as k-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Long Short-Term Memory to predict their prices. The results have shown that all of them will suffer from a roughly 4 to 5 times performance reduction in terms of the prediction ability, but when ignoring how the volatility is, LSTM will always give the best performance.

Keywords-High-volatility; Machine learning; LSTM; Stock prediction

1 INTRODUCTION

Stock is an essential part of the domestic economy, which is closely relative to the national development of the economy [1]. However, it is difficult, especially for small and medium-sized investors, to make an analysis of future stock trends to decide the investment strategy due to daily mass of data. Popular theories suggest that stock markets are essentially a random walk, so that trying to predict them will encounter great difficulties and challenges. However, there are always ambitious people who want to challenge this conundrum. Over the years, developing and testing models of future stock prices have appealed a large number of investors and researchers, in order to approach the prediction accuracy [2].

However, analyzing stock market behaviors and price changes is extremely difficult, since the stock market has naturally the following characteristics: dynamic, nonlinear, nonstationary,

FFIT 2022, October 28-30, Shenzhen, People's Republic of China

DOI 10.4108/eai.28-10-2022.2328453

nonparametric, noisy, and chaotic [3]. And changes in stock prices are influenced by multiple factors, including economy, policy, market, and investor psychology, which increase the difficulty of stock forecasting. It is supposed that the market prices of stock have reflected all information, therefore the remaining price fluctuation accounts for unpredicted random events, forming a random walk of changes in stock price.

Nowadays, technological advances promote research about stock price prediction. The main stock price prediction models are divided into several categories: traditional model, machine learning model, and deep learning model. Traditional models, such as linear regressive model and autoregressive integrated moving average model (ARIMA). The conducted study by several researchers suggested that implementing ARIMA model in stock market to forecast stock indexes was recommended [4]. With the proliferation of computer technology, machine learning has been successfully applied to predicting stock prices according to historic data [5]. Cao, Leggio and Schniederjans verified the effectiveness of forecasting stock price by comparing linear model with nonlinear Artificial Neural Network (ANN) model, where statistically significant data indicated prediction power was enhanced by the neural network [6]. Moreover, Long Short-Term Memory (LSTM), according to Raghav's research, have good results as they are able to keep track of the context-specific temporal dependencies between stock prices for a longer period of time while performing predictions [7]. Also, LSTM is widely applied to stock predicting research because it solves the defects of gradient disappearance and gradient explosion in RNN model [1].

However, after the appearance of the Covid-19 pandemic which started in 2020, the whole world is getting more and more unstable, especially the global financial market, which is sensitive and easy to be affected even if the factor is tiny. And it has been proved that the whole capital market suffered a high volatility during the recent two years. People may hope the world could be better and more stable in 2022 as many countries have introduced a series of positive policies to stimulate the economic recovery, but one thing that interrupted this is the war between Russia and Ukraine, which has deeply influenced the world, leading the stock market to a high volatility state again even more. Therefore, there is a point given that how can technical analysis like machine learning will forecast the trend of the global stock market when high volatility exists. Also, some comparison between the results from different machine models will be shown in this situation as war also presents important economic questions [8] and one explicit question of them is high volatility. Furthermore, many researchers are exploring the final results of their prediction using machine learning models on generic stocks on an average level, therefore some comparations between those high volatility stocks are necessary.

This research focuses on the comparison between the high volatility stock price prediction results from different machine learning models including linear regression, k-Nearest Neighbor (KNN), support vector machine (SVM) and Long Short-Term Memory (LSTM). The study would show a more comprehensive analysis of how each model can deal with the volatility, and will also simply compare how the range of the time will affect the prediction as sometimes validity could be diluted when the time range is giant enough.

2 METHODOLOGY

To examine how the volatility will affect different machine learning models in stock market, and compare the performance of each model, we built two common machine learning models as well as one famous time series model. They are k-Nearest Neighbour (KNN), Support Vector Machine (SVM) and Long Short-Term Memory (LSTM).

2.1 Dataset and Features

According to the average price volatility from Nasdaq in the past two years, the stocks we chose include two high-volatility stocks: NVIDA (NVDA) and Tesla (TSLA), and two low-volatility stocks: Microsoft (MSFT) and Air Products & Chemicals (APD).

The total size of the data sample is 505, where the stock prices are originated from 4 companies. The date is between 03/22/2020 to 03/22/2022 as 03/22/2020 is the day that is close to the start of the global pandemic, which led to the high volatility of the world stock market.

In the dataset downloaded from Yahoo Finance [9], we found that information included “date”, “open price”, “close price”, “the highest price”, “the lowest price” and “the adjusted closing price” daily. In this study, adjusted closing price was used because amended closing price reflected the price after corporates’ actions, including dividends, rights offerings and so on [10]. Therefore, we needed to delete the other columns and only remained date and adjusted closing price. In addition, normalizing data was significant in order to alleviate the impact of magnitude and restrict the data to a certain scale. In this report, MinMax normalization was used for more accurate data and following operations. Also, the dataset was divided into train set and test set, where the proportion was 4:1, and therefore the test set starts on 10/27/2021. A sample piece of our data is shown Table 1:

Table 1 Data structure of NVIDA stock price, from 2020 to 2022

DATE	OPEN	HIGH	LOW	ADJ CLOSE	VOLUME
2022-03-23	51.427502	54.119999	49.630001	53.061264	64387600
2022-03-24	57.250000	63.165001	57.012501	62.164680	110164400
2022-03-25	63.450001	65.455002	60.915001	61.276536	79343200
2022-03-26	63.485001	65.555000	62.500000	64.175468	69610800
2022-03-27	62.500000	65.779999	62.169998	63.050327	63882800

2.2 k-Nearest Neighbor

First, k-nearest neighbor was taken into consideration, which was a common machine learning model storing the training set and predicting the test set on the degree of similarity, such as the Euclidean Distance [11]. The most essential parameter in kNN method was k, which is automatically selected by Grid Search among several numbers (from 1-9). And according to the number of the nearest neighbors k which determined the closest k records of the training set are highly similar to unknown records, also, depending on the calculated distance between training data and test data, then a majority vote can be conducted to ensure unknown record a predicted value and give test data a class label. Finally, Root Mean Square Error (RMSE) was applied to measure the accuracy of kNN model, which is commonly utilized in machine learning to evaluate a model. The smaller the RMSE is, the smaller the deviation between the predicted value and observed value is. Overall, kNN model has its edge in simplicity, short training time and insensitivity of outliers.

2.3 Support Vector Machine

Support Vector Machine (SVM) is one of the most popular supervised learning algorithms, which is used for classification as well as regression problems. The goal of SVM is to create the best line, or the best decision boundary, that can segregate n-dimensional space into classes [12]. In this case, we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points or vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. SVM algorithm can learn to predict stock market prices and trends based on historical data and provide meaningful analysis of historical prices.

In our work, we set the kernel parameter to "Radial Basis Function (RBF) ".

2.4 Long Short-Term Memory

In the third part, we used the Long Short-Term Memory (LSTM) which is commonly used in time series problems. Long Short-Term Memory Network is an advanced Recurrent Neural Network (RNN), a sequential network, that allows information to persist [13]. Furthermore, LSTM introduces the memory cell that enables long-term dependency between time lags [14]. There are three kinds of gates in LSTM, namely input gate, forget gate and output gate.

The forget gate will help to decide whether the information from the previous timestamp should be kept. It uses input x_t as the current timestamp, the output from the previous timestamp H_{t-1}, and then outputs a value between 0 and 1 describing the cell state of the previous timestamp:

$$f_t = \sigma(x_t * U_f + H_{t-1} * W_f) \tag{1}$$

The input gate helps to quantify how to import the new information taken by the input. And it uses a different weight matrix with the forget gate:

$$i_t = \sigma(x_t * U_i + H_{t-1} * W_i) \tag{2}$$

The output gate has a similar formula as the other two, which is:

$$o_t = \sigma(x_t * U_o + H_{t-1} * W_o) \tag{3}$$

As the sigmoid function is used, the output value will also be in the interval of 0 to 1. However, o_t and the $\tanh(C_t)$ of the updated cell state can be used to calculate the current hidden state:

$$H_t = o_t * \tanh(C_t) \tag{4}$$

For the final step using LSTM, we can apply the *SoftMax* activation on the hidden state H_t to achieve the output result.

2.5 Implementation details for LSTM

Using a library called Keras is a good way to build an LSTM network. In this experiment, there are 4 layers of LSTM added, and each is followed by a dropout layer to prevent overfitting. The parameter of the dropout is 0.2, meaning that 20% of the layers will be dropped.

Additionally, we used the popular optimizer called Adam and loss function was chosen as Root Mean Square Error. Besides, we set epochs as 30 and batch size as 32 to reduce the randomness of the experiment.

After importing the test set, we plot the graph of our prediction compared with the real values and still used "Root Mean Square Error" to evaluate the performance.

3 RESULT AND DISCUSSION

Root Mean Square Error is a good manner to evaluate the effect of regression questions, so here we regard it as the indicator to evaluate our models, and we calculate the results of each experiment using this formula:

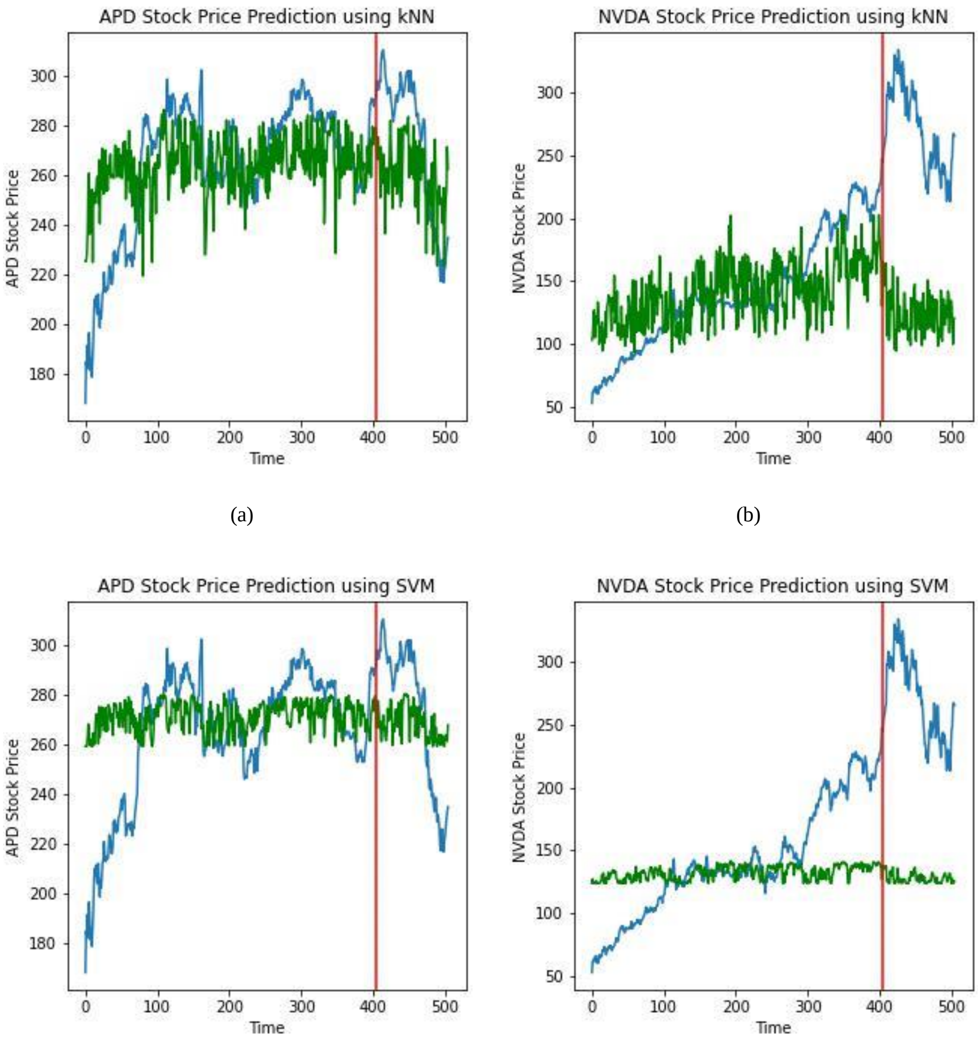
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}} \tag{5}$$

where x_i is the true value while $\hat{x}_i$ is the predicted value, and N is the size of the data sample.

In this experiment, 404 pieces of data were used for training different machine learning models and then predicted 101 pieces of data for the future which has the same dates as the testing data samples. Then the predicted output and the actual data were used to calculate the RMSE. The results for different machine learning models and different stocks (NVDA, APD, TSLA, MSFT) are shown in Table 2. However, not only the dates separated for testing were predicted, but the whole dataset which began on 2020-3-23 and ended on 2022-3-22 was predicted by the training performance to give a more intuitive observation. Among 12 graphs of stock prediction, 6 results shown in Figure 1 were chosen because they are more representative.

Table 2 RMSE of models.

Model Name	**Company Name**			
	NVDA	***TSLA***	***MSFT***	***APD***
kNN	151.551	374.489	96.044	28.221
SVM	75.765	253.833	51.769	26.053
LSTM	32.612	133.755	17.047	16.358

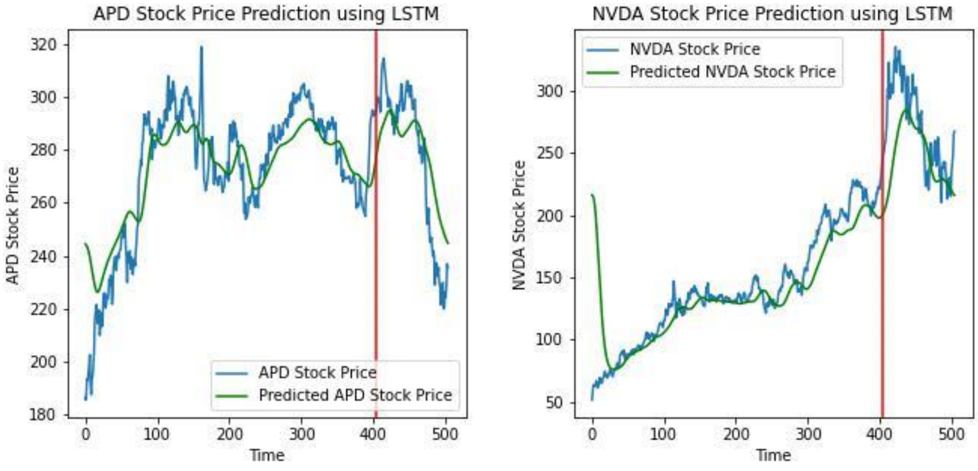

APD Stock Price Prediction using kNN
APD Stock Price
180
200
220
240
260
280
300
Time
0
100
200
300
400
500
NVDA Stock Price Prediction using kNN
NVDA Stock Price
50
100
150
200
250
300
Time
0
100
200
300
400
500
APD Stock Price Prediction using SVM
APD Stock Price
180
200
220
240
260
280
300
Time
0
100
200
300
400
500
NVDA Stock Price Prediction using SVM
NVDA Stock Price
50
100
150
200
250
300
Time
0
100
200
300
400
500

(a)

(b)

(c)

(d)



(e) (f)

Figure 1. (a) kNN for APD; (b) kNN for NVDA; (c) SVM for APD; (d) SVM for NVDA; (e) LSTM for APD; (f) LSTM for NVDA

It is clear to observe that all these three models suffered a RMSE increase because of the volatility according to the table above, although the effects are not the same for them. For k-Nearest Neighbor model, the average RMSE for low volatility stocks MSFT and APD is 62.1325, which is the highest among these three models. And the average RMSE for high volatility stocks NVDA and TSLA is 263.02, which suffers about 4 times increase, making it harder to predict the future price. Besides, analyzing figure 4 to predict the price of NVDA by kNN, we can find that the value of the predicted line does not fit the actual value a lot, but the trend has some similarities, i.e., both lines increased at the point of 500. Therefore, compared with the other two models, kNN is not a good choice.

Support Vector Machine has a better performance compared with kNN as the average RMSE value of low volatility stocks is only 38.911 while of high volatility stocks is 164.799. It is proved that SVM has a better ability to predict the price no matter what the volatility is, but as the multiple of growth of the RMSE for it is roughly equal to the increase for kNN, which are both around 4.23, we cannot identify that which model will get less effect from the volatility. Furthermore, by observing the predicted price of SVM compared with the actual price, we can neither infer the specific price value as SVM will not jump off its hyperplane once it got the margin size. It is not a wise choice to predict the actual price value using this model, but it is still a good way to find the trend as it can show which direction the price will go in the future in some ways by watching the figure we got above, i.e. in figure 5 of NVDA using SVM, the predicted line has the similar shape with the actual value between 470 and 480.

It is obvious that LSTM has the best performance in these three models. The average RMSE value for low volatility stocks is 16.70, which is a quite low value compared with the other two. However, it can be seen that the growth of the average RMSE when using high volatility stocks is huge, and the value is 83.1835, 5 times higher than the RMSE for low volatility stocks.

Therefore, we can infer that no matter how the volatility is, LSTM is a proper model to predict the price in the future, and by watching the figure 6 of using LSTM to predict the stock price of NVDA, the predicted line can well fit the true value line even in the specific value. As the multiple of growth of LSTM is just a bit higher than the other two models, it is not clear to confirm that LSTM will get more effect from the volatility than the other and more repetition should be done to avoid the probability events to reduce the difference.

4 CONCLUSION

In our work, three machine learning models are used to predict the price of 2 high volatility stocks and 2 low volatility stocks and compare the predictions, where kNN, SVM and LSTM models are applied to predict the future price of the stocks. Based on this, the impact of volatility on the model's prediction of stock prices is discussed and some improvements for the prediction in the future are mentioned. The results show that the predictions of LSTM are obviously better than the other two models, no matter what the volatility is. However, the changes in RMSE caused by the level of volatility are not clearly reflected in the three models. The multiple of growth of RMSE of the LSTM model, when changing from low volatility stocks to high volatility stocks, is just a bit higher than the other two models. In the future, more stocks will be applied to clearly confirm whether or not LSTM will get more effect from the volatility than the other two models.

REFERENCES

[1] M. Cheng, S. Gao, "Multi-scale stock prediction based on deep transfer learning", Computer Engineering and Applications, 2022.

[2] Fama. F, "Random walks in stock market prices", Financial Analysts Journal, 1965.

[3] A. Yaser, A. Amir, "Introduction to financial forecasting", Applied Intelligence, 1996.

[4] W. Dassanayake, I. Ardekani, N. Gamage, C. Jayawardena and H. Sharifzadeh, "Effectiveness of Stock Index Forecasting using ARIMA model: Evidence from New Zealand", International Conference on Advancements in Computing, 2021.

[5] Y. Huang, L. F. Capretz, D. Ho, "Machine Learning for Stock Prediction Based on Fundamental Analysis", IEEE Symposium Series on Computational Intelligence, 2021.

[6] C. Qing, B. Karyl, S. Marc, "A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market", Computers & Operations Research, 2005.

[7] N. Raghav, M. Uttamraj, R. Vishal, “Stock Price Prediction Using Long Short Term Memory”, International Research Journal of Engineering and Technology, 2018.

[8] L. Andrew, W. Justin, Z. Eric, "What Do Financial Markets Think of War In Iraq?" National Bureau of Economic Research, 2003.

[9] Yahoo, “stock mark live, quotes, business and finance”, https://uk.finance.yahoo.com/, 2022.

[10] Investopedia, “Adjusted Closing Price”, https://www.investopedia.com/terms/a/adjusted_closing_price.asp, 2020.

[11] S. Islam, M. Sikder, M. Hossain. et al. “Predicting the daily closing price of selected shares on the Dhaka Stock Exchange using machine learning techniques.” SN Bus Econ 1, 58, https://doi-org.liverpool.idm.oclc.org/10.1007/s43546-021-00065-6, 2021.

[12] Javatpoint, “Support Vector Machine Algorithm”, https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm, 2021.

[13] S. Saxena, "Introduction to Long Short Term Memory(LSTM)", Analytics Vidhya, https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm, 2021

[14] Z. Zou, Z. Qu, "Using LSTM in Stock prediction and Quantitative Trading", 2020