Predictions of Tesla Stock Price based on Linear Regression, SVM, Random Forest, LSTM and ARIMA

Renjie Fan*

Department of Shagang Iron and Steel Institute, Soochow University, Suzhou, China

*Corresponding author: 2013401006@stu.suda.edu.cn

Abstract. The stock market of a country is an important financial market. A booming stock market promotes the effective use of social capital, the prudent deployment of economic resources, and the expansion of the country's macroeconomics. Making more informed decisions as an investor is made possible by the development of trustworthy equity market models. A trading model allows market participants to select corporations that pay the highest dividend payments while lowering the risks associated with investing. However, batch processing methodologies make stock market research more challenging as a result of the strong connection between stock prices. The advent of technological achievements like universal digitization has elevated share market forecasting into a highly advanced age. Through the research and comparison of several methodologies, this article tries to discover the most accurate approach for predicting Tesla stock closing prices. Predictions are made using statistical approaches such as ARIMA and machine learning methods such as SVM , Linear Regression, Random Forests, and LSTM. Following a thorough examination of all approaches, it was discovered that the accuracy of machine learning methods in predicting stocks is higher than that of statistical methods and integrated algorithm technologies like Random Forest have excellent anti-interference and anti-overfitting characteristics, which are more suitable for evaluating high-volatility stocks like Tesla.

Keywords: Tesla; stock market; machine learning; future price prediction.

1. Introduction

The stock market is a group of stockbrokers and dealers who buy and trade stock shares. A stock exchange is where many large firms' shares are traded. This improves the stock's liquidity, making it more enticing to investors [1]. A vast number of investors put large sums of money into the stock market. However, it is risky since stock values may quickly rise or decrease [2]. Because of this, forecasting stock prices is a difficult subject on which many scholars are working.

To predict stock prices, a variety of approaches have been utilized. Because statistical methodologies are linear in nature, they perform poorly in the event of a rapid spike or decrease in stock prices. Because stock data is unpredictable, volatile, non-stationary, and dependent on a variety of technical characteristics, statistical techniques have been proven to be insufficiently accurate [3].

The open stock exchange trades stocks of publicly listed companies, whereas the private share exchange trades stocks of privately held businesses. Investments made through the mixed-ownership share trade are made in companies with ordinary shares that may only be exchanged publicly on rare occasions. Several countries, like the UK's London Stock Exchange and the US's New York Stock Exchange, have stock exchanges with mixed ownership [4-9].

Thus, all previous work has been centered on reducing some metric that leads to estimates near the true stock price. However, this does not indicate that these forecasts will be profitable.

The experiment first collects data from Kaggle. Reprocess the data to remove any null values in the time series. And transform and normalize the data to make it compatible with the input of the model. Then use Matlab to implement the Linear Regression model [10], SVM model [11, 12], Random Forest model [13, 14], LSTM model [15], and ARIMA model [16, 17]. Finally, compare the performance of different models and propose further work and project expansion.

The purpose of the experiment is to successfully realize the linear regression model, SVM model, Random Forest model, LSTM model, and ARIMA model of Tesla stock data. Finally, assess the model's accuracy and provide any updated recommendations. The model is trained and tested using

daily data from Tesla stock from June 29, 2010, to July 12, 2022. The data collection consists of 6 columns and 3031 rows. Every line reflects the day's data.

Data preparation comprises looking for missing values and removing them from the data collection, as well as looking for category values and removing extraneous information from the data source. Training data and test data make up the two sections of the data set. In this case, 2121 data points were considered as training data, while the remaining 909 data points were maintained for testing. The training data includes 2121 days, from June 29, 2010, to September 7, 2018, while the test data extends 909 days, from July 10, 2010, to July 12, 2022. Furthermore, to reduce the variable ranges, all the data is scaled with a typical scaler. Data from various methodologies can be scaled and compared in similar situations.

2. Methodology

2.1 Linear Regression

The association of two variables can be predicted using linear regression by applying by using the observed data and a linear formula. Independent variables and dependent variables are the two types of variables that are referred to as "variables". One well-liked technique for conducting predictive analysis is linear regression.

2.1.1 Evaluate the training set part of the model

Figure 1 shows a comparative chart of Linear Regression training set prediction performance (R^2=0.99739; MSE=1.3784; RMSE=1.1741).

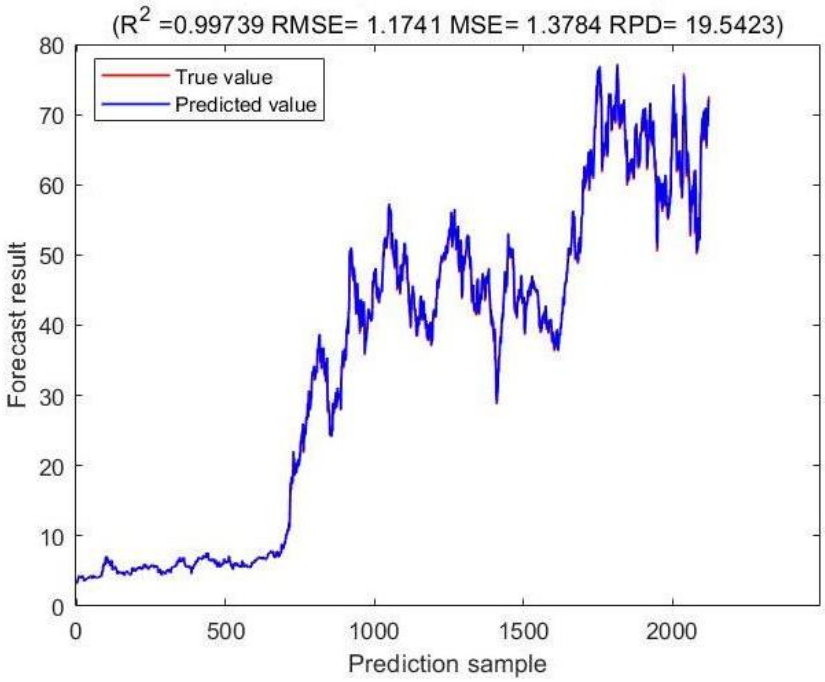

Figure 1. Comparison chart of training set prediction of Linear Regression

2.1.2 Evaluate the test set part of the model

According to Figure 2, the Linear Regression model's predictive ability is great (R^2=0.99586; MSE=529.2284; RMSE=23.005).

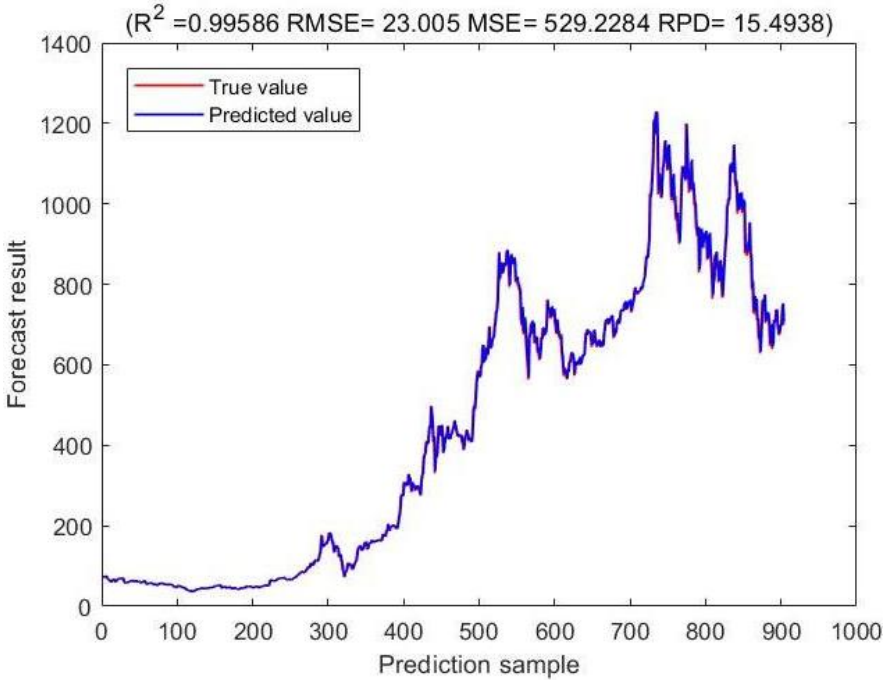

Figure 2. Comparison chart of test set prediction of Linear Regression

2.2 SVM

SVM is a type of two-way classification model. Figure 3 illustrates the use of SVM, a specially created machine learning algorithm, for classification and regression tasks. Its fundamental model is a linear classifier with the perceptron's smallest interval variances and the feature space's greatest interval set. The SVM algorithm, which is most commonly used for classification tasks, performs best in high-dimensional environments or when the number of samples surpasses the number of variables. This algorithm gives investors irregular returns on investments and also functions as a risk management tool. Convex quadratic programming is optimized using the SVM learning methodology.

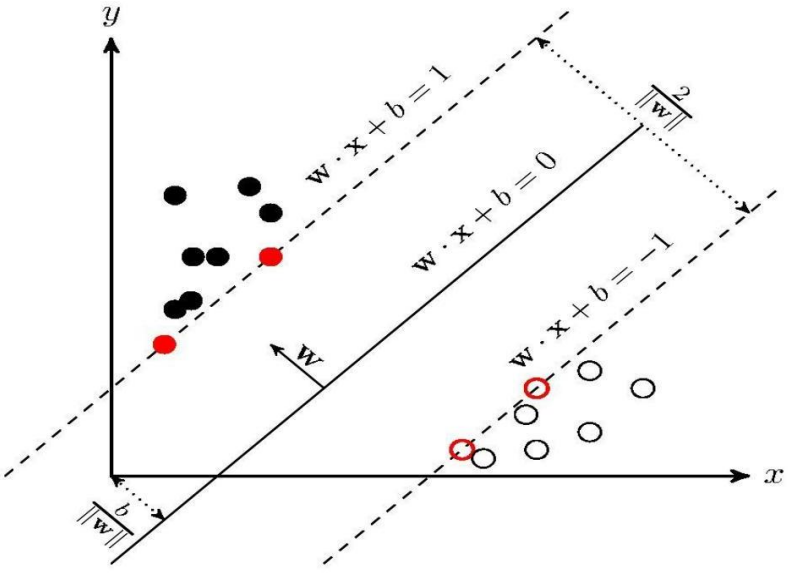

Figure 3. Schematic diagram of SVM

2.2.1 Evaluate the training set part of the model

The SVM training set prediction comparison chart is shown in Figure 4. The SVM model performed well (R^2=0.99738; MSE=1.3643; RMSE=1.168).

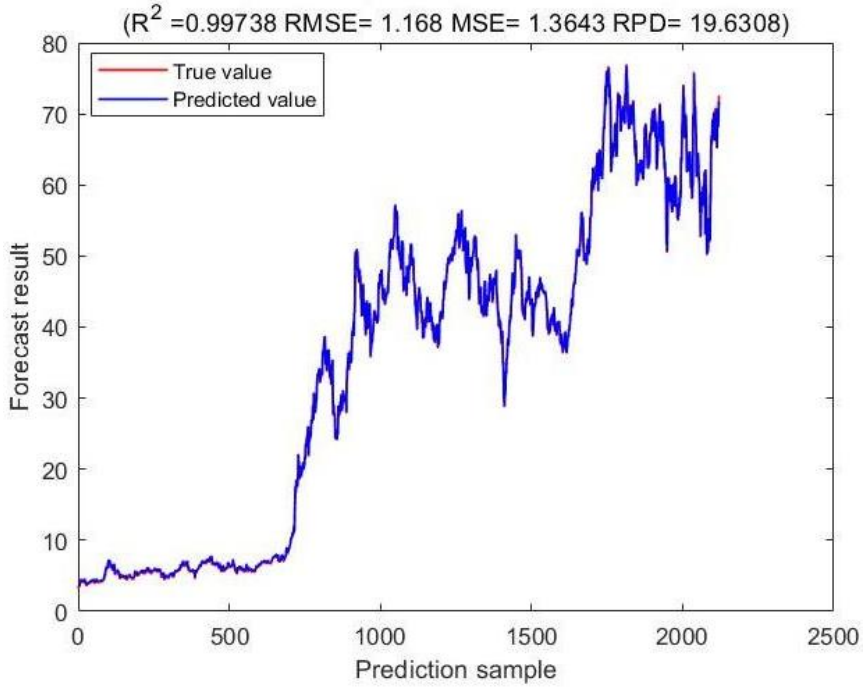

Figure 4. Comparison chart of training set prediction of SVM

2.2.2 Evaluate the test set part of the model

From Figure 5, it can be seen that the prediction effect of the SVM model is good (R^2=0.99605; MSE=497.7178; RMSE=22.3096).

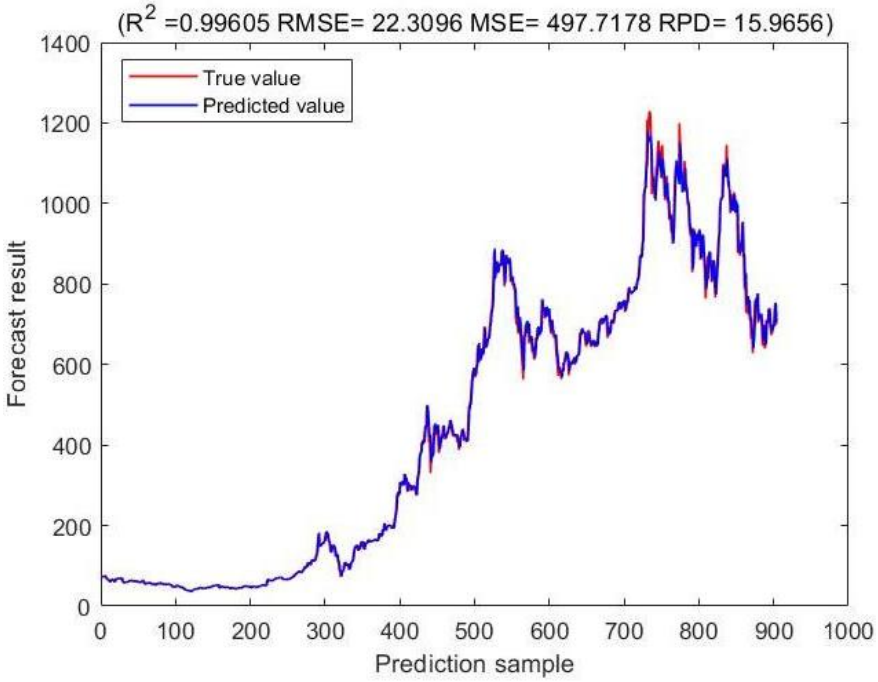

Figure 5. Comparison chart of test set prediction of SVM

2.3 Random Forest

Using a combination of bagging and characteristic randomization, the random forest methodology is a variant of the bagging methodology that creates an uncorrelated forest of decision trees. Figure 6 demonstrates how feature randomness, sometimes referred to as "the random subspace technique" or feature bagging offers a random selection of features that helps make sure little association between decision trees. Compared to random forests, decision trees are distinguished by this. Decision trees

evaluate every feature split, while random forests just choose a small portion of those that are accessible.

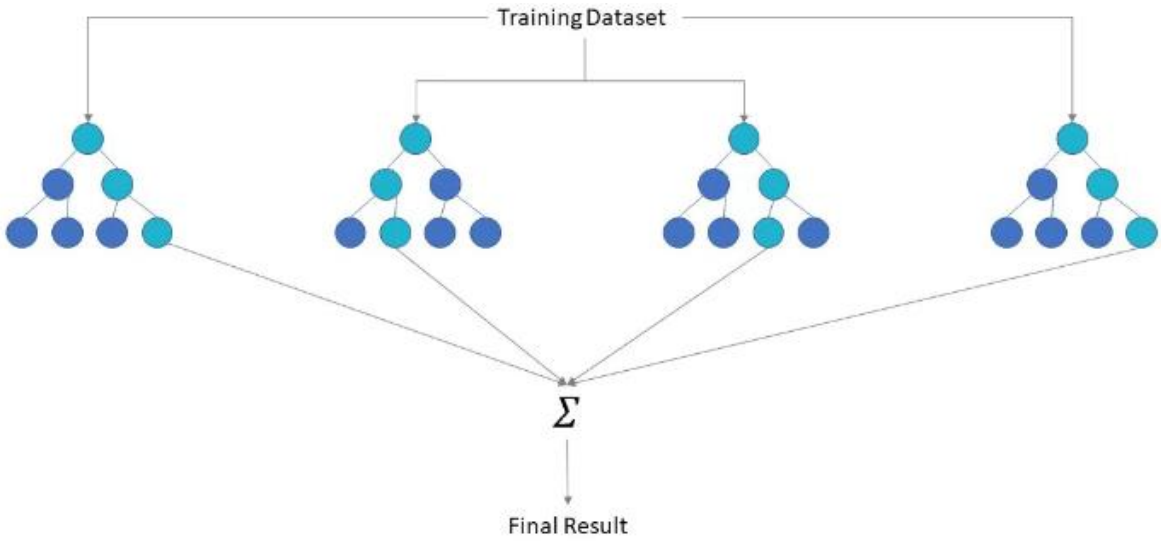

Figure 6. Random Forest architecture

2.3.1 Evaluate the training set part of the model

It can be seen from Figure 7 that compared with Linear Regression and SVM models, the Random Forest model has achieved the best training results (R^2=0.99939; MSE=0.31638; RMSE=0.56248).

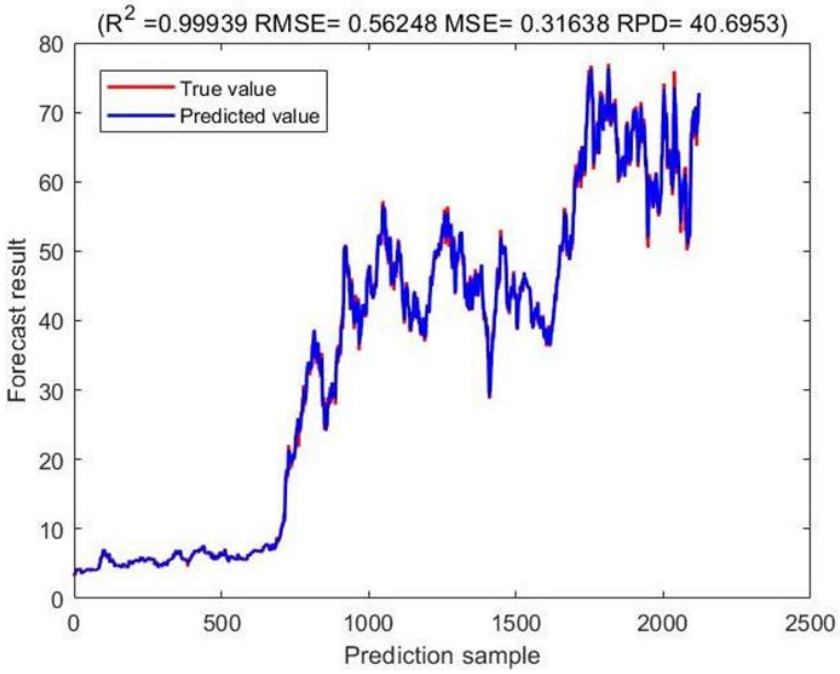

Figure 7. Comparison chart of training set prediction of Random Forest

2.3.2 Evaluate the test set part of the model

Similarly, it can be seen from Figure 8 that the prediction results of the random forest model are also the best (R^2=0.999; MSE=116.2317; RMSE=10.7811).

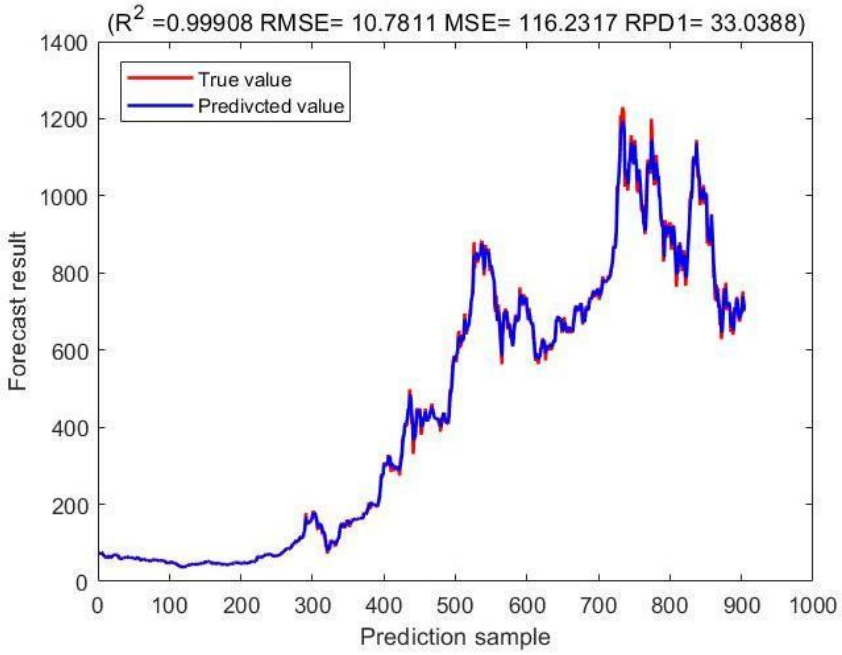

Figure 8. Comparison chart of test set prediction of Random Forest

2.4 LSTM

LSTM is a powerful RNN architecture. As shown in Figure 9, in the buried layer of the network, the LSTM introduces the memory cell. These memory cells may be efficiently connected with memories and information from distant points in time through networks, which enables them to foresee properly and understand the data structure dynamic over time.

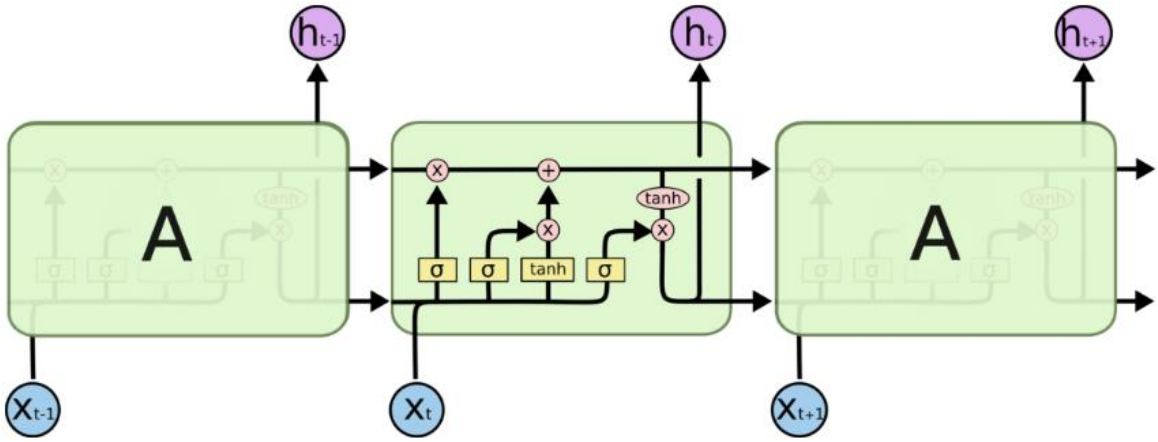

Figure 9. LSTM architecture

2.4.1 Evaluate the training set part of the model

Figure 10 shows a comparative chart of LSTM training set prediction performance (RMSE=1.1197).

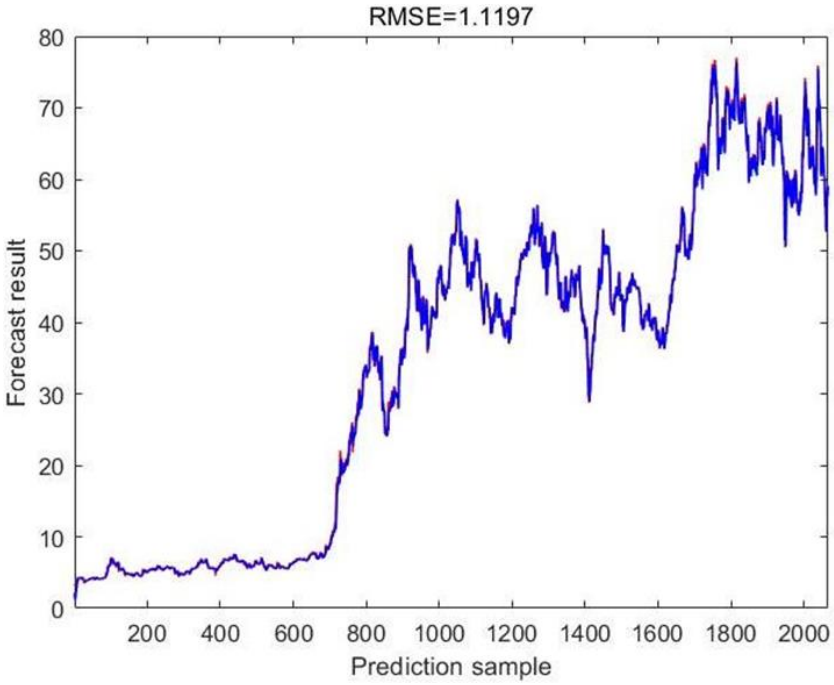

Figure 10. Comparison chart of training set prediction of LSTM

2.4.2 Evaluate the test set part of the model

The LSTM model appears to have underestimated the stock price of Tesla during its recent price boom, as seen in Figure 11.

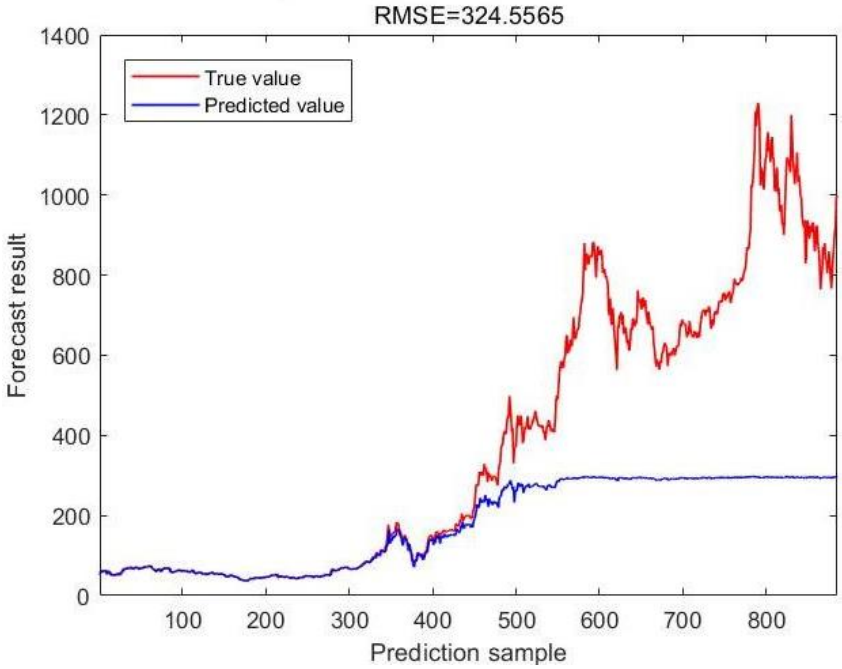

Figure 11. Comparison chart of test set prediction of LSTM

2.5 ARIMA

ARIMA is a more advanced version of the ARMA that combines the AR and MA processes to form a composite model.

AR: Auto-regression. A regression model based on the associations between such numerous delayed observations and a single observation (p).

I: Integrated. Determining the disparities between observations made at different periods in order to render the time series steady (d).

MA: Moving Average. A method for using a moving average model to analyze lagged observations that consider the dependence between the residual error terms and the observations (q).

2.5.1 Evaluate the training set part of the model

Figure 12 shows that the black line indicates the average of the forecasts, while the chosen region with the red dotted line represents the 95% confidence intervals and the ARIMA model predicted an increasing trend.

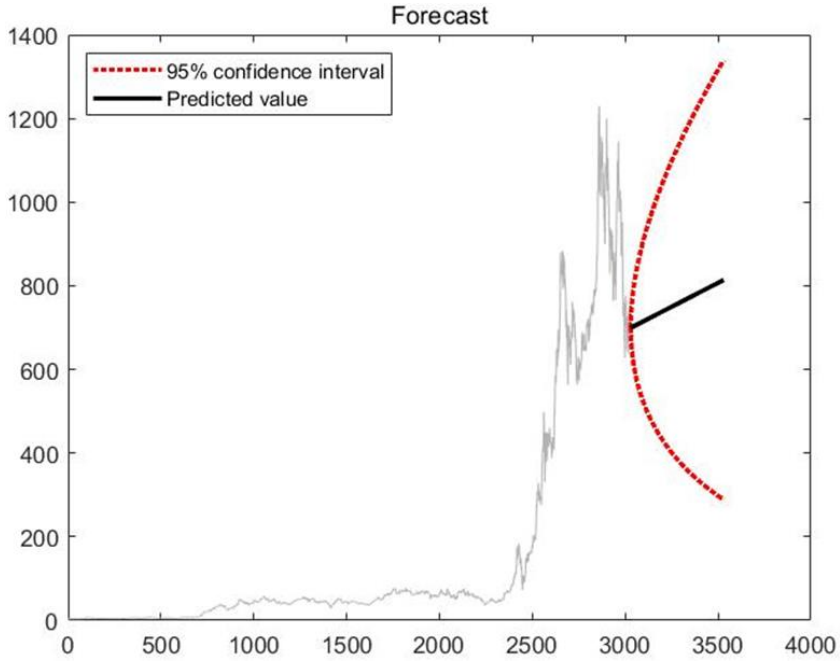

Figure 12. ARIMA (0,1,1)

3. Results and Discussion

The stock values of the firm Tesla Inc. are listed in the research's data. On June 29, 2010, Tesla began trading publicly. The business has raised US $226 million through its IPO. The original share price of the stock was $17.00 USD. In the eleven years since the stock became public, it has grown by over 4144%, and it is currently listed at roughly US$ 721 per share.

Table 1. Model results

	Train score (RMSE)	Test score (RMSE)	Train:Test
Linear Regression	1.1741	23.005	7:3
SVM	1.168	22.3096	7:3
Random Forest	0.56248	10.7811	7:3
LSTM	1.1197	324.5565	7:3
ARIMA	-	-	-

The NASDAQ index lists Tesla's shares. Elon Musk, the richest man in the world at the time of its founding, has served as its CEO and has overseen operations in several nations. Since becoming the company's CEO in 2008, Musk has consistently surpassed expectations and overcome obstacles.

The LSTM, SVM, Random Forest, and Linear Regression models' training outcomes are shown in Tables 1, 4, 7, and 10. The results show how reliable and strong the deep learning model is at predicting the current value based on the training data. Low forecast error rates are observed in Random Forest, as measured by the RMSE (0.56248).

The remaining 30 percent of the data were utilized in testing after training to validate the predictions generated by the models using deep learning. For assessing the advantages of the suggested machine learning models for estimating Tesla's current value, testing is essential. Figures 2, 5, and 8 indicate that the current and forecasted prices of Tesla's stock for the share marketing set of data throughout the training stage are precisely in line with one another. This is demonstrated by the fact that there isn't a big discrepancy between the actual and anticipated figures. Because the R-square values of Linear Regression, SVM, and Random Forest are all quite high, and the RMSE values are all very low, they were ready to be tested throughout the training period. These figures demonstrate how the system is capable of achieving the set objectives. The best results were obtained with the Random Forest model (R^2 =0.99908; MSE = 116.2317; RMSE = 10.7811).

Figure 11 illustrates how the LSTM model appears to have underestimated the stock price of Tesla during its recent price surge. When compared to other firms like AAPL and others, Tesla's training data for the LSTM model was substantially less, which may account for its low performance in predicting the stock price of the company. Consequently, its significant volatility might be blamed for the poor LSTM projection.

Figure 12 illustrates the ARIMA model. The black line in Figure 12 indicates the average of the forecasts, while the red dotted line in the specified area denotes the 95% confidence intervals. Despite an upward trend in the ARIMA model's prediction, the experimental data's actual value differs from what was predicted. Compared with other machine models, ARIMA, as a popular and widely used statistical method of time series prediction, is more suitable for predicting stocks with low noise and volatility, so there is no need to compare the results with other models.

The stock market is a very significant issue in today's society. Investors can easily acquire more stocks and stand to profit significantly from dividends issued as a component of the corporation's shareholder incentive plan. Investors may trade their personal equities on the stock market with other traders using stock brokerages and computerized trading platforms. Traders try to acquire stocks with rising prices and sell stocks with dropping prices on the share market. Investors in stocks must therefore be capable of accurately foreseeing the share's general behavior attributes prior to choosing an investment strategy to purchase or trade a share. They would get more money by making more accurate predictions about a stock's behavior. The creation of an autonomous algorithm that is capable of accurately forecasting market shifts is essential to assisting traders in maximizing their earnings. Predicting stock market trends, on the other hand, is difficult due to a range of elements such as market performance, corporate media and performance, economic variables, and investor attitudes. In this study, the capacity of different machine learning algorithms to forecast Tesla stock prices was explored. Figures 2, 5, 8, and 11 show how effective these models were during Tesla's testing and training stages. Random Forest outperformed other models in both the training (Random Forest: RMSE = 0.56248) and testing (Random Forest: RMSE = 10.7811) stages, as shown in Table 1. As a consequence, in terms of accuracy, Random Forest beat other models.

Because of its potential advantages, it could make it possible to foresee the future with credibility, which is something that most economies and people have long wished for. Understanding how to predict price fluctuations may be helpful for forecasting stock market enthusiasts. The use of artificial intelligence enables scientists to foresee with more accuracy than ever before. Its accuracy will also increase with time as both technological advancements and computational accuracy do.

4. Conclusion

To predict the Tesla share's closing price, this paper employed five different algorithms, including statistical and deep learning methods. It can be seen that the Random Forest prediction algorithm has the greatest forecast accuracy for the Tesla stock among machine learning algorithms.

Because it's possible that changes in the share market don't often adhere to a recognizable pattern or a constant cycle. The existence and longevity of trends will differ based on the organizations and industries. Investment results will increase with an awareness of these cycles and trends. We should

use an integrated algorithm technique like Random Forest, which has excellent anti-interference and anti-over-fitting characteristics, to assess highly volatile equities like Tesla. Deep learning models that include news stories about the economy and monetary factors like income statements, trade volume, etc. can be constructed for future work to produce potentially better outcomes.

References

[1] Dhankar R S. Stock Market Operations and Long-Run Reversal Effect in Capital Markets and Investment Decision Making India. Springer, 2019.

[2] Usmani M, et al. Ali Stock market prediction using machine learning techniques. 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), 2016, 322 - 327.

[3] Grigoryan H. A stock market prediction method based on support vector machines (svm) and independent component analysis (ica). Database Systems Journal, 2016, 7 (1): 12 - 21.

[4] Nti I K, Adekoya A F, Weyori B A. A systematic review of fundamental and technical analysis of stock market predictions. Artif. Intell. Rev, 2019, 53: 3007 – 3057.

[5] Sengupta A, Sena V. Impact of open innovation on industries and firms—A dynamic complex systems view. Technol. Forecast. Soc, 2020, 159: 120199.

[6] Terwiesch C, Xu Y. Innovation Contests, Open Innovation, and Multiagent Problem Solving. Manag. Sci, 2008, 54: 1529 – 1543.

[7] Blohm I, et al. Idea evaluation mechanisms for collective intelligence in open innovation communities: Do traders outperform raters? In Proceedings of the 32nd International Conference on Information Systems, Cavtat, Croatia, 2010, 21 – 24.

[8] Del Giudice, et al. The human dimension of open innovation. Manag. Decis, 2018, 56: 1159 – 1166.

[9] Daradkeh M. The Influence of Sentiment Orientation in Open Innovation Communities: Empirical Evidence from a Business Analytics Community. J. Inf. Knowl. Manag, 2021, 20: 2150031.

[10] Seber G A F, Lee A J. Linear regression analysis. John Wiley & Sons, 2012, 329.

[11] Sunil Ray. Svm | support vector machine algorithm in machine learning [internet]. Available from:https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/. Google Scholar. 2017.

[12] Henrique B M, Sobreiro V A, Kimura H. Stock price prediction using support vector regression on daily and up to the minute prices. J. Financ. Data Sci., 2018, 4 (3): 183 - 201.

[13] Liaw Andy, Matthew Wiener. Classification and regression by Random Forest. R news, 2002, 2 (3): 18 - 22.

[14] Kumar, Manish, Thenmozhi. Forecasting stock index movement: A comparison of support vector machines and random forest. In Indian institute of capital markets 9th capital markets conference paper, 2006.

[15] Roondiwala M, Patel H, Varma S. Predicting stock prices using LSTM. International Journal of Science and Research (IJSR), 2017, 6 (4), 1754 - 1756.

[16] Zhang Peter. Time series forecasting using a hybrid ARIMA and neural network mode. Neurocomputing, 2003, 50: 159 - 175.

[17] Siami-Namini S, Tavakoli N, Namin A S. A comparison of ARIMA and LSTM in forecasting time series. In 2018 17th IEEE international conference on machine learning and applications (ICMLA), 2018, 1394 - 1401.