Stock Price Prediction

Mandy Wong
Department of Computer Engineering
San Jose State University
California, United States

Shengtao Li
Department of Computer Engineering
San Jose State University
California, United States

Kangjun Lou
Department of Computer Engineering
San Jose State University
California, United States

Tian Lan
Department of Computer Engineering
San Jose State University
California, United States

***Abstract*—Nowadays, the stock market is the place where funds from all over the world gather. If people invest wisely, anyone can earn additional funds. However, it is difficult for people who are not familiar with the stock market to determine the types and prices of stocks. Therefore, the correct prediction of short-term and long-term stock prices is significant for all investors. With the development of machine learning and deep learning technology, there are now many algorithms for prediction, and the prediction accuracy of each algorithm is different. This article uses different algorithms to predict stock prices and analyzes and compares the results to get the most accurate prediction algorithm. The algorithms used in this article include K-Nearest Neighbor, Support Vector Machine, Random Forest, Auto Regressive Integrated Moving Average, Long Short Term Memory, Linear Regression, Logistic Regression, and Ensemble Learning.**

Keywords—stock price prediction, machine learning, deep learning, accuracy, SVM, K-NN, random forest, LSTM, ARIMA, linear regression, logistic regression, ensemble learning

I. INTRODUCTION

Stock market investment relies on fast and accurate information. Since there are many influencing factors in stock market trading, many of them are complex and uncontrollable, leading to frequent stock price fluctuations and challenging predictions. On the other hand, much wealth is flowing in the current stock market. People can get much wealth if people can accurately predict the trend of stocks and make corresponding investments. Therefore, stock price forecasting is a valuable and challenging process.

Nowadays, machine learning has become a powerful analytical tool for financial markets. By analyzing a variety of influencing factors, modeling the stock market, and using past data for training, the model finally obtained can predict the future short-term trend of the stock market to a certain extent. This model has been widely used in the financial field and helps investors make better investment and management decisions to maximize investment benefits.

Although it has been possible to make specific predictions about the stock market trend, the stock market trend determinants are very complex. Some factors that are excluded from the model building process due to technical or other reasons may also be significant for analyzing stock market fluctuations. The prediction accuracy of existing models has not been outstanding. Simultaneously, the stock market's uncertainty and volatility make the stock market always a high-risk area, and its high risk and possible huge cost make people always pursue more accurate prediction results. People have increased the research on the application of machine learning in the financial field, which has led to more and more algorithm applications in the financial field and the emergence of new models. More choices bring more opportunities and more confusion. Determining which of the multiple existing algorithms is the most appropriate has become an issue of investors' concern.

This article analyzes various algorithms in the existing financial field and explores the best-performing model among existing models by comparing the prediction accuracy of different algorithms. This project first gathers and preprocesses the stock data from *yfinance*, a python library of Yahoo Finance API, to generate five years of stock history dataset. The machine learning models used the stock dataset to train. A total of eight models were implemented to perform N days stock price prediction and stock classification.

The Related Work section described multiple machine learning models and their algorithms for making a prediction. The Methodology section states all the proposed technical aspects for stock price prediction and stock classification. The evaluation of each model performance is discussed in the Results section. Lastly, it would be the conclusion of this study.

II. RELATED WORK

With the successful application of machine learning in many fields, people continue to invest in machine learning research in the financial field. There are many related articles about stock price prediction algorithms. This article refers to some of them.

Work in [1] reviewed the research on machine learning techniques and algorithms used to improve stock price predictions' accuracy. This article first analyzes some general

influencing factors in the financial field and points out which parameters should be considered when modeling the financial field. Secondly, a very detailed analysis of the feasibility of existing algorithms in the financial field is carried out, and the advantages and disadvantages of each algorithm are listed. Finally, it gave a simple analysis of the algorithm suitable for stock price analysis. This article provides extensive background knowledge related to machine learning and finance and builds an overall knowledge framework for readers, but it does not profoundly analyze stock price prediction algorithms' implementation methods and performance.

Paper [2] analyzed and compared traditional statistical methods and machine learning methods used in stock price forecasting. This article first analyzes a variety of traditional statistical methods, such as exponential smoothing and naive methods, and then analyzes many machine learning methods, such as linear regression, K-NN, SVM, and LSTM. Afterward, the above methods are briefly compared in terms of predictive performance and accuracy. Although this article analyzes and compares various methods, especially machine learning methods, the comparison method is relatively simple, and the data used is not representative.

Paper [3] introduced the integrated time-varying effective transfer entropy (ETE) and analyzed the feasibility of combining it with various machine learning algorithms to predict stock price trends. This article analyzes how the logistic regression algorithm, the multilayer perceptron algorithm, the random forest algorithm, and the LSTM algorithm use ETE to make stock price predictions and conduct a horizontal comparison through experiments to analyze each algorithm's performance under ETE. The ETE mentioned in this article provides us with a new idea of extracting stock market features during data processing.

Work in [4] proposed and analyzed the feature-weighted support vector machine (FWSVM) and feature-weighted K-nearest neighbor (FWKNN) algorithms modified on the basic SVM and K-NN, and then tested the prediction accuracy of the two algorithms through experiments. This article explains the principles of SVM and K-NN and the design process of FWSVM and FWKNN so that readers can have an in-depth understanding of the principles and performance of these two new algorithms, but this article only involves these two algorithms and lacks a horizontal comparison.

Work in [5] [6] analyzed the application of the Long Short-Term Memory (LSTM) algorithm in stock price prediction. These two articles analyzed the LSTM algorithm's characteristics in detail and tested the accuracy of the LSTM algorithm in predicting stock prices through experiments. These two articles help us understand and use the LSTM algorithm.

Work in [7] [8] analyzed the ARIMA algorithm's application in stock price forecasting. These two articles analyze the ARIMA algorithm's characteristics in detail and compare them with the artificial neural network to analyze the prediction accuracy of the ARIMA algorithm and possible influencing factors. These two articles help us understand the implementation and use of the ARIMA algorithm.

III. METHODOLOGY

A. Stock Dataset

The stock history of Standard & Poor (S&P) 500 companies is used in this project. Yahoo Finance API gathers all the S&P 500's stock history, which the data are in *pandas* DataFrame. Instead of using the request function, there is a python library *yfinance.* If it is unable to get the history data, the invalid stock ticker will be removed from the S&P 500 ticker list. In this case, when updating the stock data, we do not need to deal with the empty list again. Fig. 1 shows the format of the stock dataset. The total size of the dataset is 623244 rows × 7 columns.

	Ticker	Date	Open	High	Low	Close	Volume
0	MMM	2015-11-16	134.812716	137.108538	134.812716	137.065231	2395000.0
1	MMM	2015-11-17	137.195155	137.342432	135.601074	136.042908	2393000.0
2	MMM	2015-11-18	136.195540	137.346642	136.108344	137.250717	2226900.0
3	MMM	2015-11-19	137.459997	138.227385	136.657720	138.114029	1518200.0
4	MMM	2015-11-20	138.375644	139.387216	138.079153	138.611099	1891100.0

Fig. 1. The Format of S&P 500's stock dataset

B. Prediction Overview

There are two methods to make the stock price prediction: regression and classification. For regressive prediction, it will predict one day, 30 days, 60 days, and even 1-year stock price. For stock classification, it will estimate if the next day's stock price goes up or goes down compared to the previous close price. Moreover, the prediction is performing on three stocks in different business categories: Apple Inc. (AAPL), JP Morgan Chase & Co. (JPM), Pfizer Inc. (PFE). Fig. 2 shows the portion of 5 years of stock data that is used for training the model and predicting the stock price in regression.

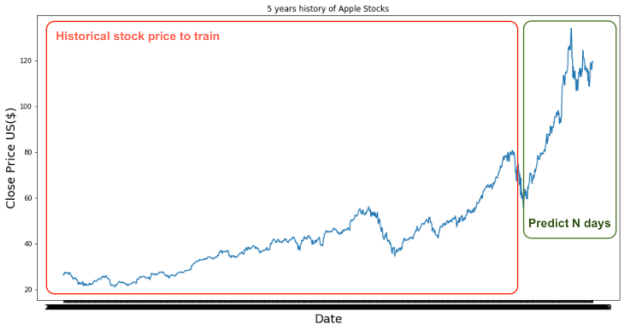

Fig. 2. The portion of 5 years stock data used for training and predicting

C. Prediction Regression

For daily stock price prediction, instead of using the historical close price, past open, high and low price will be used as an input to perform the prediction. The predicted value would be the daily stock closing price in this case. For 30 days and 60 days' stock price prediction, the historical close price is used as an input to make the prediction. The output value would be 30 days or 60 days' stock price.

There are four predictive models used to perform the one day, 30 days, and 60 days' stock price prediction. They are Linear Regression, K-NN Regressor, SVM RBF Regressor, and Random Forest Regressor. Basic four machine learning models

were applied using sci-kit learn python library. Since this approach is making a prediction based on regression analysis, the training model uses the regressor function of classification algorithms (K-NN, SVM, and RF) instead. The evaluation of daily price predictive models is based on Root Mean Square Error (RMSE), R2 score, and Accuracy. On the other hand, the assessment of N days price predictive models is based on the predicted trend and model accuracy.

We have also predicted a 1-year stock price. However, stock market data are non-stationary data and usually display seasonality and trend. Basic machine learning models cannot handle the time-series data well. Thus, we have also used ARIMA and LSTM to predict the stock price for a long time. For time series data, we can also use the historical data to predict future stock price. Here we choose ARIMA and LSTM models to perform time series analysis for target stock market data. We divide the data based on timestamp. The first 80% of the data will be used to train models, and the last 20% data will be used to perform test and prediction. Out of all the relevant data, the closing price at each day is selected.

For ARIMA, we take the logarithmic of the data to reduce the value and rising trend. Then, for each stock, we use "pmdarima.arima" auto_arima function to find out the optimal p,q, and d that is used for final ARIMA model for final training and prediction. The final prediction and test data are compared visually with a plot, and 4 metrics are calculated to measure the error: mean squared error, mean absolute error, root mean squared error, and mean absolute percentage error.

For LSTM, we have reformatted the data to map the previous 60 data points as input, and the current data as output. We designed neural networking using tensor flow keras library, with 4 hidden layers, each layer has 25 neurons, and the dropout rate is set to 0.2. There is one output neuron in the output layer. We use the mean squared error loss function and the Adam stochastic gradient descent optimizer. The model is trained for 50 epochs, and is used to predict the test data. The final prediction and test data are compared visually with a plot, and 4 metrics are calculated to measure the error: mean squared error, mean absolute error, root mean squared error, and mean absolute percentage error.

D. Prediction Classification

For stock prediction, it will predict whether the stock price will go up (+1) or go down (-1) the next day. The machine learning models use the historical open, high, and low stock prices to train. The predicted output is either +1 or -1. The value of +1 indicates that the stock price will go up the next day. There are six predictive models used to perform the stock classification. They are RBF SVM, Linear Regression, Logistic Regression, K-NN, Random Forest, and Ensemble learning using the Bagging Classifier with Decision Tree. The evaluation of stock classification is based on the training error and testing error for each machine learning model.

IV. Results

A. Stock Trend

Fig. 3 shows the five years stock price history of 3 stocks (Apple, JP Morgan Chase, and Pfizer). The plot illustrates the overview of the stocks' trend. The trend for both Apple Inc. and JP Morgan Chase & Co. is clearly in a positive direction with few big drops occasionally. However, the stock trend of Pfizer Inc. is not noticeable. Thus, the individual plot for Pfizer Inc. would be better for analyzing the stock trend. Fig. 4 shows the five years stock price history of Pfizer Inc. The stock trend is now distinct. The stock trend helps to understand how the historical stock price data should feed into the machine learning models. The accuracy of the predicted value would be affected if using incorrect training data.

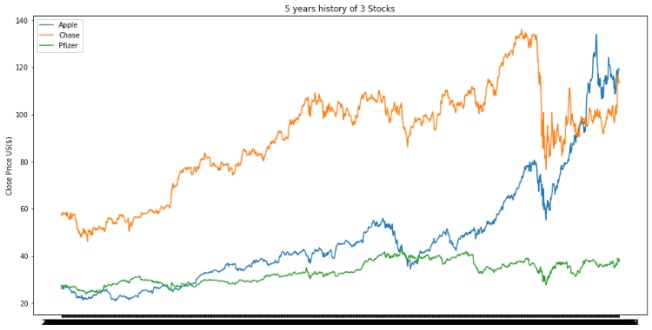

Fig. 3. Five years history of 3 stocks (Apple, Chase, and Pfizer)

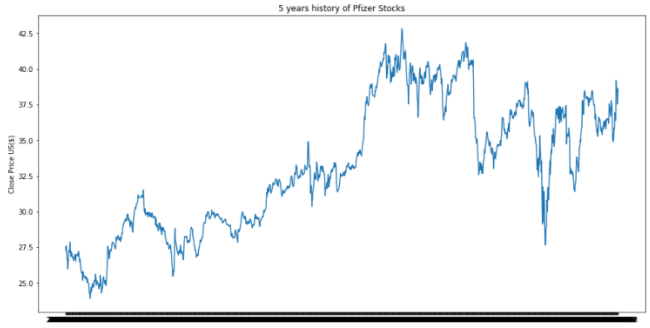

Fig. 4. Five years stock price history of Pfizer Inc

B. Prediction Regression

There are four predictive models used for the daily stock price prediction. They are linear regression, SVM RBF, K-NN, and Random Forest. Moreover, Fig. 5 displays the summary of the error measure for each model and each stock. Compared with the accuracy, surprisingly all the models perform so well. Yet, accuracy is not appropriate to decide the effectiveness of the predictive model in this situation. The predictive model may estimate the most similar price, but the small difference in price may impact a lot in the stock market. Therefore, the predicted price should be precise but not accurate. R^2 score shows that the data fit the model and the regression. From Fig.5, the R^2 score of all models is incredibly high, which means the data is well fitted. Lastly, by comparing the RMSE, RBF kernel SVM contains a higher error rate than the other three models.

Error Measure of Daily Stock Predication

Apple

	RMSE	R2	Accuracy
LinReg	0.432816	0.999685	0.999685
K-NN	0.651733	0.999285	0.999285
SVM_RBF	2.448366	0.989910	0.989910
RanForest	0.592638	0.999409	0.999409

JP Morgan Chase

	RMSE	R2	Accuracy
LinReg	0.604663	0.999172	0.999172
K-NN	0.794015	0.998572	0.998572
SVM_RBF	1.315082	0.996084	0.996084
RanForest	0.853867	0.998349	0.998349

Pfizer

	RMSE	R2	Accuracy
LinReg	0.165957	0.998727	0.998727
K-NN	0.201243	0.998128	0.998128
SVM_RBF	0.210889	0.997944	0.997944
RanForest	0.207013	0.998019	0.998019

Fig. 5. Error Measure of Daily Stock Prediction

For N days stock price prediction, there are four predictive models used to predict 30 days and 60 days separately. Four predictive models are linear regression, SVM RBF, K-NN, and Random Forest. Moreover, ARIMA and LSTM were used to predict one-year stock prices because of their ability to well handle the time-series data.

showing in Fig.3, both Apple stock and JP Morgan Chase stock experienced a wide increasing price range within these five years. It may be challenging when predicting the stock price for a long period of time using a basic machine learning model. Fig. 7 shows the accuracy of these two predictions for each model and each stock. The Apple stock performs the best on all four models compared to the other two company stocks.

Accuracy of 30 Days Prediction

	Apple	Chase	Pfizer
LinReg	0.930080	0.819259	0.746580
K-NN	0.944150	0.876220	0.771005
SVM_RBF	0.934879	0.863838	0.783859
RanForest	0.939475	0.862834	0.676653

Accuracy of 60 Days Prediction

	Apple	Chase	Pfizer
LinReg	0.855934	0.721500	0.618651
K-NN	0.893706	0.892199	0.682245
SVM_RBF	0.879668	0.894687	0.701652
RanForest	0.851232	0.870691	0.606602

Fig. 7. Accuracy of 30 days and 60 days prediction

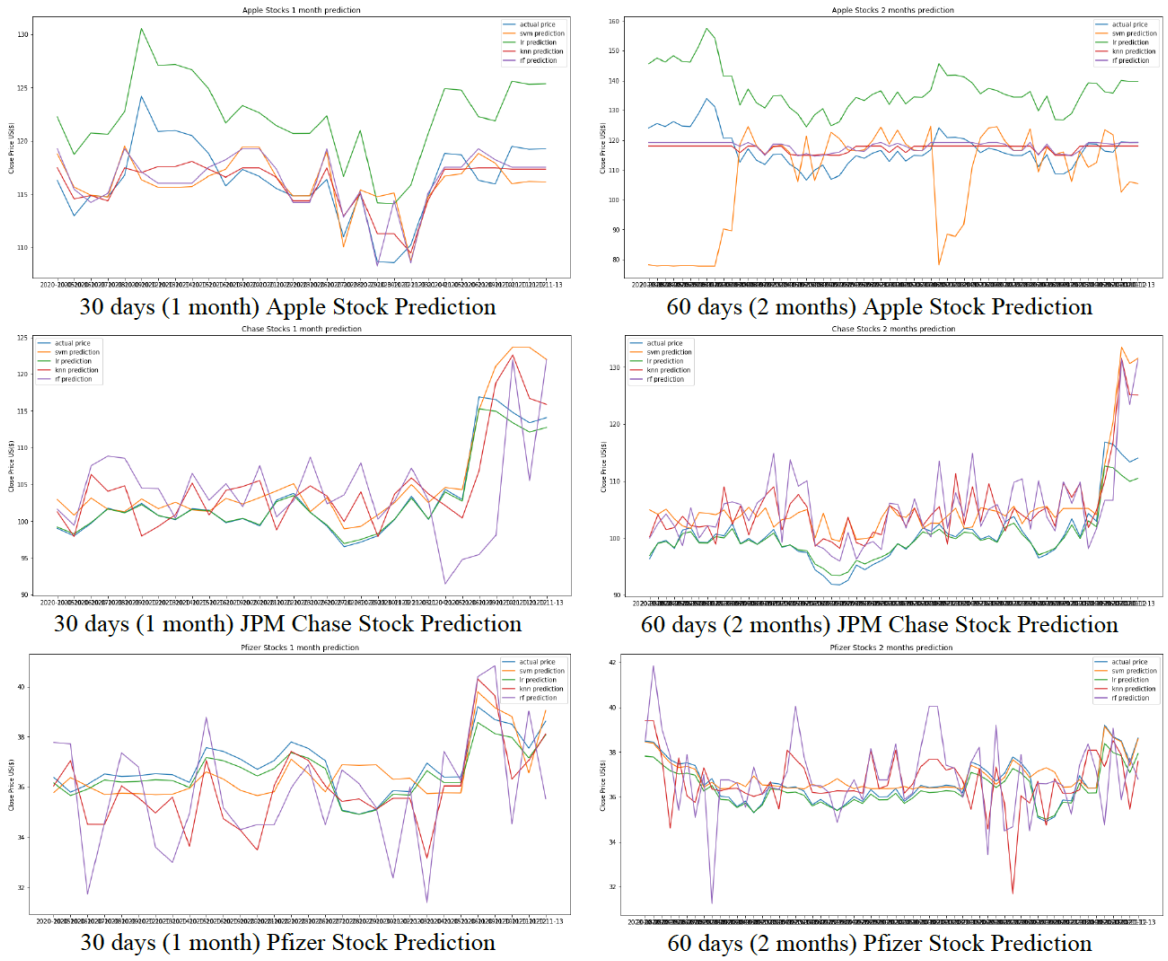

Fig. 6. All price predictions for 30 days and 60 days (Blue line: actual value; Orange line: SVM; Green line: Linear Regression; Red line: K-NN, Purple: Random Forest)

Fig. 6 demonstrates the 30 days and 60 days stock prediction of each stock. The blue line indicates the actual stock price. Most of the models follow the trend and are similar to the real stock price. For the Apple stock, it is noticeable that the linear regression model (the green line) followed the trend but predicted a higher value. According to the 5-years historical data

Furthermore, ARIMA and LSTM models were used to predict one-year stock prices. Fig. 8-10 shows the prediction result for the 3 different stocks using ARIMA. Although the ARIMA model can somehow predict the overall trend, the overall error rate is somewhat high (6-12% MAPE). LSTM can predict both the overall trend and the fluctuations better (3-6% MAPE). Given that we only use 4 layers of 25 neurons and epochs of 50 times, there are still rooms for improvement for the LSTM model.

Time Series: ARIMA, table 1 shows the result.

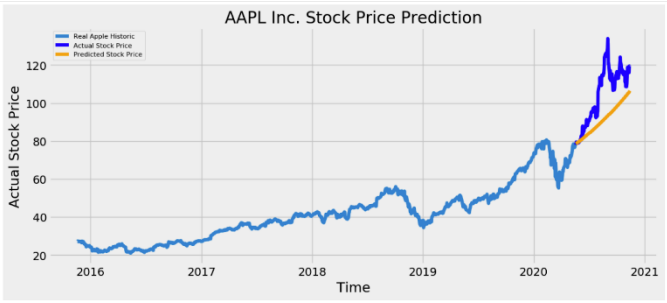

Fig. 8. Apple Inc stock price prediction using ARIMA

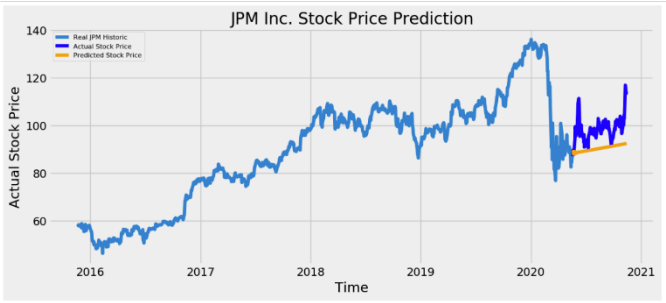

Fig. 9. JPMorgan Inc stock price prediction using ARIMA

Fig. 10. Pfizer Inc stock price prediction using ARIMA

TABLE I. ARIMA TIME SERIES RESULT

ARIMA	Error		
	Apple Inc Stock	***JPMorgan Inc Stock***	***Pfizer Inc Stock***
MSE	265.21	91.87	6.73
MAE	13.46	8.31	2.02
RMSE	16.29	9.59	2.59
MAPE	0.12	0.08	0.06

Time Series: LSTM. Fig. 11-13 shows the prediction result for the 3 different stocks using LSTM. Table 2 shows the result.

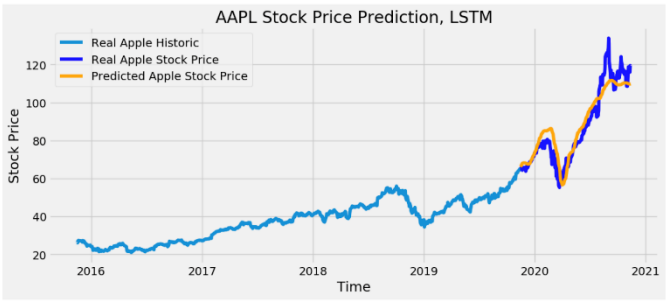

Fig. 11. Apple Inc stock price prediction using LSTM

Fig. 12. JPMorgan Inc stock price prediction using LSTM

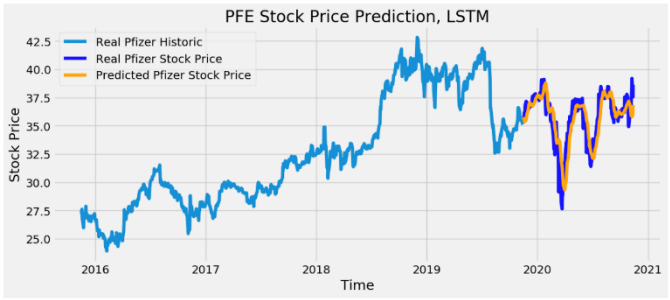

Fig. 13. Pfizer Inc stock price prediction using LSTM

TABLE II. LSTM TIME SERIES RESULT

LSTM	Error		
	Apple Inc Stock	***JPMorgan Inc Stock***	***Pfizer Inc Stock***
MSE	45.52	50.13	1.85
MAE	5.38	5.16	1.03
RMSE	6.75	7.08	1.36
MAPE	0.06	0.05	0.03

C. Prediction Classification

There are six predictive models used for the stock classification. They are linear regression, SVM RBF, K-NN, Logistic Regression, Random Forest, and Ensemble Learning. The Bagging Classifier with Decision Tree used in Ensemble Learning. Fig. 14 shows the training error and test error of each model. Linear Regression performs the worst among the other six models. The error indicates that basic machine learning algorithms are not suitable for doing the stock classification as the stock price is time-series data.

Error Measure of Stock Classification

Apple Stock

	RBF SVM	LinReg	LogReg	K-NN	RandForest	EnsembleL
Train Error	0.463754	0.996503	0.453823	0.298908	0.022840	0.527245
Test Error	0.452381	1.053467	0.472222	0.460317	0.452381	0.501119

JP Morgan Chase

	RBF SVM	LinReg	LogReg	K-NN	RandForest	EnsembleL
Train Error	0.486594	0.999620	0.486594	0.295929	0.034757	0.513364
Test Error	0.503968	1.000779	0.503968	0.511905	0.523810	0.460078

Pfizer

	RBF SVM	LinReg	LogReg	K-NN	RandForest	EnsembleL
Train Error	0.465740	0.998834	0.480636	0.318769	0.030785	0.501527
Test Error	0.559524	1.024887	0.555556	0.571429	0.472222	0.547854

Fig. 14. Error Measure of Stock Classification

V. CONCLUSION

This project implemented a total of eight models for stock price prediction and stock classification. Interestingly, the accuracy of predictive models is usually much higher than the classification models. However, the small difference in price may impact a lot in the stock market. In this situation, accuracy may not be appropriate to decide the effectiveness of the predictive model. For stock classification, the linear regression model performs the worst among other models by comparing the training and testing error. Furthermore, ARIMA and LSTM can predict the general trend of the stock market based on previous historical data. Although the accuracy of both models can be still improved, LSTM shows the ability to provide both trend and fluctuation predictions. Maybe in the future it's a good idea to investigate different parameter optimizations to optimize the overall trend and fluctuation prediction. Predicting the stock price is challenging because of the complex structure of the finance market. It is not possible to predict the price only based on the historical price data. Other features like the trend and impacts of news should also take into consideration when developing the stock price predictive model.

VI. CONTRIBUTION

Name	Tasks
Mandy Wong	Data preprocessing, SVM, Linear Regression, Logistic Regression, Ensemble Learning Paperwork
Shengtao Li	Data preprocessing, Random Forest, Neural Network, ARIMA, LSTM Paperwork
Kangjun Lou	Data preprocessing, KNN, Neural Network Paperwork
Tian Lan	Data preprocessing

REFERENCES.

[1] Obthong M, Tantisantiwong N, Jeamwatthanachai W, et al. A survey on machine learning for stock price prediction: algorithms and techniques[J]. 2020. Web.

[2] I. Bhattacharjee and P. Bhattacharja, "Stock Price Prediction: A Comparative Study between Traditional Statistical Approach and Machine Learning Approach," 2019 4th International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh, 2019, pp. 1-6, doi: 10.1109/EICT48899.2019.9068850

[3] S. Kim, S. Ku, W. Chang and J. W. Song, "Predicting the Direction of US Stock Prices Using Effective Transfer Entropy and Machine Learning Techniques," in IEEE Access, vol. 8, pp. 111660-111682, 2020, doi: 10.1109/ACCESS.2020.3002174.

[4] Chen, Yingjun, and Hao, Yongtao. "A Feature Weighted Support Vector Machine and K-nearest Neighbor Algorithm for Stock Market Indices Prediction." Expert Systems with Applications 80 (2017): 340-55. Web.

[5] Rana M, Uddin M M, Hoque M M. "Effects of Activation Functions and Optimizers on Stock Price Prediction using LSTM Recurrent Networks[C]". Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence. 2019: 354-358.

[6] Wen, Yulian, Peiguang Lin, and Xiushan Nie. "Research of Stock Price Prediction Based on PCA-LSTM Model." MS&E 790.1 (2020): 012109.

[7] Mondal, Prapanna, Labani Shit, and Saptarsi Goswami. "Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices." International Journal of Computer Science, Engineering and Applications 4.2 (2014): 13.

[8] Ayodele Ariyo Adebiyi, Aderemi Oluyinka Adewumi, Charles Korede Ayo, "Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction", Journal of Applied Mathematics, vol. 2014, Article ID 614342, 7 pages, 2014. https://doi.org/10.1155/2014/614342

SUPPLEMENTARY.

Code Link: wmymandy/stock_price_prediction: CMPE 257 Team Project (github.com)