Abstract

The impact of numerous factors on stock prices makes stock prediction a complex and time-consuming endeavour. Predicting the price of a stock is computationally hard because of its non-stationary nature and also it depends on many factors like News Headlines, Tweets, Historical Trends, Social Media News etc. In this paper, Machine Learning Algorithms and Neural Networks are implemented on various Companies like Apple, Amazon, Pfizer, Walmart Stores etc to overcome the difficulties and to achieve better accuracy in predicting the price of a stock. Artificial Intelligence algorithms like Random Forest, XG Boost (Extreme Gradient Boosting), LSTM (Long Short Term Memory), GRU (Gated Recurrent Units) etc are developed and their RMSE (Root Mean Square Error) are compared in predicting the price of a stock. The Dataset is an open-source Time Series dataset and consists of stock prices for 88 different companies that fall under 9 different sectors for around 5 years.

Key Words

Stock Market, Regression, MinMax Scaler, Supervised Learning, Boosting, Neural Networks, Machine Learning.

Table of Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
ARIMA	Autoregressive Integrating Moving Average
GRU	Gated Recurrent Units
LSTM	Long Short Term Memory
ML	Machine Learning
NLP	Natural Language Processing
RMSE	Root Mean Square Error
RNN	Recurrent Neural Networks
SARIMAX	Seasonal Autoregressive Integrating Moving Average
SVM	Support Vector Machines
XG Boost	Extreme Gradient Boosting

1. Introduction

The well-being of every developing economy mainly hinges on their market economies and stock price, with the financial market being the pivot. Thus, it is essential to study and learn about the financial market extensively. A stock market is a place for trading equity and other financial instruments of public listed companies, where the price of shares is termed "share" or "stock price". In reality, the stock price level of a firm, to a large extent, reflects how it "cuts its pie."

Stock price prediction is very crucial for a country's economic growth and also plays a vital role for developing countries than the developed countries but predicting the price of a stock is computationally hard [1] because of its non-stationary nature (such as general economic conditions, social factors, and political events at both homegrown and international levels) and also it depends on several other factors like Social media news, Historical trends [2], Tweets etc. In recent years due to the COVID-19 pandemic, there were a lot of fluctuations in the stock market due to the rise and fall of several new cases daily. The Efficient Market Hypothesis states that stock prices reflect all current information, and new information leads to unpredictable stock prices. The Random Walk concluded that stock prices could not be accurately predicted using historical values [36]. However, the attraction of good returns has led to numerous methods for price prediction. In the last three decades, much research has been done in this area. But still, researchers are of the view that prediction of stocks on non-linear non-stationary financial time series is one of the most challenging tasks. Several mathematical models have been developed, but the results are still dissatisfying [37]. Studies focusing on forecasting the stock markets have been mostly preoccupied with forecasting volatilities [38].

Some form of prediction often guides investments in the stock markets - two main stock market prediction approaches: fundamental and analysis, technical analysis. Our work used a technical analysis approach to predict the future movements of stock market.

The prediction methods in stock market analysis can play a crucial role in bringing awareness to more investors and people around the world to invest in the stock market. Existing works [3] mainly focuses on ARIMA and SARIMAX Machine learning algorithms however other popular models like Support vector machines, Random forests and Naive Bayes [4] has outperformed in terms of performance. Also, LSTM can be implemented that considers the whole previous data to predict the future whereas traditional algorithms consider a set of continuous previous data to predict the future. So, LSTM in RNN is better than ARIMA and SARIMAX models.

This research seeks to apply machine learning techniques and propose new models that generate an acceptable prediction accuracy when predicting popular stocks in U.S stock market. Succinctly, this work contributes to the body of knowledge as summarised :-

1. Used Random Forest to predict 88 stocks from 9 industries and gotten good results for 79 of them.
2. Fifty-five companies have outperformed historical models and gotten <=0.5 RMSE, and 24 companies achieved an RMSE around 0.5 to 1, and 9 companies are getting RMSE more than one but below ten.

2. Literature Survey

Market analysis' primary objective is to comprehend the market's behaviour in order to aid investors in making more informed decisions. Numerous market attributes and characteristics associated with time series data on stock prices have been investigated. Market analysis can be classified into two types based on the market factors that are used: fundamental and technical analysis [6].

2.1 Fundamental Analysis

Fundamental analysis [Figure 1] assumes that the related factors are the internal and external attributes of a company. These attributes include the interest rate, product innovation, the number of employees, the management policy and etc. [7]. In order to improve the prediction, other information such as the exchange rate, public policy, the Web, and financial news are used as features.

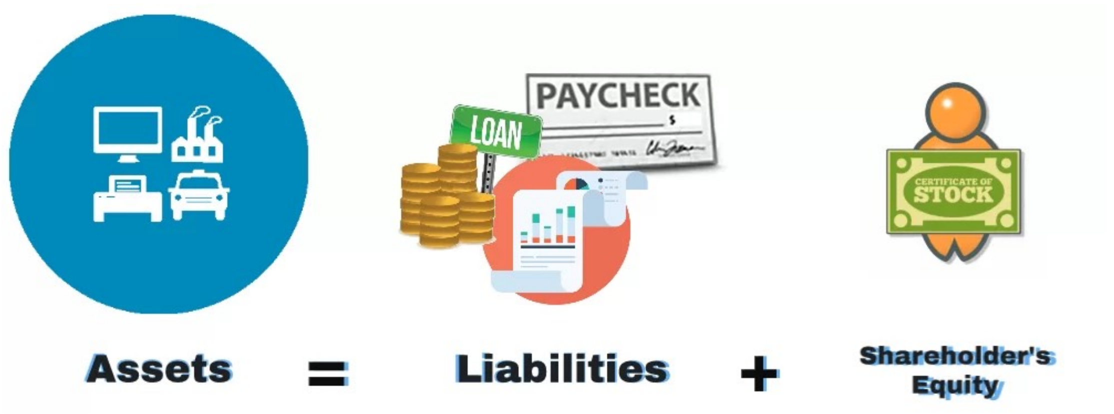


Figure 1: Fundamental Analysis of Stocks

For example, Nassirtoussi et al. used news headlines as features and proposed a “semantics-sentiment” dimension reduction algorithm with multi-layer perception to predict intra-day movements of the market [8].

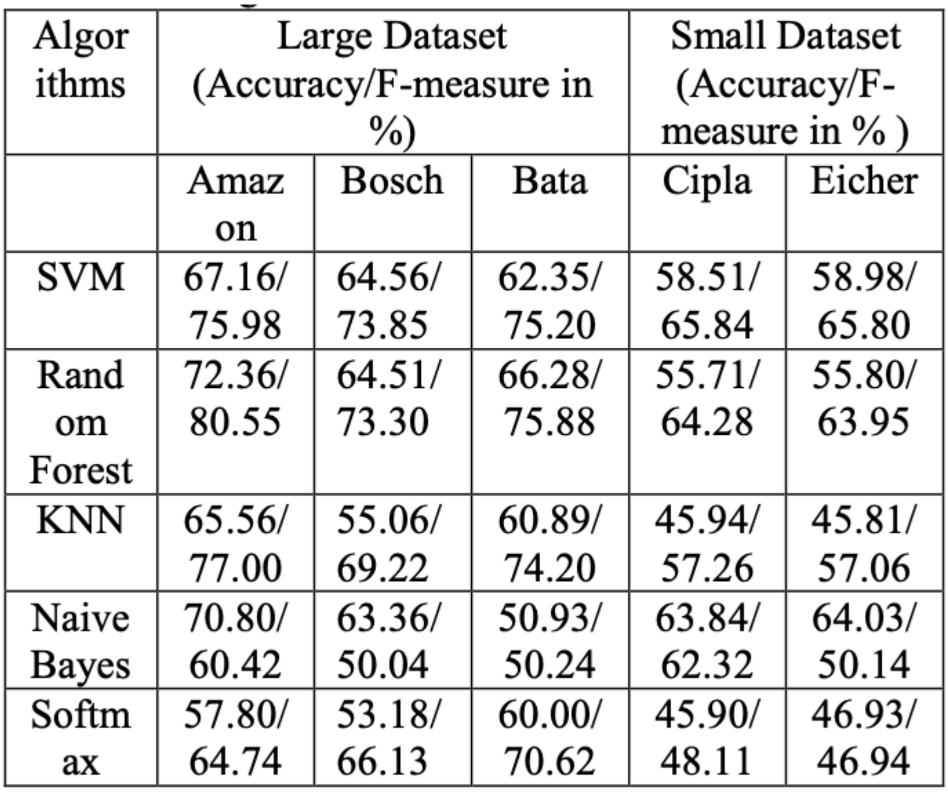
Due to the unstructured nature of fundamental factors, automation of fundamental analysis is difficult. On the other hand, the emergence of machine learning has enabled researchers to automate stock market prediction based on unstructured data, which in some cases has reported higher prediction accuracy. Nonetheless, fundamental analysis is useful for long-term stock-price movement, but not suitable for short-term stock-price change [39].

2.2 Technical Analysis

On the other hand, technical analysis frequently relies solely on historical prices as market characters to denote price movement patterns. The studies assume that relative factors are incorporated into the market price movement and that history repeats itself. Certain investors have had considerable success forecasting stock prices using technical analysis [9]. However, the Efficient Market Hypothesis (EMH) [10] assumes that market efficiency always results in prices that incorporate and reflect all market information, such that only new information affects the movement of market prices, and new information is unpredictable.

In the technical analysis literature, a variety of stock movement prediction approaches have been proposed, ranging from Autoregressive Integrated Moving Average (ARIMA) to ensemble methods [11]. Huang et al. [12] used Support Vector Machines (SVM) to forecast the NIKKEI 225 index's weekly movement directions, and Lin et al. [13] combined decision trees and neural networks to achieve a prediction accuracy of 70% for stock movement prediction. Recently, Khuwaja et al. proposed a framework for predicting stock price movement by combining phase space reconstruction (PSR) and extreme learning machines (ELM). The experimental results indicate that this approach is capable of greater predictive accuracy than conventional machine learning methods [14]. However, the methods discussed above focus exclusively on linear characteristics and overlook the non-linear and non-stationary nature of stock prediction, limiting their performance in practice [15]. With the development of the deep learning framework, a new wave of stock prediction methods has been proposed [16–22]. The Long Short-Term Memory (LSTM) recurrent neural network, in particular, has been shown to be extremely effective at stock prediction [23-25]. From the previous papers [4] the average Accuracy scores and F1 Scores for few companies are mentioned in the following Table 1:-

Table 1: Performance Metrics from Previous Papers [4]



Algorithms	Large Dataset (Accuracy/F-measure in %)			Small Dataset (Accuracy/F-measure in %)	
	Amazon	Bosch	Bata	Cipla	Eicher
SVM	67.16/75.98	64.56/73.85	62.35/75.20	58.51/65.84	58.98/65.80
Random Forest	72.36/80.55	64.51/73.30	66.28/75.88	55.71/64.28	55.80/63.95
KNN	65.56/77.00	55.06/69.22	60.89/74.20	45.94/57.26	45.81/57.06
Naive Bayes	70.80/60.42	63.36/50.04	50.93/50.24	63.84/62.32	64.03/50.14
Softmax	57.80/64.74	53.18/66.13	60.00/70.62	45.90/48.11	46.93/46.94

2.3 Combining Fundamental and Technical Analysis

According to Lui et al., technical analysis is more advantageous for short-term forecasting. Thus, it is applicable to high-frequency trading, whereas fundamental analysis provides a more accurate forecast of trends [26]. 85 percent of respondents rely on fundamental and technical analysis [26]. Numerous researchers have attempted to improve forecasting accuracy by combining fundamental and technical analyses. For example, Ding et al. described a method for event-driven deep learning prediction that incorporates a knowledge graph into the event embeddings [27]. Chen et al. enhanced Ding et almethod .'s by factoring in specific semantic information about diverse event types derived from coarse-grained events. [28] To summarize, the work of Ding et al. and Chen et al. established the advanced nature of leveraging event semantics for stock price movement prediction. Apart from event-driven approaches, prediction using superior neural networks has made tremendous strides as well. For example, Wen et al. proposed a novel end-to-end model dubbed the multi-filters neural network (MFNN) for extracting features from financial time series samples and forecasting price movement [29]. Its novel multi-filter structure effectively captures data from a variety of feature spaces and market perspectives. Feng et al. hypothesized that input features based on stock prices are stochastic in nature and change over time. They used adversarial training on a simple Attentive LSTM model to add perturbations to simulate stochasticity and achieved state-of-the-art performance on the target dataset [30]. Attention mechanisms combining NLP techniques [31] have been widely applied to the stock prediction task in these proposed advanced neural frameworks. Hu et al. proposed a framework for stock trend prediction using deep learning and hybrid attention networks (HAN) [32].Xu et al. proposed a deep generative model StockNet based on HAN that utilizes tweet and historical price data and achieves state-of-the-art performance [33].

3. Dataset

The dataset [5] is a Time Series data [Figure 2] and consists of stock prices of 88 different companies like Apple, Amazon, Chevron Corporation, Sanofi, Duke Energy Corporation, Visa, Alphabet etc. These companies fall under 9 different categories namely Basic Materials, Consumer Goods, Healthcare, Services, Utilities, Conglomerates, Financial, Industrial Goods and Technology. In total there are 88 files in the dataset and each file consists of features like Date, Open price of a stock when it was opened on a particular day, High and Low price of a stock within a period, Volume of the stocks and the Adjusted closing price of an individual company. The output (or) the predicted variable is the Closing price of a stock for a particular day.

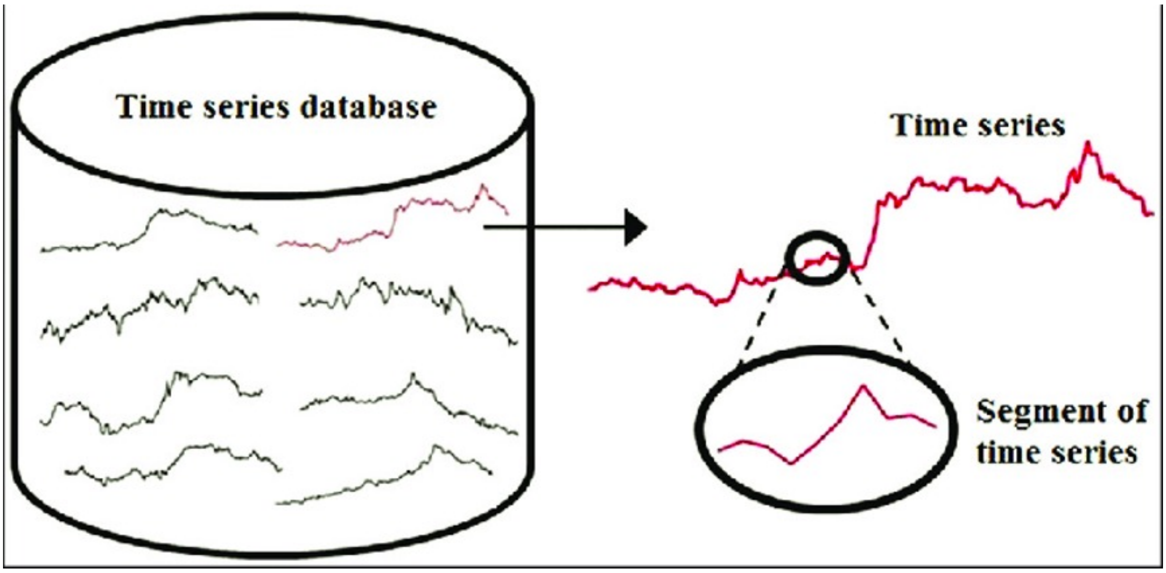


Figure 2: Time Series Data

To train the Machine Learning algorithm like Random Forest Regressor, initially, all the CSV files are loaded and converted into Data frames then Scaling is done on all the Data frames such that each feature is translated to a given range. MinMax Scaler can be implemented for scaling to normalise all the features because each feature in the dataset is of a different scale and it is very important to scale each feature before it is sent to the model for training. Also, the dataset consists of a date feature but this is not understood by the algorithm so, the datasets have to be pre-processed by splitting the date column into three different columns (Year, Month and Day). At last, the dataset is split for training and testing data. The stock symbols for all the companies under different sectors are as follows:-

Sector - Basic Materials:

Exxon Mobil Corporation (XOM), Royal Dutch Shell plc (RDS-B), Petro China Company Limited (PTR), Chevron Corporation (CVX), TOTAL S.A. (TOT), BP p.l.c. (BP), BHP Billiton Limited (BHP), China Petroleum & Chemical Corporation (SNP), Schlumberger Limited (SLB), BHP Billiton plc (BBL).

Sector - Consumer Goods:

Apple Inc (AAPL), The Procter & Gamble Company (PG), Anheuser-Busch InBev SA/NV (BUD), The Coca-Cola Company (KO), Philip Morris International Inc (PM), Toyota Motor Corporation (TM), Pepsico, Inc (PEP), Unilever N.V (UN), Unilever PLC (UL), Altria Group, Inc (MO).

Sector - Healthcare:

Johnson & Johnson (JNJ), Pfizer Inc (PFE), Novartis AG (NVS), UnitedHealth Group Incorporated (UNH), Merck & Co., Inc (MRK), Amgen Inc (AMGN), Medtronic plc (MDT), AbbVie Inc (ABBV), Sanofi (SNY), Celgene Corporation (CELG).

Sector - Services:

Amazon, Inc (AMZN), Alibaba Group Holding Limited (BABA), Wal-Mart Stores, Inc (WMT), Comcast Corporation (CMCSA), The Home Depot, Inc (HD), The Walt Disney Company (DIS), McDonald's Corporation (MCD), Charter Communications, Inc (CHTR), United Parcel Service, Inc (UPS), The Priceline Group Inc (PCLN).

Sector - Utilities:

NextEra Energy, Inc (NEE), Duke Energy Corporation (DUK), Dominion Energy, Inc (D), The Southern Company (SO), National Grid plc (NGG), American Electric Power Company, Inc (AEP), PG&E Corporation (PCG), Exelon Corporation (EXC), Sempra Energy (SRE), PPL Corporation (PPL).

Sector - Conglomerates:

Icahn Enterprises L.P. (IEP), HRG Group, Inc (HRG), Compass Diversified Holdings LLC (CODI), REX American Resources Corporation (REX), Steel Partners Holdings L.P. (SPLP), PICO Holdings, Inc (PICO), AgroFresh Solutions, Inc (AGFS), Global Medical REIT, Inc (GMRE).

Sector - Financial:

Banco de Chile (BCH), Banco Santander-Chile (BSAC), Berkshire Hathaway Inc (BRK - A), JPMorgan Chase & Co (JPM), Wells Fargo & Company (WFC), Bank of America Corporation (BAC), Visa Inc (V), Citigroup Inc (C), HSBC Holdings plc (HSBC), Mastercard Incorporated (MA).

Sector - Industrial Goods:

General Electric Company (GE), 3M Company (MMM), The Boeing Company (BA), Honeywell International Inc (HON), United Technologies Corporation (UTX), Lockheed Martin Corporation (LMT), Caterpillar Inc (CAT), General Dynamics Corporation (GD), Danaher Corporation (DHR), ABB Ltd (ABB).

Sector - Technology:

Alphabet Inc (GOOG), Microsoft Corporation (MSFT), Facebook, Inc (FB), AT&T Inc (T), China Mobile Limited (CHL), Oracle Corporation (ORCL), Taiwan Semiconductor Manufacturing Company Limited (TSM), Verizon Communications Inc (VZ), Intel Corporation (INTC), Cisco Systems, Inc (CSCO).

4. Proposed Methods

In this work methods for performing time series prediction on real-world time series stock market datasets will be examined. In this study, two different metrologies have been proposed, machine learning-based methodology and deep learning-based methodology. These two techniques are as follows:

4.1 Machine Learning

The proposed methods mainly consists of three steps: Loading all the datasets and pre-processing it for feature extraction, Applying Supervised Regression learning models on the training and testing dataset, Comparing the obtained results with the actual test data [Figure 3]. The prediction of a stock price can be done using various Regression methodologies that are available in Machine Learning and Deep Learning [Figure 4]. Some of the popular ML Algorithms like K-Nearest Neighbors, Support Vector Regressor, Bagging Techniques, Random Forest Regressor etc are implemented but Random Forest Regressor has outperformed for most of the companies that is for 79 companies because Random Forest creates multiple trees for a given dataset and only a subset of random features are considered for each tree so, that always only the strong predictor variable is not used for the first split. Each time different trees with different predictor variables are created and when we average them all the variance will be minimal.

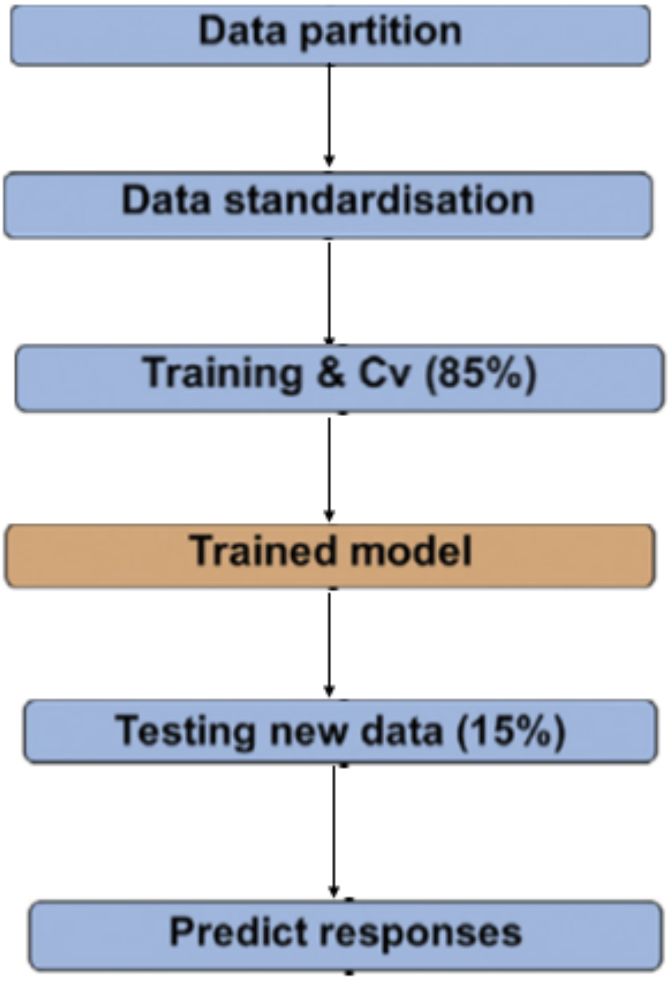


Figure 3: Workflow of a Supervised Learning Algorithm

4.2 Deep Learning

To our knowledge, the related studies do not have high accuracy in predicting the stock market price, we believe that by using ensemble learning [34] and combining the output of few different AI algorithms the result can be improved. Therefore, the finalised model (a combination of few algorithms) can be used for predicting the values for the test (unseen) dataset. If the result is satisfying and the mean absolute error between the predicted and actual values is a low number, so we can finalise the model with the chosen hyper parameters. But if it is not satisfying, we need to change and update the hyper parameters to get higher accuracy. It should be noted that in this case, we decided to use several different AI algorithms and then only pick the three algorithms with higher accuracies (Algorithm #1, #2 and #3 in Figure 5).

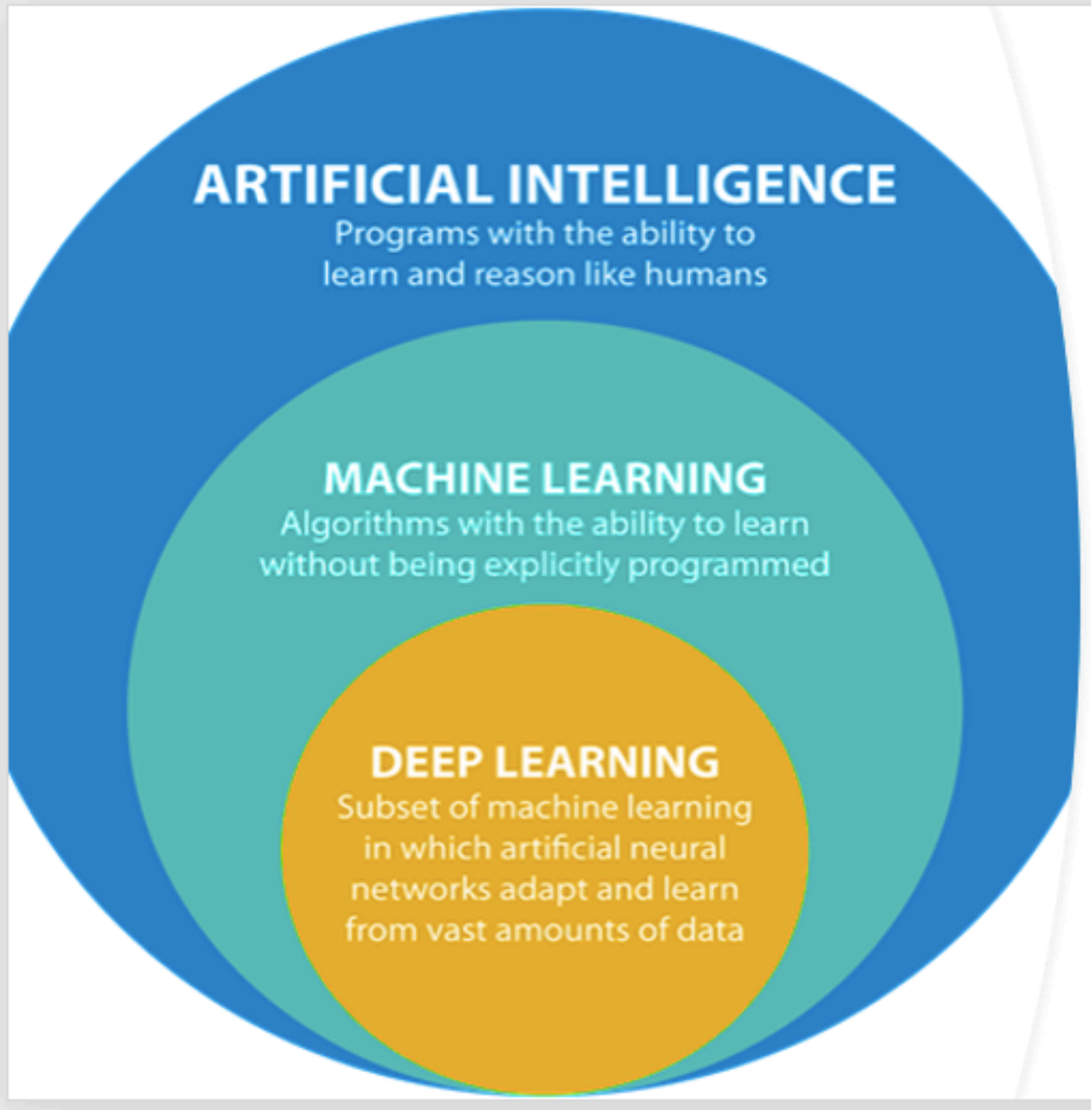

ARTIFICIAL INTELLIGENCE
Programs with the ability to learn and reason like humans
MACHINE LEARNING
Algorithms with the ability to learn without being explicitly programmed
DEEP LEARNING
Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data

Figure 4: Deep Learning (vs) Machine Learning

In the first step, data will be separated into three different groups, train, validation, and test data sets. While train and validation sets will be used to train the model, the test dataset will be used to check the performance of the trained model. In the second step, data should be normalised [35]. It is because the scales for different features are wildly different, this can have a knock-on effect on the learning process. In other words, ensuring standardised feature values implicitly weights all features equally in their representation. Data after pre-processing will be applied to the AI-based algorithms. As is mentioned earlier few different methods with different strategies (distance-based, angle-based, …) will be used and then the combination of the results of these methods will give us the final results (ensemble learning). Since this problem is a supervised classification and the supervisor exists, so the result will be checked using both validation and test datasets while training and testing processes, respectively.

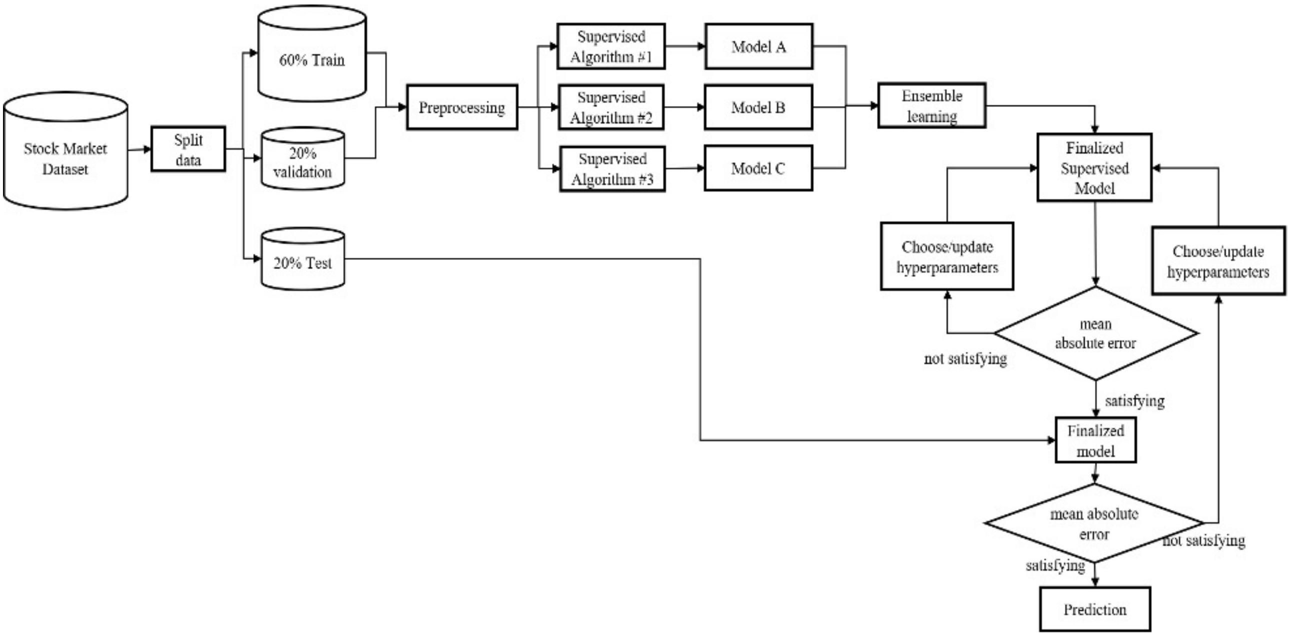


Figure 5: Flowchart for the Deep Learning based Proposed Technique

5. Results

5.1 Machine Learning

Machine Learning Algorithm like Random Forest Regressor has outperformed for 79 companies among 88 companies and the RMSE values ranges from 0 to 1 for those 79 companies. The RMSE values for test data on 79 best performing companies is shown in the Figure 6.

Nearly, for 8 companies the Random Forest Regressor didn't perform as expected and the RMSE values for those 8 companies range from 0 to 10 whereas for 1 company the algorithm has performed poorly and the RMSE value ranges from 750 to 830. The better and worst-performing companies are shown in Figure 7 and Figure 8. Overall the RMSE values for all the companies are shown in Table 2.

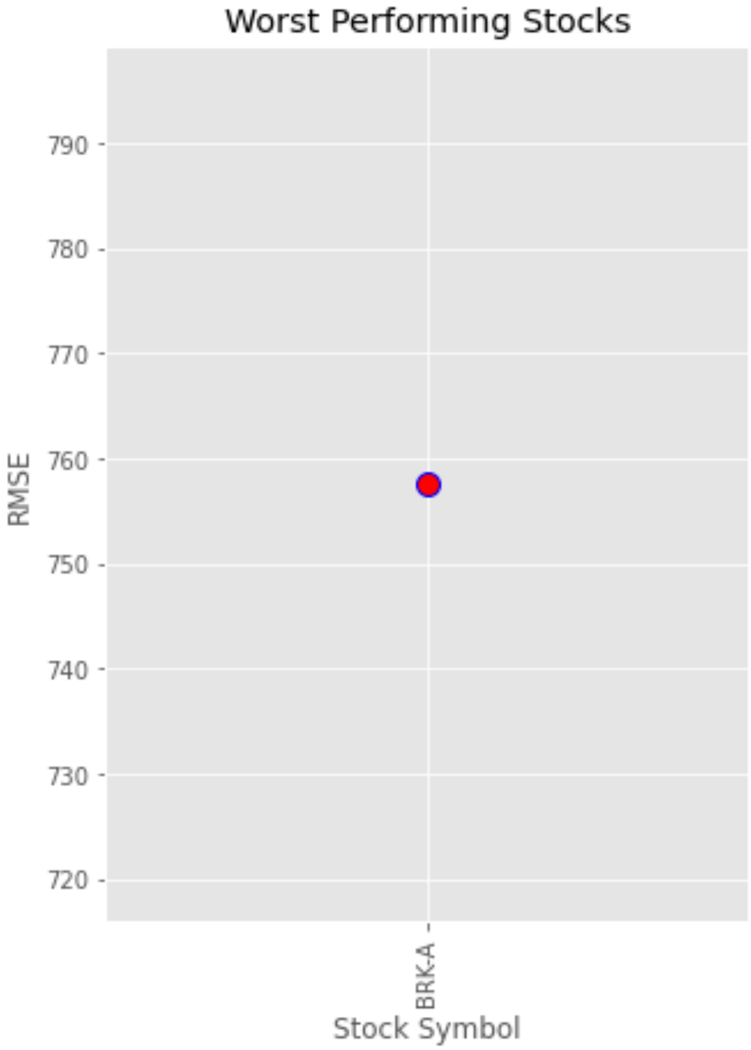


Figure 8: Worst Performing Companies

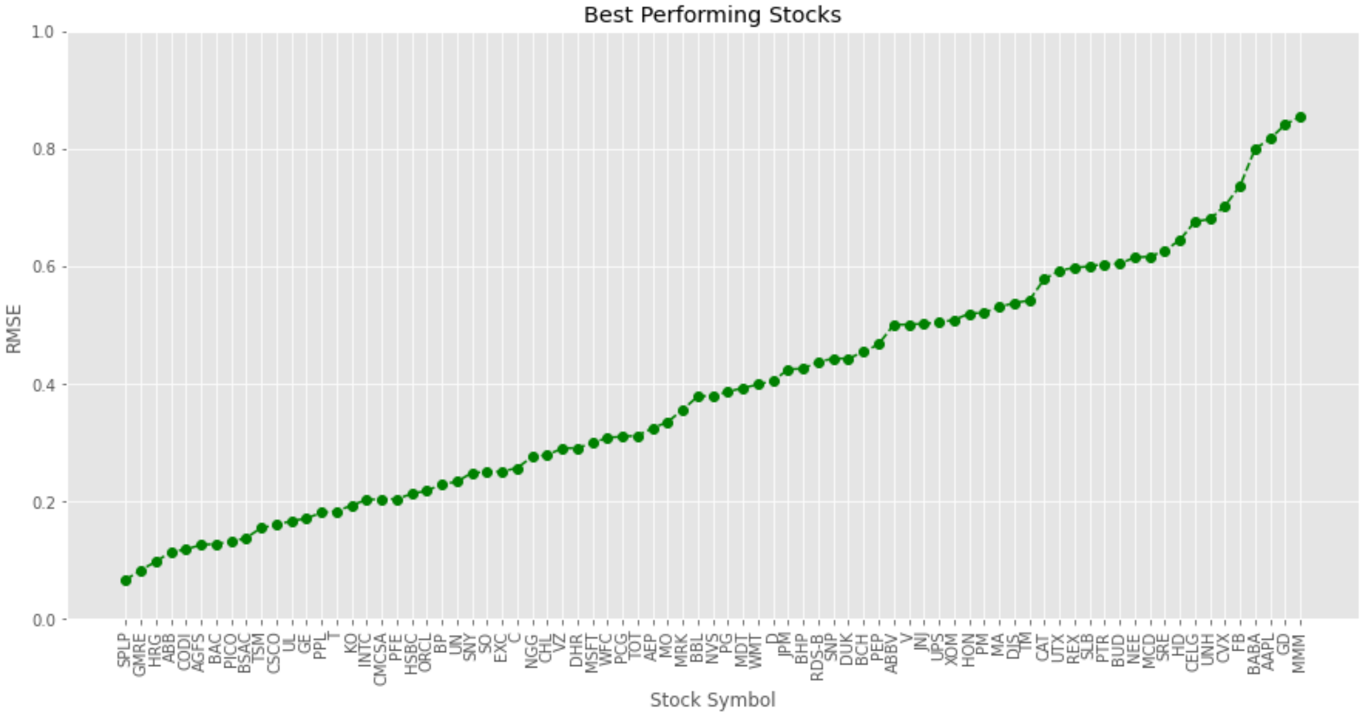


Figure 6: Best Performing Companies

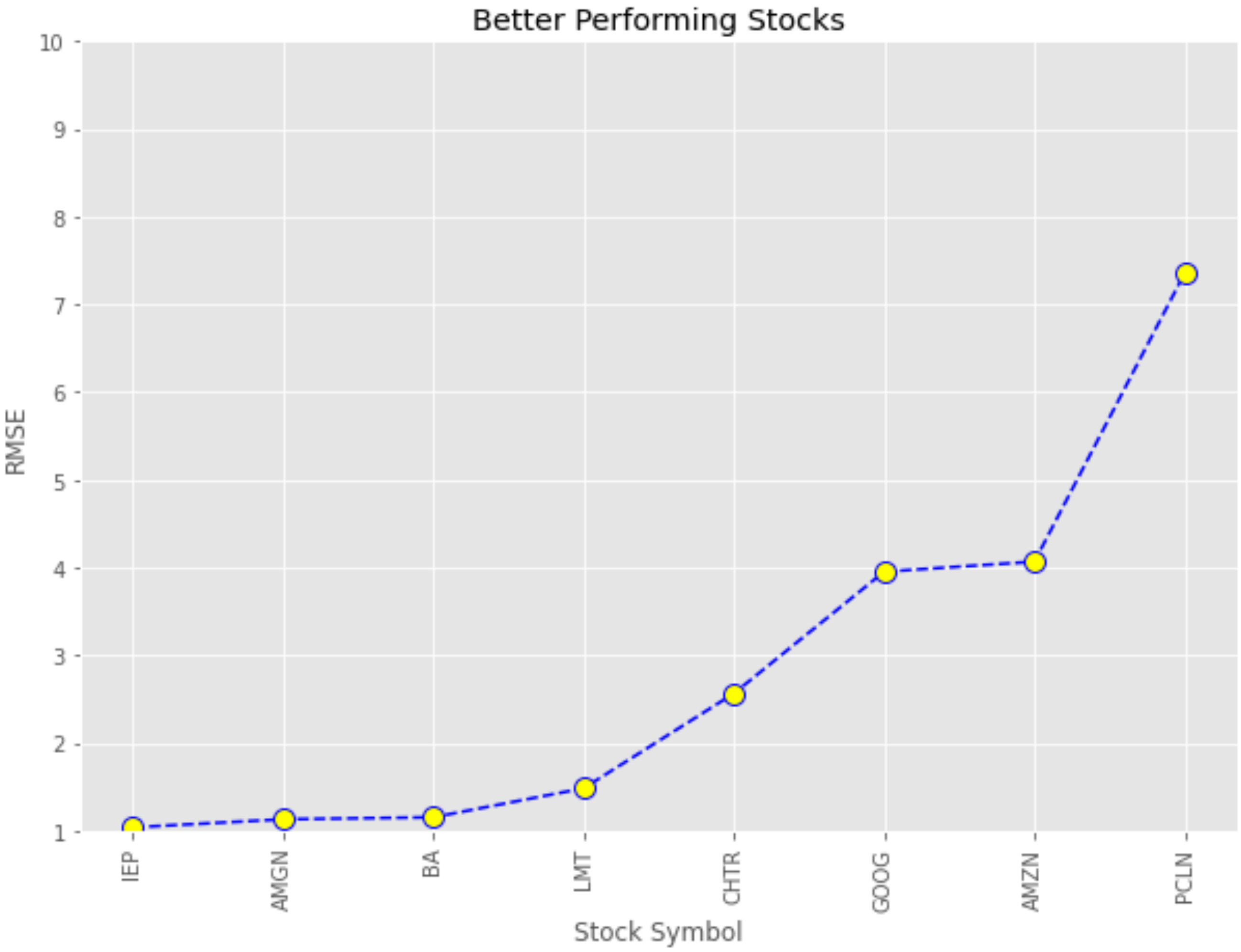


Figure 7: Better Performing Companies

RMSE values of 88 Companies

Companies	Range of RMSE Values
Best Performing Companies(79 stocks)	[0 - 1]
Better Performing Companies(8 stocks)	(1 - 10]
Worst Performing Companies(1 stock)	>10

5.2 Deep Learning

In this study, we have used the Apple company database. The goal is to predict the “close” price for each day. We started to train a model with 2 different deep learning-based approaches, Long short-term memory (LSTM) [40] and Gated recurrent units (GRU) [41]. As mentioned earlier. This is a supervised learning problem and we already know the actual prices.

After training the models and make the predictions, the root means square error (RMSE) can be calculated by the below equation :-

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

Where i, N,   and show variable index, number of non-missing data points, actual observations time series and estimated time series respectively. The below table [Table 3] shows the results for 10 epochs when using LSTM and GRU. This table shows that there is an inverse relationship between the number of epochs and the calculated loss.

Table 3: Loss for 10 epochs when using LSTM and GRU Algorithms

# of epochs	Loss when using LSTM	Loss when using GRU
1	0.0253	0.0135
2	5.7749E-04	3.4058E-04
3	3.4907E-04	2.9140E-04
4	2.9300E-04	2.7506E-04
5	2.8483E-04	2.6240E-04
6	2.8081E-04	2.5788E-04
7	2.8306E-04	2.5780E-04
8	2.7823E-04	2.6618E-04
9	2.8888E-04	2.6389E-04
10	2.8410E-04	2.6386E-04

In our case using 7 is enough to decrease the loss to a reasonable number. Below table shows the final RMSE and the processing time for LSTM and GRU algorithms when using 7 epochs.

Table 4: LSTM and GRU Algorithms

	RMSE for Train data	RMSE for Test data	# of neurons	Time taken to Process	# of epochs
LSTM	1.78	3.89	4	25s	7
GRU	1.88	2.72	4	10s	7

Table 4 shows better performance in terms of both RMSE and processing time compare to the LSTM.

As it is mentioned in the proposed technique (section 4.2), the goal is to use ensemble learning to improve the results. The below table shows the final results after using ensemble learning.

Table 5: RMSE values after using Ensemble Learning

	RMSE for Train data	RMSE for Test data
LSTM	1.78	3.89
GRU	1.88	2.72
Ensemble Learning	0.89	3.19

Table 5 shows that while RMSE for the training dataset is improved but for the test, it is not good compared to the GRU algorithm. That can be because of overfitting. Because overfitting happens when a model learns the detail and noise in the training data, and it decreases the error on the training dataset but increases the error on the test side. The Figure 9 shows the results when using ensemble learning. While the blue trend (shown in the background) shows the actual prices, the orange trend shows predictions on the training dataset and the green trend represents predicted values on the test dataset. The difference between the actual and predicted values give the error which shows how accurately the model is performing.

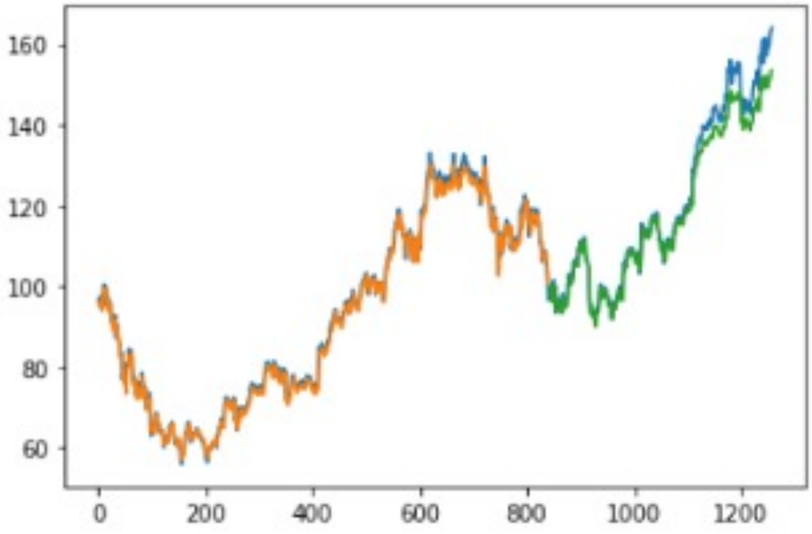


Figure 9: Price Prediction for Apple stock market using Ensemble Learning

From Section 5.1, it is clear that Machine Learning has failed to perform well for 10 companies i.e Better and Worst performing companies so, Deep Learning was implemented on those companies for making better predictions. The results for Better and Worst performing companies using Deep Learning are mentioned as follows, While the blue trend (shown in the background) shows the actual prices, the orange trend shows predictions on the training dataset and the green trend represents predicted values on the test dataset:-

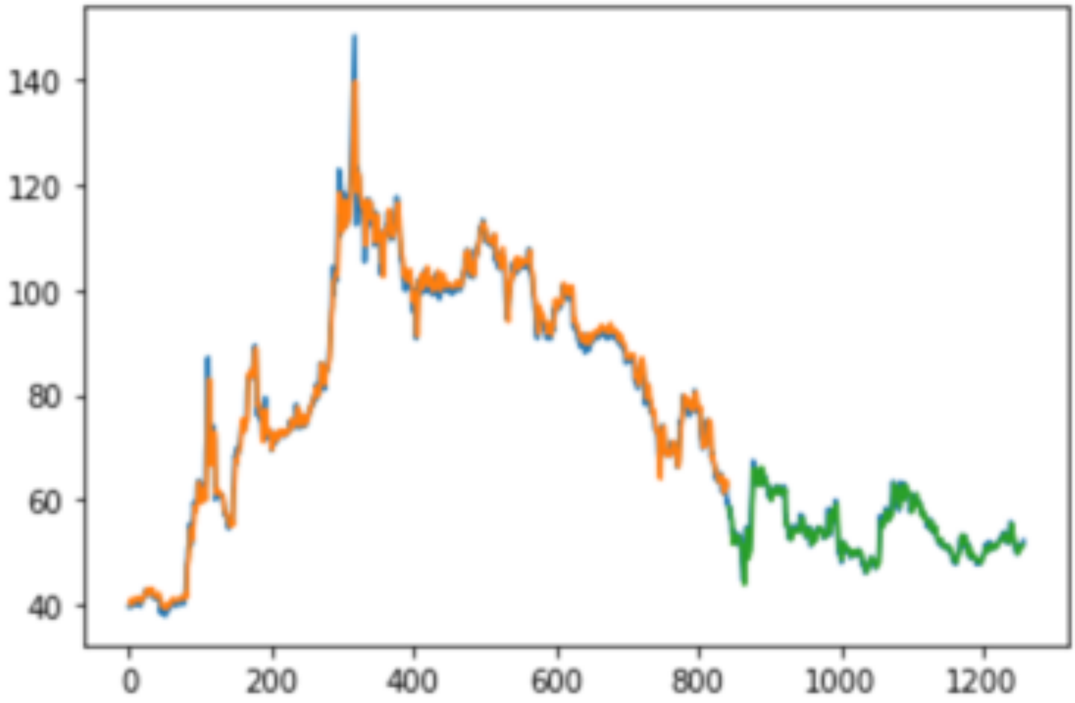


Figure 10: Price Prediction for Icahn Enterprises L.P (IEP)

Table 6: RMSE values for Icahn Enterprises L.P (IEP)

	RMSE for Train data	RMSE for Test data
LSTM	2.44	1.66
GRU	2.76	1.65
Ensemble Learning	1.38	1.40

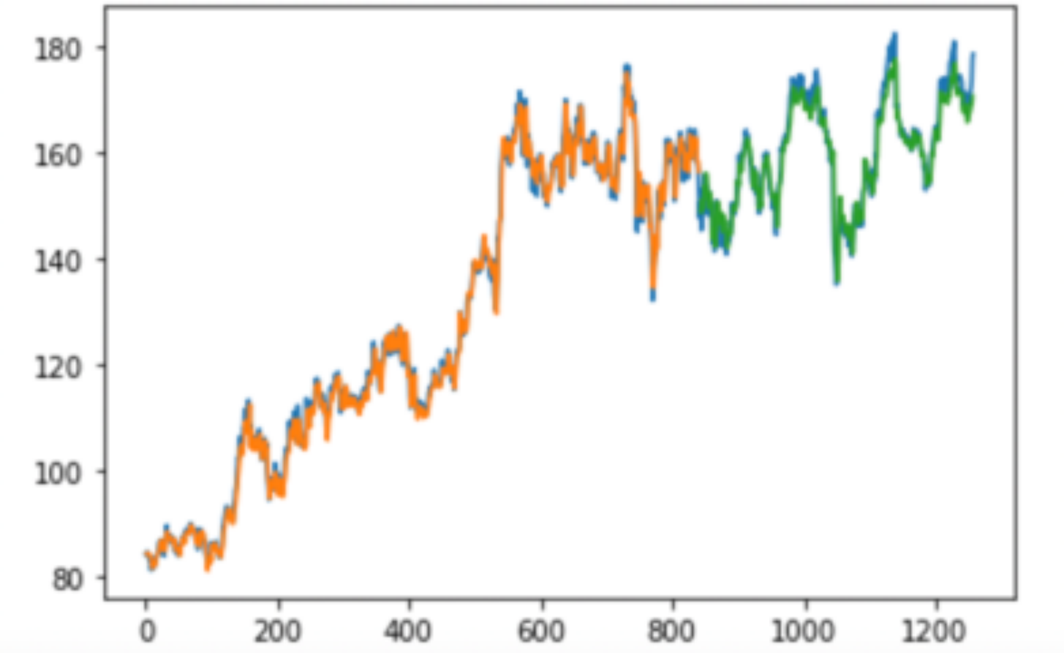


Figure 11: Price Prediction for Amgen Inc (AMGN)

Table 7: RMSE values for Amgen Inc (AMGN)

	RMSE for Train data	RMSE for Test data
LSTM	2.55	3.06
GRU	2.32	2.44
Ensemble Learning	1.16	2.67

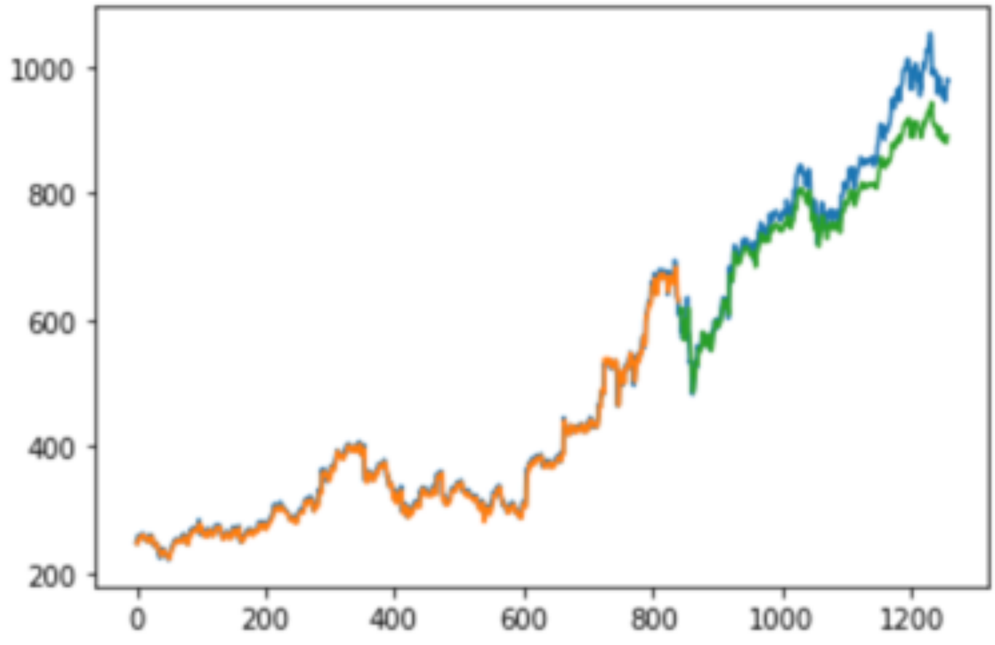


Figure 12: Price Prediction for Amazon, Inc (AMZN)

Table 8: RMSE values for Amazon, Inc (AMZN)

	RMSE for Train data	RMSE for Test data
LSTM	8.33	47.24
GRU	8.22	42.58
Ensemble Learning	4.11	44.88

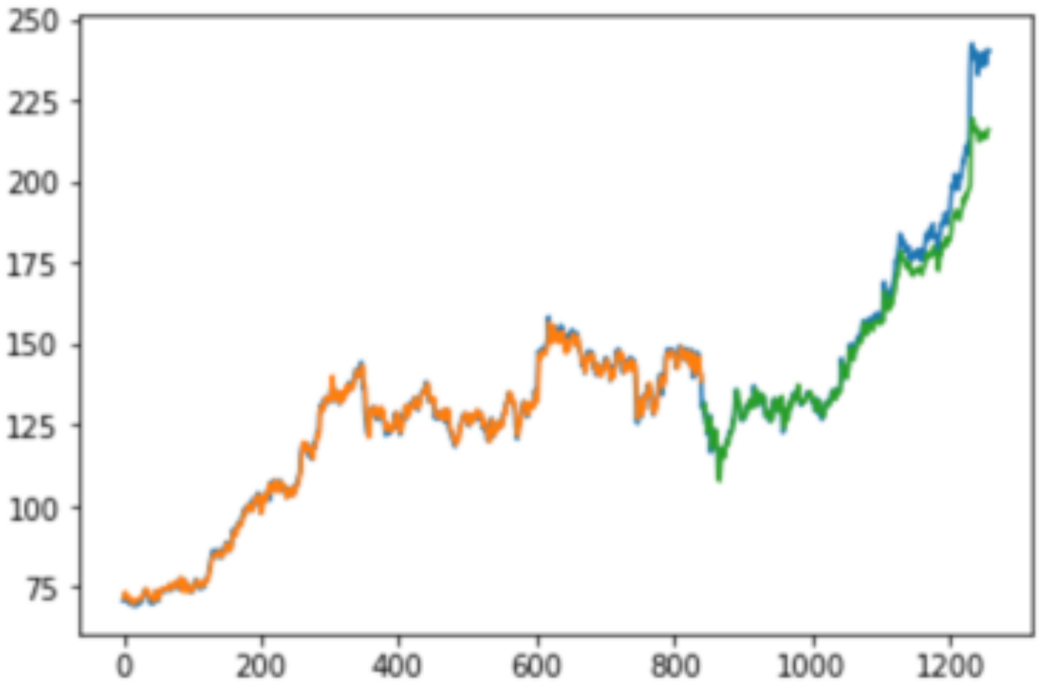


Figure 13: Price Prediction for The Boeing Company (BA)

Table 9: RMSE values for The Boeing Company (BA)

	RMSE for Train data	RMSE for Test data
LSTM	1.90	9.64
GRU	1.76	5.47
Ensemble Learning	0.88	7.52

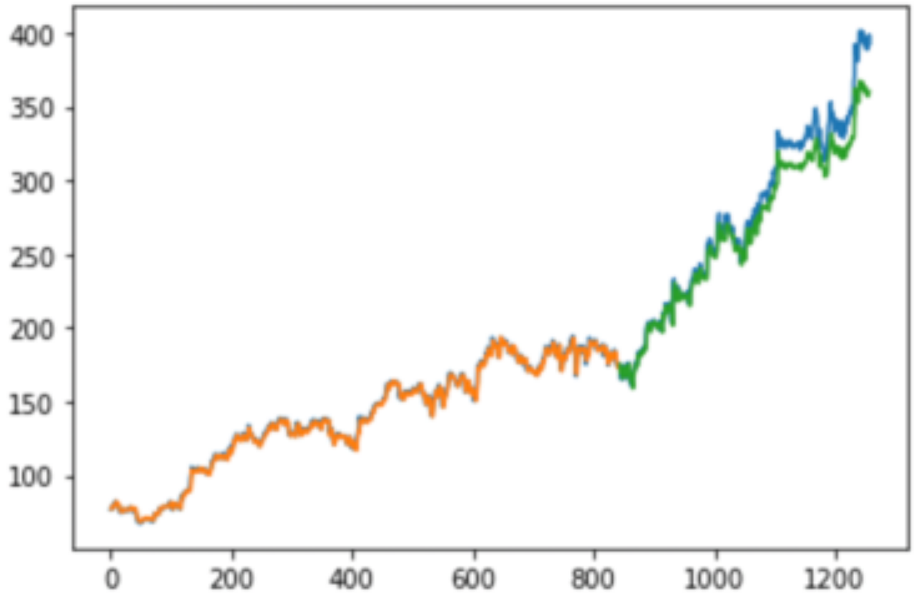


Figure 14: Price Prediction for Charter Communications, Inc (CHTR)

Table 10: RMSE values for Charter Communications, Inc (CHTR)

	RMSE for Train data	RMSE for Test data
LSTM	2.65	22.41
GRU	2.52	5.77
Ensemble Learning	1.26	13.36

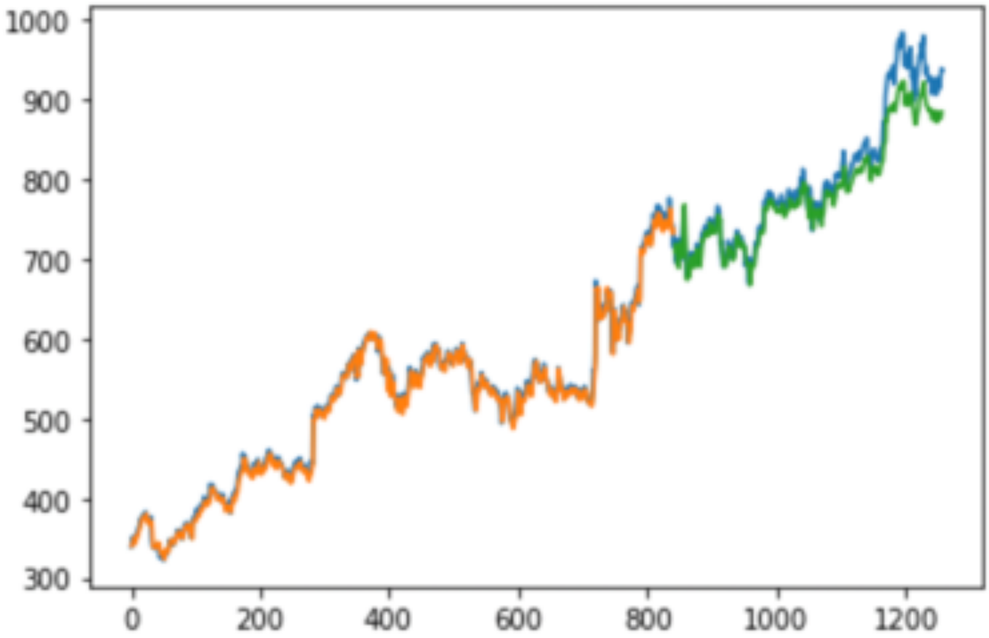


Figure 15: Price Prediction for Alphabet Inc (GOOG)

Table 11: RMSE values for Alphabet Inc (GOOG)

	RMSE for Train data	RMSE for Test data
LSTM	10.59	23.66
GRU	8.94	26.72
Ensemble Learning	4.47	25.17

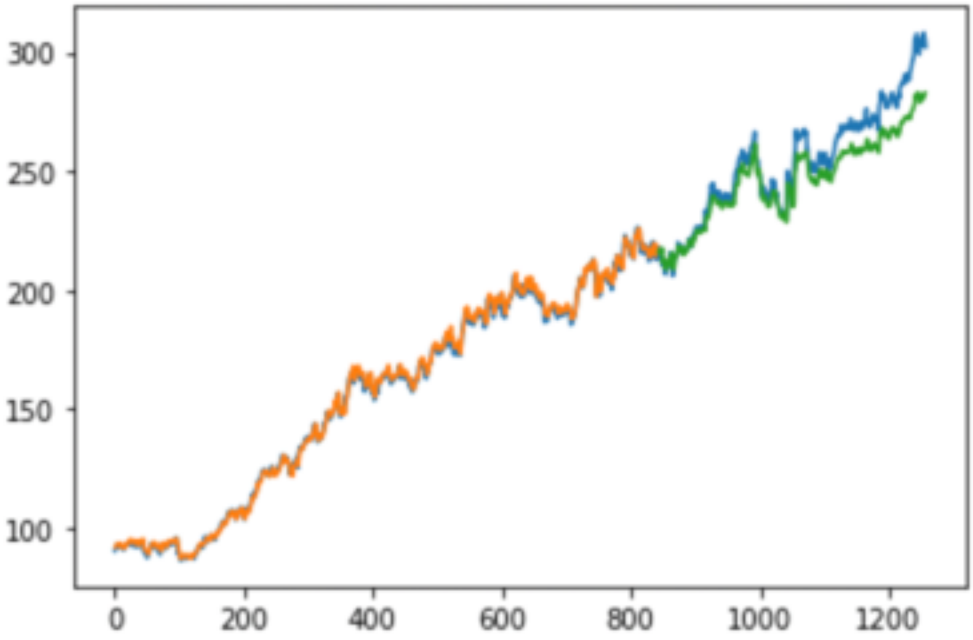


Figure 16: Price Prediction for Lockheed Martin Corporation (LMT)

Table 12: RMSE values for Lockheed Martin Corporation (LMT)

	RMSE for Train data	RMSE for Test data
LSTM	2.17	11.98
GRU	1.96	7.12
Ensemble Learning	0.98	9.52

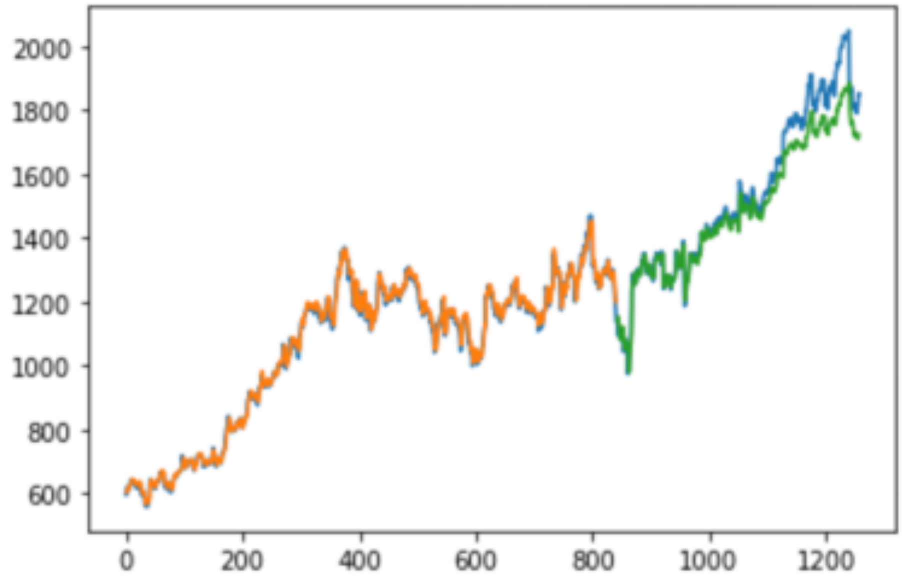


Figure 17: Price Prediction for The Priceline Group Inc (PCLN)

Table 13: RMSE values for The Priceline Group, Inc (PCLN)

	RMSE for Train data	RMSE for Test data
LSTM	20.56	86.97
GRU	20.86	44.04
Ensemble Learning	10.43	64.70

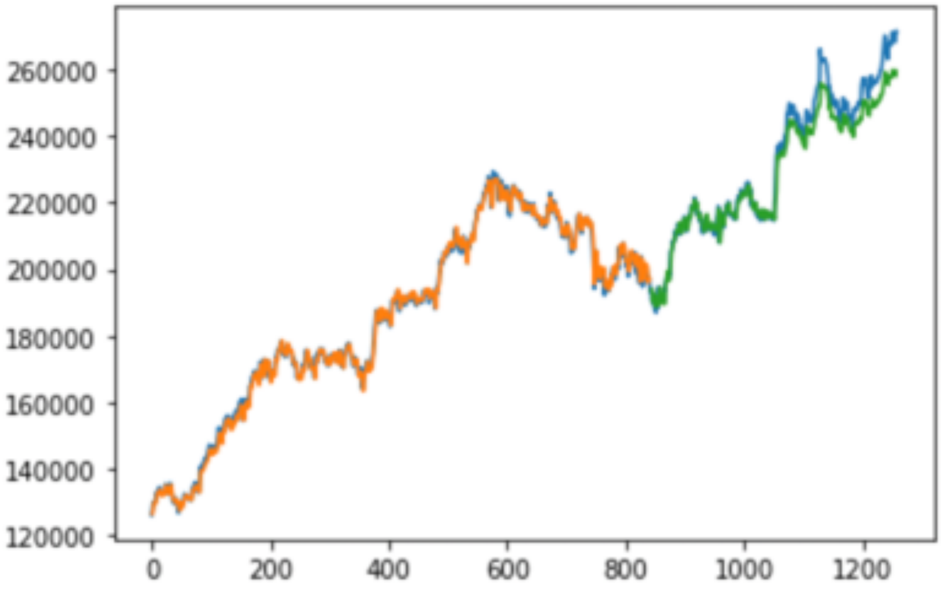


Figure 18: Price Prediction for Berkshire Hathaway Inc. (BRK-A)

Table 14: RMSE values for Berkshire Hathaway Inc. (BRK-A)

	RMSE for Train data	RMSE for Test data
LSTM	1.05	4.43
GRU	1.53	1.44
Ensemble Learning	0.52	1.76

From the above Deep Learning Results it is clear that the results obtained from Better performing Companies [Figure 7] using Machine Learning are much better than the results obtained using Deep Learning Techniques. Whereas, Deep Learning has outperformed the company BRK-A when compared with Machine Learning results.

6. Discussion

As mentioned in Section 5.1, after applying Random Forest Regressor for all the companies only a few companies didn't perform well and one company i.e

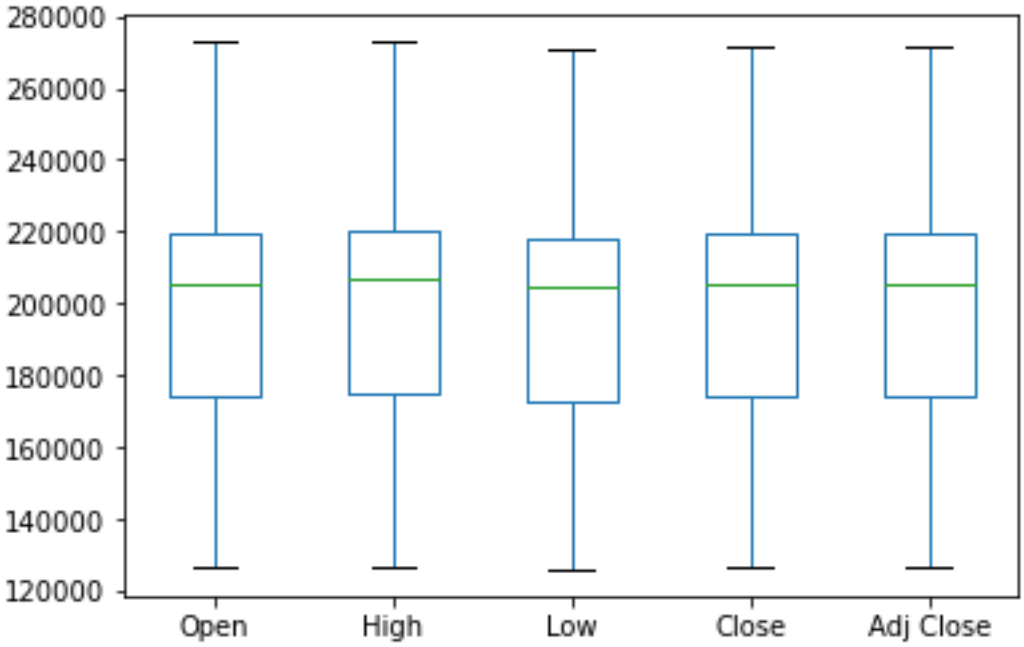


Figure 19: Box Plot of Berkshire Hathaway Inc

Berkshire Hathaway Inc (BRK - A) has performed very poorly because the minimum and maximum price of one stock is from 170,000 US dollars to 220,000 US dollars [Figure 19].

Hence, Random Forest Regressor didn't perform well for large stock price values so, other boosting algorithms like Ada boost and XG boost are implemented on the Berkshire Hathaway Inc (BRK-A) dataset. Initially, we tried using Ada boost, KNN and Support Vector Regressor, Logistic Regressor on the dataset by implementing different values to the hyper-parameters on the dataset but they failed to predict the output with an optimal value. Finally, we tried using XG Boost on the dataset, the RMSE value was around 300 which is almost 60% less than the value obtained using Random Forest Regressor but the model is getting overfitted i.e the RMSE value for train data is only 75 whereas for test data it is 300 even after performing hyper-parameter tuning so, Neural Networks has to be implemented on the companies with high stock price values. Also, Linear Regression won't work well for Stock Market Prediction because the prices in Stock Market are dynamic in nature so, fitting a linear line to the input data won't make optimal predictions.

The better-performing companies like Icahn Enterprises L.P (IEP), Amgen Inc (AMGN), The Boeing Company (BA), Lockheed Martin Corporation (LMT), Charter Communications, Inc (CHTR), Alphabet Inc (GOOG), Amazon, Inc (AMZN) and The Priceline Group, Inc (PCLN) didn't perform well as expected using Random Forest Regressor so, boosting algorithms like Ada boost and XG boost are implemented on these datasets. Initially, we tried using Ada boost on these datasets by implementing different values to the hyper-parameters on the datasets but the algorithm has failed to predict the output with an optimal value. Finally, we tried using XG boost on the datasets, the RMSE values were low than

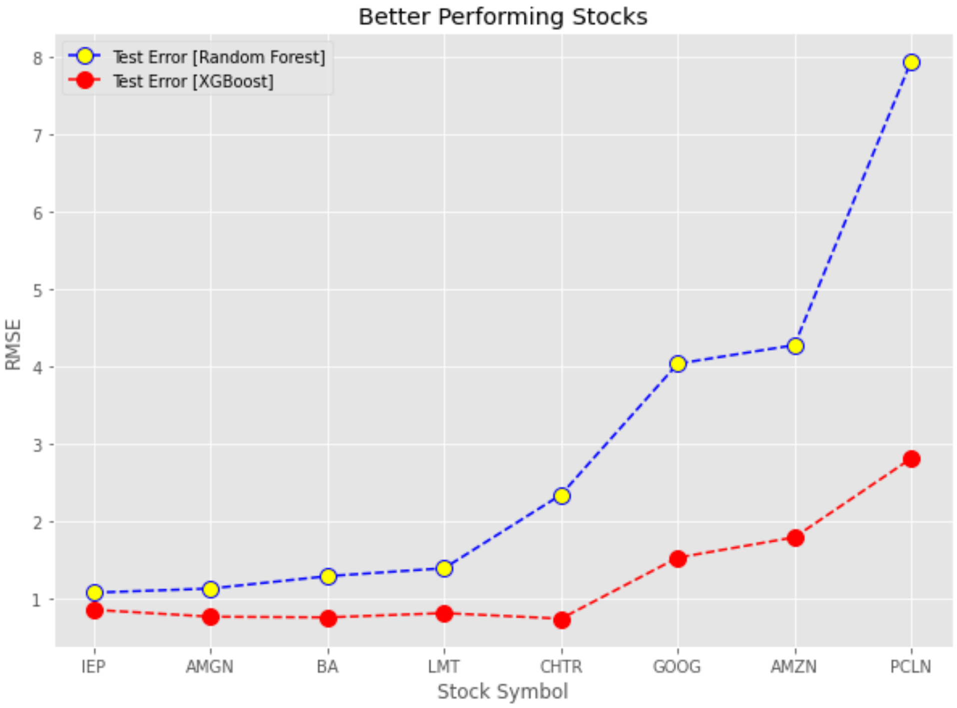


Figure 20: Random Forest (vs) XG Boost

the previous algorithm as shown in Figure 20 but the model is slightly getting overfitted even after tuning the hyper-parameters as shown in Figure 21.

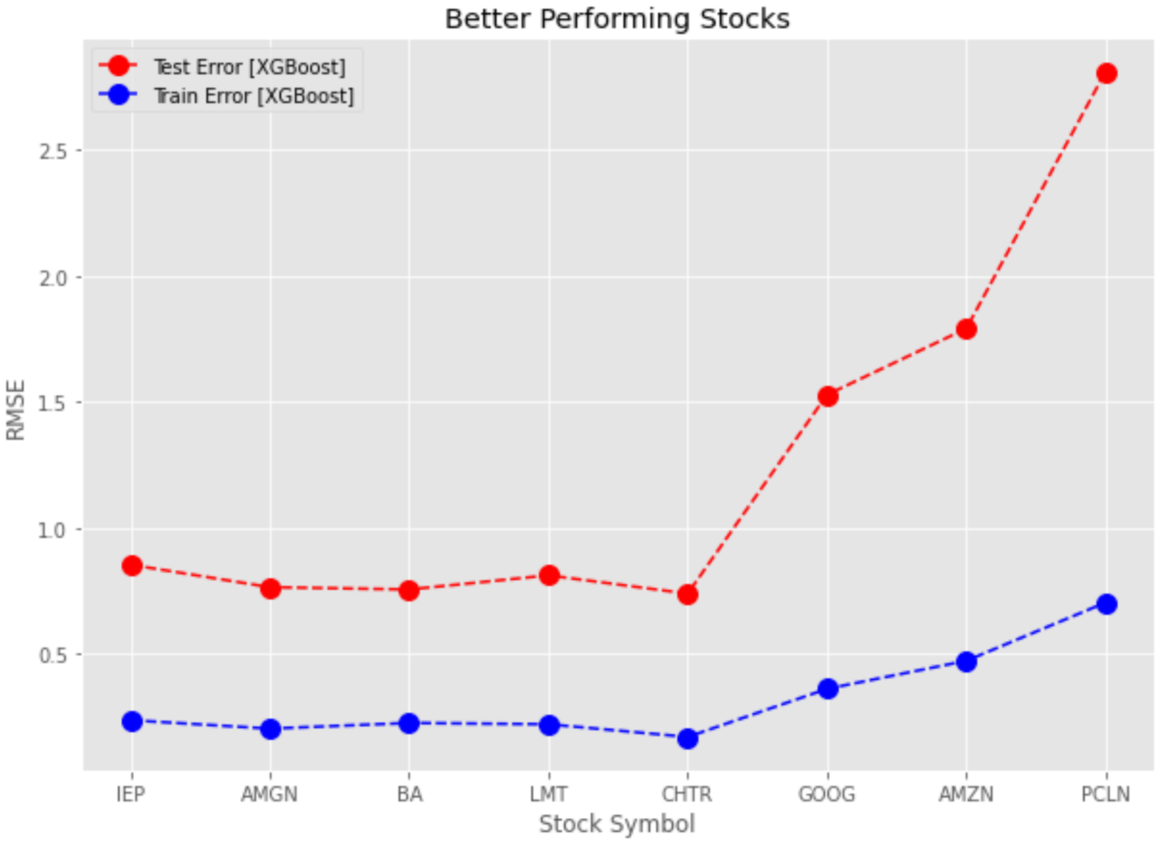


Figure 21: Test Error (vs) Train Error using XG Boost

Better performing Companies like The Boeing Company (BA), Charter Communications, Inc (CHTR) and The Priceline Group, Inc (PCLN) had many outliers in the dataset whereas other Better performing Companies didn’t had outliers however only Icahn Enterprises L.P (IEP) company had lot of fluctuations in the Stock Market. Box Plots for the companies BA, CHTR and PCLN are shown in Figures 22, 23 & 24. Thereby, Neural Networks has to be implemented on these better-performing companies too.

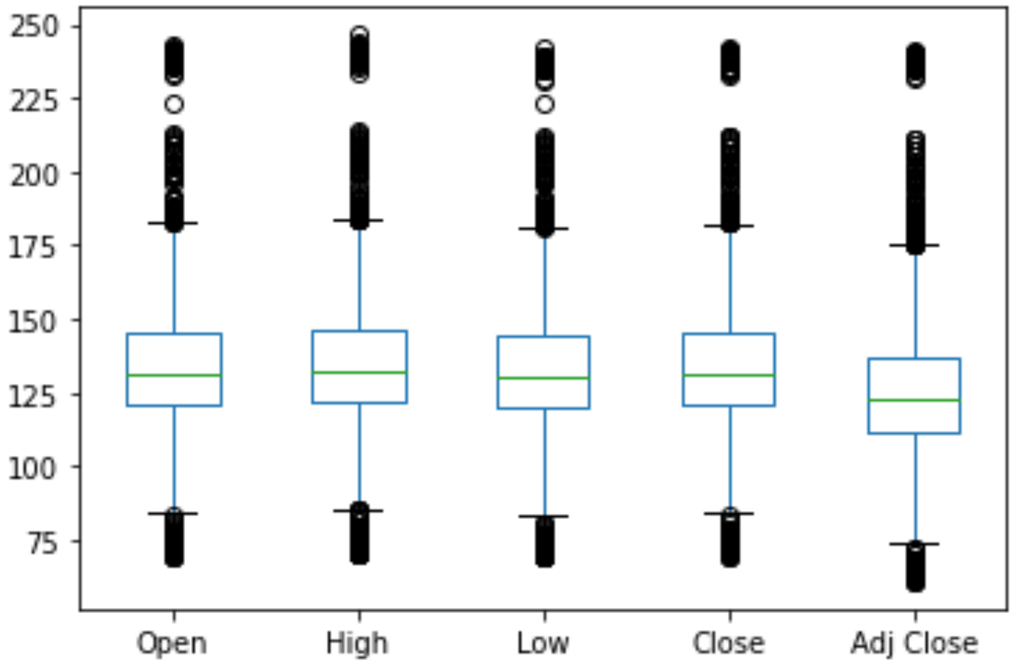


Figure 22: Box Plot of The Boeing Company

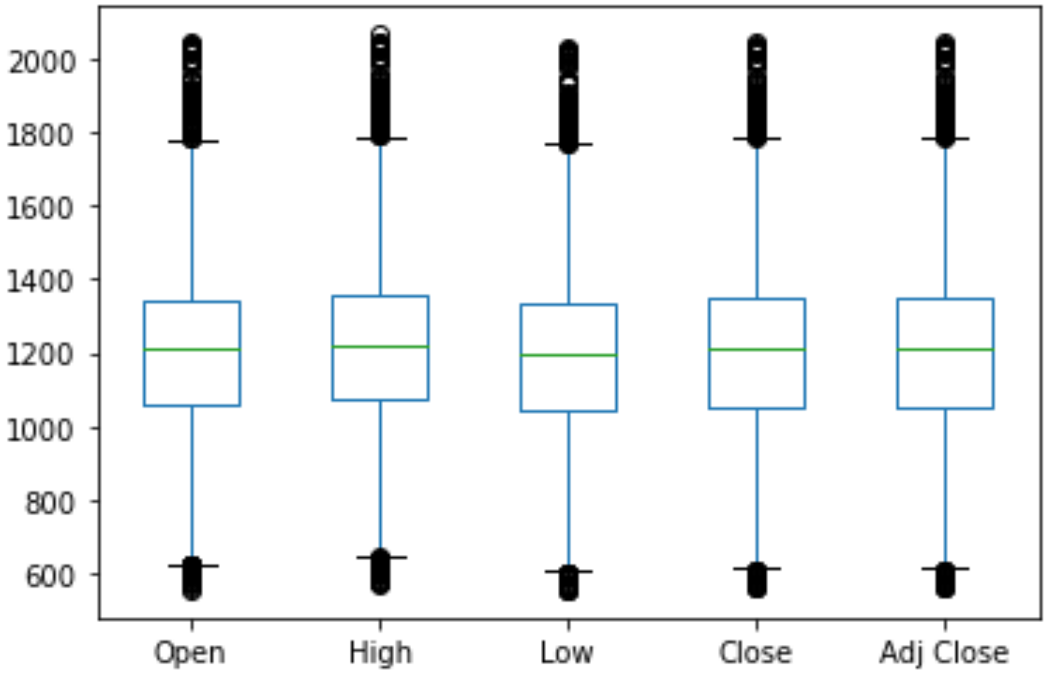


Figure 23: Box Plot of The Priceline Group, Inc

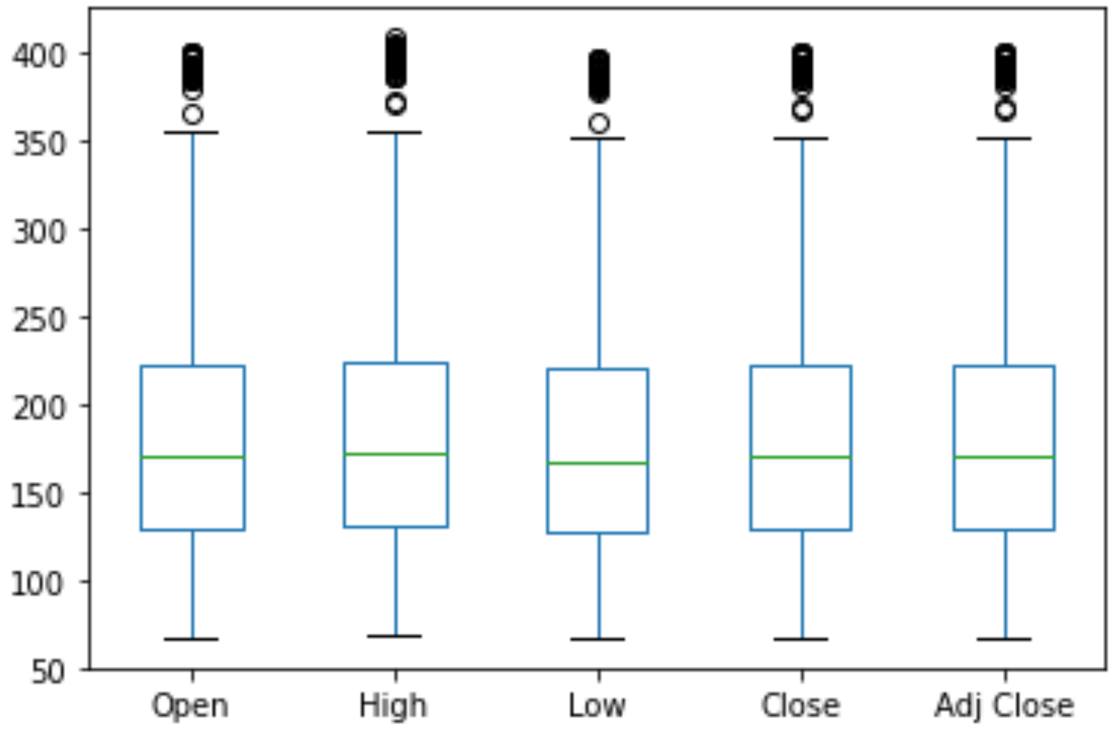


Figure 24: Box Plot of Charter Communications, Inc

7. Conclusion

In this study, Artificial intelligence is used to make predictions about stock market prices. A stock market is a place where buying and selling of shares happen for companies. The data that is used in this work is a time Series dataset and consists of stock prices of 88 different companies as described in Section 3. However, while the time component adds additional information, it also makes time series problems more difficult to handle compared to many other prediction tasks. In this study, we proposed two methods, Machine Learning-based, and Deep Learning-based. As the proposed methods show in this study different AI-based algorithms together with ensemble learning are used to make the predictions and make a comparison between the results of different methodologies. It has been shown that GRU performs better than Deep Learning based methods in terms of both accuracy and processing time. Also, Machine Learning-based methods perform pretty well for most of the companies but it fails when it comes to large stock price values so, Deep Learning methods were implemented on the stocks with high price values and the results were far better using Deep Learning Methods. In future, different AI algorithms can be implemented to further decrease the error value.

References

1. Klaus Adam, Albert Marcet, and Juan Pablo Nicoli. 2016. Stock market volatility and learning. The Journal of Finance, 71(1):33–82.
2. Kazuhiro Kohara, Tsutomu Ishikawa, Yoshimi Fukuhara, and Yukihiro Nakamura. 1997. Stock price prediction using prior knowledge and neural networks. Intelligent Systems in Accounting, Finance & Management, 6(1):11–22.
3. A. A. Ariyo, A. O. Adewumi and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, 2014, pp. 106-112, doi: 10.1109/UKSim.2014.67.
4. I. Kumar, K. Dogra, C. Utreja and P. Yadav, "A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1003-1007, doi: 10.1109/ICICCT.2018.8473214.
5. Available online at https://github.com/yumoxu/stocknet-dataset/tree/master/price .
6. R. C. Cavalcante, R. C. Brasileiro, V. L. Souza, J. P. Nobrega, and A. L. Oliveira, “Computational intelligence and financial markets: A survey and future directions,” Expert Systems with Applications, vol. 55, pp. 194–211, 2016.
7. M. Lam, “Neural network techniques for financial performance prediction: integrating fundamental and technical analysis,” Decision Support Systems, vol. 37, no. 4, pp. 567–581, 2004.
8. A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, “Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment,” Expert Systems with Applications, vol. 42, no. 1, pp. 306–324, 2015.
9. A. Rodr´ıguez-Gonza´lez, A´ . Garc´ıa-Crespo, R. Colomo-Palacios, F. G. Iglesias, and J. M. G´omez-Berb´ıs, “Cast: Using neural networks to improve trading systems based on technical analysis by means of the rsi financial indicator,” Expert Systems with Applications, vol. 38, no. 9, pp. 11 489–11 500, 2011.
10. E. F. Fama, “The behavior of stock-market prices,” The journal of Business, vol. 38, no. 1, pp. 34–105, 1965.

11. B. Qian, K. Rasheed, Stock market prediction with multiple classifiers, Applied Intelligence 26 (1) (2007) 25–33.
12. W. Huang, Y. Nakamori, S.-Y. Wang, Forecasting stock market movement direction with support vector machine, Computers & Operations Research 32 (10) (2005) 2513–2522.
13. X. Lin, Z. Yang, Y. Song, Short-term stock price prediction based on echo state networks, Expert Systems with Applications 36 (3) (2009) 7313–7317.
14. P. Khuwaja, S.A. Khowaja, I. Khoso, I.A. Lashari, Prediction of stock movement using phase space reconstruction and extreme learning machines, Journal of Experimental and Theoretical Artificial Intelligence 32 (1) (2020) 59–79.
15. R.K. Nayak, D. Mishra, A.K. Rath, A naïve svm-knn based stock market trend reversal analysis for indian benchmark indices, Applied Soft Computing 35 (2015) 670–680.
16. A.A. Adebiyi, A.O. Adewumi, C.K. Ayo, Comparison of arima and artificial neural networks models for stock price prediction, Journal of Applied Mathematics 2014 (2014) 1–7.
17. M. Göçken, M. Özçalıcı, A. Boru, A.T. Dosdog˘ru, Integrating metaheuristics and artificial neural networks for improved stock price prediction, Expert Systems with Applications 44 (2016) 320–331.
18. K.-J. Kim, H. Ahn, Simultaneous optimization of artificial neural networks for financial forecasting, Applied Intelligence 36 (4) (2012) 887–898.
19. J.L. Ticknor, A bayesian regularized artificial neural network for stock market forecasting, Expert Systems with Applications 40 (14) (2013) 5501–5506.
20. J. Patel, S. Shah, P. Thakkar, K. Kotecha, Predicting stock market index using fusion of machine learning techniques, Expert Systems with Applications 42 (4) (2015) 2162–2172.
21. H. Jang, J. Lee, An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information, IEEE Access 6 (2017) 5427–5437.
22. L. Zhang, C. Aggarwal, G.-J. Qi, Stock price prediction via discovering multifrequency trading patterns, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13–17, 2017, ACM, 2017, pp. 2141–2149.
23. R. Akita, A. Yoshihara, T. Matsubara, K. Uehara, Deep learning for stock prediction using numerical and textual information, in: 15th IEEE/ACIS

International Conference on Computer and Information Science, ICIS 2016, Okayama, Japan, June 26–29, 2016, IEEE, 2016, pp. 1–6.

24. J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, X. Deng, Exploiting topic based twitter sentiment for stock prediction, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4–9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers, vol. 2, 2013, pp. 24–29.
25. S.M. Idrees, M.A. Alam, P. Agarwal, A prediction approach for stock market volatility based on time series data, IEEE Access 7 (2019) 17287–17298. on Economics and Natural Language Processing, Association for Computational Linguistics, Hong Kong, 2019, pp. 31–40.
26. Y.-H. Lui and D. Mole, “The use of fundamental and technical analyses by foreign exchange dealers: Hong kong evidence,” Journal of International Money and Finance, vol. 17, no. 3, pp. 535–545, 1998.
27. X. Ding, Y. Zhang, T. Liu, J. Duan, Knowledge-driven event embedding for stock prediction, in: COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11–16, 2016, Osaka, Japan, ACL, 2016, pp. 2133–2142.
28. D. Chen, Y. Zou, K. Harimoto, R. Bao, X. Ren, X. Sun, Incorporating fine-grained events in stock movement prediction, in: Proceedings of the Second Workshop on Economics and Natural Language Processing, Association for Computational Linguistics, Hong Kong, 2019, pp. 31–40.
29. W. Long, Z. Lu, L. Cui, Deep learning-based feature engineering for stock price movement prediction, Knowledge-Based System 164 (2019) 163–173.
30. F. Feng, H. Chen, X. He, J. Ding, M. Sun, T. Chua, Enhancing stock movement prediction with adversarial training, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10–16, 2019, IJCAI, 2019, pp. 5843–5849.
31. D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473.
32. Z. Hu, W. Liu, J. Bian, X. Liu, T.-Y. Liu, Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5–9, 2018, ACM, 2018, pp. 261–269.

33. Xu, S.B. Cohen, Stock movement prediction from tweets and historical prices, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 1: Long Papers, 2018, pp. 1970–1979.
34. Dietterich, Thomas G. "Ensemble learning." The handbook of brain theory and neural networks 2 (2002): 110-125.
35. Sola, Jorge, and Joaquin Sevilla. "Importance of input data normalization for the application of neural networks to complex industrial problems." IEEE Transactions on nuclear science 44.3 (1997): 1464-1468.
36. Malkiel BG (2007) A random walk downWall Street: the time-tested strategy for successful investing.WW Norton & Company.
37. Patel J, Shah S, Thakkar P, Kotecha K (2015) Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Syst Appl 42:259–268.
38. Guresen E, Kayakutlu G, Daim TU (2011) Using artificial neural network models in stock market index prediction. Expert Syst Appl 38:10389–10397.
39. Khan HZ, Alin ST, Hussain A (2011) Price prediction of share market using artificial neural network “ANN”. Int J Comput Appl 22(2):42–47.
40. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.
41. Chung, Junyoung, et al. “Empirical evaluation of gated recurrent neural networks on sequence modelling.” arXiv preprint arXiv:1412.3555(2014).