Comparison of ARIMA, LSTM and CNN for Stock Market Prediction

ALI AKBAR REHMAN and AZIZ ZAFAR, Universitet i Stavanger, Norway

Any company that is listed on a stock market means it is a public company and people can buy and sell shares of the said company in the stock market. These shares represent a small ownership or a stake in the company. Companies mostly make a portion of their stocks public and owners retain a majority of shares to have a say in the companies direction. So stock markets in addition to be a market where people could invest in companies also is a strong indicator of the state of the economy. Essentially a Stock Market is an essential part of today's global economy, and so analysing stock market trends and seasonality and predicting future prices can present itself to be extremely useful. Prediction of future trends in a stock market is an extremely challenging task since there are so many factors that cannot be modelled such as sentiment of the populace, populace irrational behaviour in times of crises, global news and so on. Which makes is very difficult to predict stock market prices. In this project we will look at Auto-regressive integrated moving average (ARIMA) statistical model, Long Short-Term Memory (LSTM) and Artificial Neural Networks (ANNs) to predict future pricing and compare and discuss their results. One thing to note here is that these models only take historic prices into account and not external factors or noise when modelling the stock market trends and seasonalities.

ACM Reference Format:
Ali Akbar Rehman and Aziz Zafar. 2023. Comparison of ARIMA, LSTM and CNN for Stock Market Prediction. 1, 1 (April 2023), 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Forecasting future values can be invaluable in many fields. Predicting future stock prices can not only be extremely beneficial to individuals to invest wisely but can be valuable for predicting the direction of the economy. That is why stock market analysis and prediction is such a popular topic. On the flip side modelling stock prices and forecasting future values is an immensely difficult task at hand and many financial experts believe it to be an impossible task. This difficulty in modelling stock prices is because of extreme volatility of the stock market and a lot of external factors that cannot be modelled accurately play a huge role in the trends of the stock market. Even still with the advances in machine learning a lot of research has been put into modelling stock market prices. The most common models normally used to model such time series data are Auto-regressive Integrated Moving Average (ARIMA), Artificial Neural Networks (ANNs) and Long Short-Term Memory (LSTM) with a relatively higher accuracy.

In this project we will study the theoretical background of these models and test their forecasting on a few of the publicly listed companies on the stock market. We will use the three previously mentioned models to forecast the values for these stocks and compare & discuss the results. One important thing to note here is that all these models don't take into account external factors and consider them noise and they only focus on the historic values of a companies stock. We gather stock data using Yahoo Finance API for the past 10 years and test the model performance on the data.

ARIMA is a statistical model that can be used to analyse any time series data and forecast future values. The ARIMA model forecasts are based on the input time series data. ARIMA is particularly powerful for short term forecasting here we use the ARIMA model to forecast long term values and see how it performs in comparison to other complex models. Recently ANNs have seen numerous uses in every field of science particularly because of their ability to model complex problems and detect patterns in them. Unlike ARIMA, ANNs try to learn the underlying model from the original input data to make observations and then use that model to forecast values. LSTM is a variant of recurrent neural network. Unlike other methods, its feedback connection makes it easier to find development trends through the back propagation of current historical prices and current prices [Ma 2020]

2 BACKGROUND

Because of the randomness displayed by stock market prices, experts divide the common strategies towards stock market investment into 3 main categories.

- Fundamental Analysis
- Technical Analysis
- Random Walk Theory

Fundamental analysis takes into account economic conditions of the market as a whole, companies prior history and values to forecast the stocks value and invest accordingly. Technical analysis only looks at the historical prices of a stock and try to infer trends from that. And Random Walk theory suggests that the stock market fluctuates independent of prior stock values and is impossible to predict. On the surface of it, stock market tends to follow the random walk theory but fundamental analysis is the most viable approach towards stock market prediction but since fundamental analysis consists of factors that are harder to model machine learning follows technical analysis approach to look at historical prices to predict future stock values.

Machine learning takes input data and learns the trends and patterns in the data to automatically develop a complex model that can then be used to perform forecasting of future unseen data. The stage of allowing the algorithm to learn patterns on input data and creating a complex model is called training. The machine learning models can be trained to perform regression as well as classification problems. After training the resultant models can then be used to predict unseen data and the difference between predicted and actual values on the unseen data can be used to calculate accuracy of the

Authors' address: Ali Akbar Rehman, aa.rehman@stud.uis.no; Aziz Zafar, az.zafar@stud.uis.no, Universitet i Stavanger, Stavanger, Rogaland, Norway, 4021.

XXXX-XXXX/2023/4-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

model. Next up we discuss the 3 models / algorithms we will use for stock market prediction.

2.1 Autoregressive Integrated Moving Average (ARIMA)

The ARIMA model is a model to analyse and predict time series data. ARIMA model can be used for short-term forecasting based on historic data because it only incorporates historic input and error from the time series. ARIMA model is applicable only to the kind of time series where the time series data is stationary mean wise as well as variance wise. What is stationarity and why is it important in the analysis of time series data? In simple words stationarity means that the statistical properties of the process that outputs the time series data does not change as we progress further in the time series. Or that as the time series changes the way it changes stays consistent. Noise presents a problem in the stationarity of the time series normally.

So following the above hypothesis we want the properties like mean and variance to stay consistent throughout the series in order to be able to use ARIMA model. The ARIMA model is essentially a combination of a few modelling techniques. The first part of the ARIMA model that is Auto-Regressive part says that the value of the variable in question progresses on its own prior values. The Moving-Average Part says that the error is a linear combination of past repeatedly occurring errors. And the Integrated part of the model indicates that the current values at any point in the time series is a difference of the prior current and previous values (Maybe be performed multiple times to and is used to make the time series stationary).

ARIMA models depend on three parameters p, d and q which are a representation of the 3 parts of the ARIMA model discussed earlier in this section. To choose correct values for p, d and q we mainly use the Auto-Correlation and Partial Auto-Correlation properties of the time series. Essentially by examining the graphs Auto-Correlation graph for differencing of order 1 and order 2 we can determine the value of d. Similarly we can plot the Auto-Correlation and Partial Auto-Correlation plot for the differenced time series and upon examining the graphs choose the values of p and q that lie within the range shown in the graph. [Verma 2022] & [Wikipedia [n. d.]]

2.2 Artificial Neural Networks (ANNs) & Long Short-Term Memory (LSTM)

Artificial Neural Networks fall under a sub-branch of machine learning where the model for the underlying process is build from the input data using a layered architecture. Each layer consists of a number of nodes or neurons which performs some operation on the input and then applies a non-linear function on it before sending the result as output. The output can then be passed onto other neurons or used for either classification or regression of the test variable. Each neuron has a weight which determines how important that neuron's feature is to determine the output variable. The process of calculating and finding the optimal weights for the neurons is called training of the artificial neural network. The biggest advantage of using ANN is that it is a general process that can be used to model any kind of process and is trained automatically.

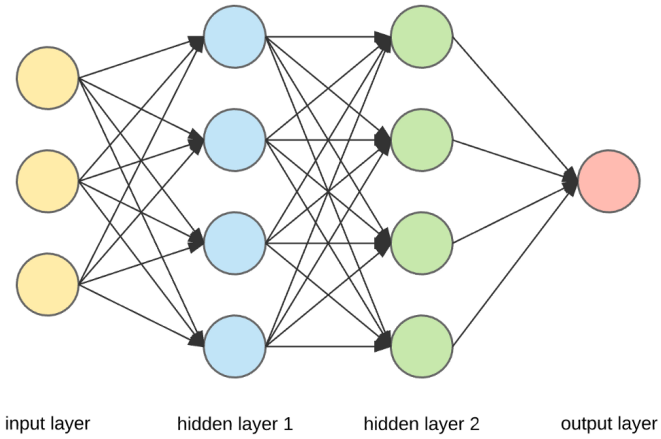

Fig. 1. An Example of a Neural Network

For ANNs part of the problem is determining the optimal network architecture for the neural network. Generally the deeper the architecture it will seem to perform better but then we run into the problem of overfitting the model. So we choose a shallow architecture maybe just 1 or 2 layers for the time series modelling but with a relatively higher number of neurons to cater for detecting all the patterns. 1 shows an example of a neural network architecture with 2 hidden layers and 4 neurons in each layer. We will use a similar architecture for our network but with 100 neurons in each layer.

The choice of the optimal architecture is the most important step of training an good neural network to model a process and there is not general rule to determine a good architecture. We can study the underlying data to guess good parameters for layers and nodes. But the most used and prevailing method is testing different architectures on the data and calculated the error at the end of a model to choose the best one.

LSTM is a type of artificial neural network architecture specially designed and developed to learn patterns that take place over a longer period of time. LSTM architecture can learn a more complex model for the underlying process and learn patterns over a longer period of time such as for time series data by using an architecture that has a smaller units over a few layers to control whaty to learn and what features to forget as part of the training process. [Gustav Gerholm 2021]

2.3 Convolutional Neural Networks (CNNs)

Convolusional Neural Nework (CNNs) are primarily used for image classification, object detection and segmentation. However, 1D CNN have also shown to be a usefull tool for forecasting when working with time-series data. It can be used to identify patterns in the historical stock price data and make predictions based on those patterns. The 1D CNN model architecture for stock price prediction typically consists of a series of 1D convolutional layers, followed by pooling layers, and then fully connected layers to make the final predictions. The convolusional layers learn patterns in the time series data, and the pooling layers reduce the dimensionlity of the data to prevent overfitting.

3 APPROACH

In this project we analyse and compare different models to predict stock prices. We gather data using the Yahoo API for the past 10 years of stock data. We then visualize the stocks and decompose then into their components and are mainly interested in the trend so visualize that. Next up we try to choose the p, d and q parameters for the ARIMA model and fit ARIMA model to the stock data and visualize its predictions. Next up we use the LSTM model to fit to the stock data for different companies, visualize the predictions and calculate Root Mean Squared Error (RMSE) of the predictions. Followed by CNN for the stock data and prediction using CNN.

3.1 Data Gathering

We start off by gathering stock data for the past 10 years. For the period 01.01.2013 - 31.12.2022, we gather stock price data. We used *yfinance* package for python that uses the Yahoo Finance API to gather stock historic stock prices from Yahoo Finance. We gathered the past 10 years of data instead of smaller time frames for two reasons. One we would like to have as much data as possible for training neural networks and secondly so as to avoid or minimize the effect of impact of COVID-19 on the stock markets all over the world. We have chosen the following stocks for seasonality analysis and prediction. Some of the stocks we have chosen are in currencies other than USD so we did not perform any aggregated analysis of all the stocks which might have provided better insight to the global stock market trend.

3.2 Companies Chosen

3.2.1 Equinor ASA. Equinor ASA (formerly known as Statoil) is a Norwegian multinational energy company that primarily operates in the oil and gas industry. Equinor's activities include exploration, production, transportation, refining, marketing of oil, gas and petroleum products.

As of the end 2021, Equinor ASA employed approximately 21,000 people globally. This includes both permanent and temporary employees, as well as contractors. The financial results for Equinor ASA shows a total revenues of approximatley 90,000 million US dollars and 149,000 million US dollars in 2021 and 2022 respectively. [eqn [n. d.]]

Equinor was founded in 1972 and headquartered in Stavanger, Norway, Equinor was formerly known as Statoil ASA and changed its name to Equinor ASA in may 2018

3.2.2 Marriott International, Inc. Marriott International, Inc. is a global hostpitality company that operates and franchises a broad portfolio of hotels and lodging facilities. The company operates over 7,600 properties across more than 130 countries and territories, with over 1.4 million rooms available for guests. Marriots business model involves both owning and operating hotels, as well as franchising its brands to third-party owners and operators.

According to the yahoo finacne, Marriott International reported a total revenue of approximatley 20,770 million US dollars in 2022 and emplyed 377,000 people worldwide. [mar [n. d.]]

The company was founded in 1927 and is headquartered in Bethesda, Maryland, USA.

Table 1. Stocks Chosen for Analysis and Prediction

Company	Yahoo Ticker	Currency
Equinor	EQNR	USD
Marriot Hotel	MAR	USD
Microsoft	MSFT	USD
Nestle	NESN.SW	CHF
Turkish Airlines	THYAO.IS	TRY

3.2.3 Microsoft Corporation. Microsoft Corporation is technology company that develops, licenses and sels a range of computer software, consumer electronics and personal computers. Microsoft is best known for its flagship Windows operating system, which is used by millions of users worldwide.

Microsoft Corporation reported a total revenue of 198,270 million US dollars in 2022 and emplyed 221,000 people worldwide. [msf [n. d.]]

The company was founded in 1975 by Bill Gates and Paul Allen and is headquartered in Redmond, Washington.

3.2.4 Nestle S.A.. Nestle S.A., together with its subsidiaries, operates as a food and beverage company. Nestle produces a wide range of food and beverage products, including baby food, bottled water, breakfast cereals, coffee and tea, dairy products, frozen food, pet food and snacks. The company's brand include NesCafe, KitKat, Gerber, Maggi, Nespresso and Purina among many others.

Nestle S.A. reported a total revenue of 98,000 million Swiss franc in 2022 and employed 275,000 people worldwide. [nes [n. d.]]

The company was founded in 1886 by Henri Nestle and is headquartered in Vevey, Switzerland.

3.2.5 Turkish Airlines - Türk Hava Yolları A.O.. Turkish Airlines provides air transport and aircraft technical maintenance services in Turkey, and internationally. It operates scheduled flights to more than 300 destinations in over 120 countries, making it one of the largest airlines in the world by number of destinations served.

According to the yahoo finacne, Turkish Airlines reported a total revenue of approximatley 311,169 million US dollars in 2022 and emplyed 40,264 people worldwide. [thy [n. d.]]

The company was founded in 1933 as State Airlines Administration and later renamed as Turkish Airlines in 1955. It's headquarter is located in Istanbul, Turkey.

We also intended to perform analysis on commodities such as Natural Gas and Wheat but Yahoo does not keep historic prices on commodities like it does for stocks so we excluded the commodity analysis from the final project.

After getting the historical data from Yahoo we saved the resultant data frames in csv files so as to avoid calling Yahoo API over and over again, since Yahoo has introduced a number of measures to stop scrappers and reduce usage of the API.

After gathering the data we visualize the the resultant datasets and the stock prices. We will only be working with the closing prices of each stocks as shown in 2. First of all we will analyze time series's components. Most time series data can be broken down into 2 kinds of components.

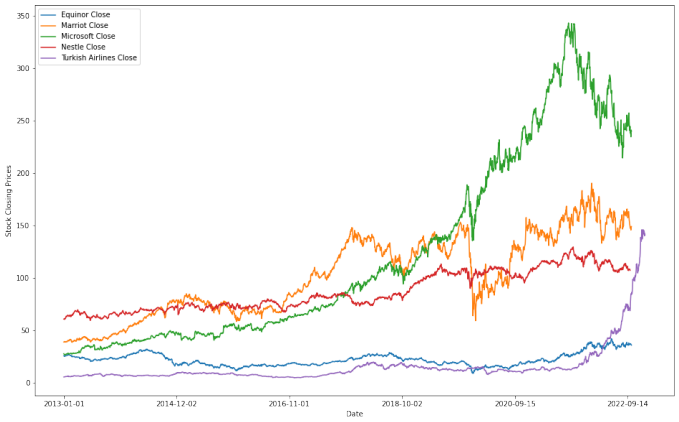

Fig. 2. Closing prices of the chosen stocks in the period (01.01.2013 - 31.12.2022)

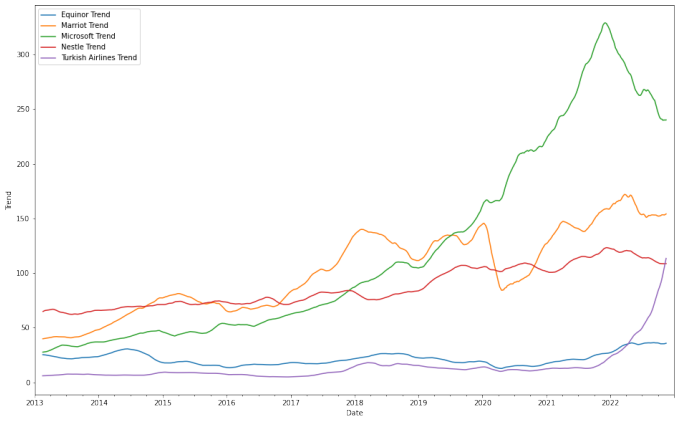

Fig. 3. Trend of the Time Series Stock Data between period (01.01.2013 - 31.12.2022)

- Systematic components are components that can be modelled. Systematic components of the Stock data are Trend and Seasonality.
- Non-Systematic components that cannot be modeled like noise. Stock data contains a lot of noise which can be variation to the stock prices from external factors like global phenomena not accounted for in the model.

There are 2 kinds of models to break a time series into components.

- Additive Model: This is the model where any point in the time series the value can be represented by a sum of components.
- Multiplicative Model: This is the model where any point in the series is represented by a product of the components.

We tried both models for decomposition of stock data and both presented similar results. So we are only including Additive Model decomposition results here. The seasonality component is too noisy and hard to read for the entire period.

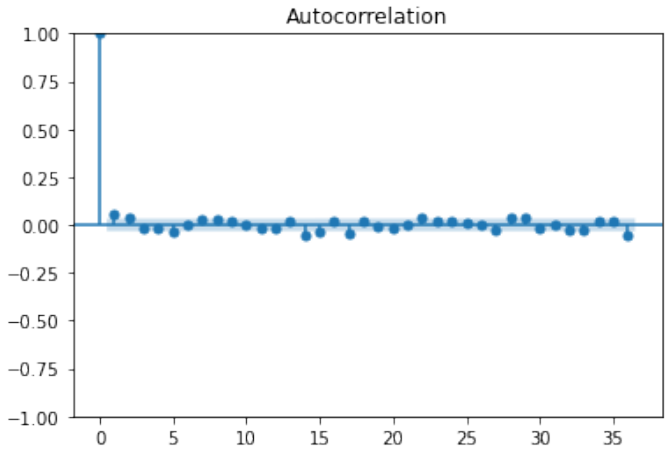

Fig. 4. Auto-Correlation of 1n Differenced Equinor Closing Prices between period (01.01.2013 - 31.12.2022)

3.3 Pre-Processing

The data gathered from Yahoo has missing values all through the dataset over the entire period. So as the first step of pre-processing we re-sample the time series values to fill in missing values where Yahoo does not have historic data for the stocks. We use linear interpolation function provided by pandas to fill in the missing data after re-sampling. Since we will only be working with closing prices of the stocks so creating a dataframe with closing prices of the 5 stocks we have chosen.

3.3.1 ARIMA. For ARIMA models the time series data needs to be stationary, so we perform stationarity test to see verify that the data is not stationary and then compute difference to make the time series stationary. Next up we plot the Auto-Correlation and Parial Auto-Correlation of the time series data to determine the acceptable values for p, d and q.

4 shows the auto correlation plot of 1n Differenced Closing prices of Equinor and 5 shows the partial auto correlation. From the plots we see that the 3rd value lies within the range and hence we use p and q equals to 2. The plots of all the stocks give the same value for p and q except Turkish Airlines and Nestle which might relate to those stocks being in currency other than USD.

3.3.2 LSTM & CNN. For LSTM and CNN we need to have input data and labels for the input data to be able to train the Artificial Neural Networks on the data. So first we split the dataset with a 80:20 percent ration between training and test dataset. And next up we use windows of 30 days so 30 days as input and the next 30 days as the labels for both training and test datasets.

4 RESULTS

We show plots for the stocks we have gathered. The plots show the actual stock value as well as the predicted stock value. As we can see from the graphs ARIMA model captures the trend of the stocks completely correctly given that the p, d and q parameters are selected correctly however ARIMA model fails to capture the

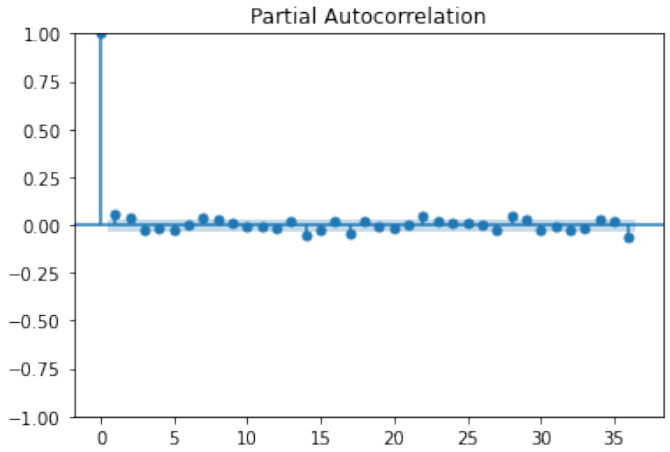

Fig. 5. Partial Auto-Correlation of 1n Differenced Equinor Closing Prices between period (01.01.2013 - 31.12.2022)

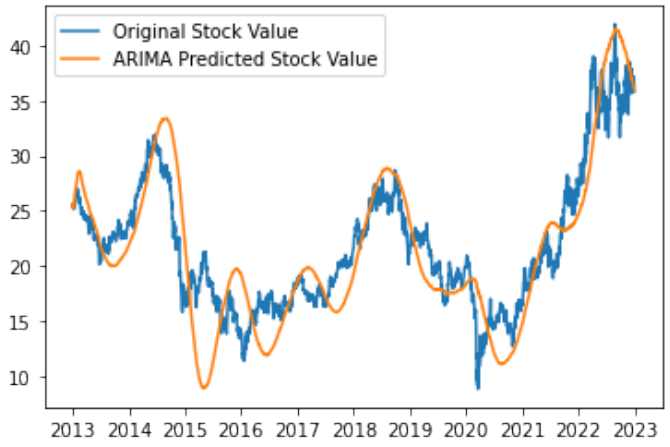

Fig. 6. Equinor ARIMA Model

granularity of the noise in the stock data. Which in turn is captured both by CNN and LSTM owing to their greater potential to fit non-linear models to the underlying dataset. However the results of the CNN and LSTM are very similar or one might even say identical. **The Complete figures of all the company stocks and their predictions are listed in the Appendix**

We furthermore show the root mean squared error of CNN and LSTM for the predictions of all the stocks. And from the RMSE we can clearly see that LSTM performs slightly better than CNNs but that in part can be attributed to LSTM overfitting to that data since we use 3 epochs. 2

REFERENCES

[n. d.]. *Equinor (EQNR) Yahoo Finance.* https://finance.yahoo.com/quote/EQNR/financials?p=EQNR
[n. d.]. *Marriot (MAR) Yahoo Finance.* https://finance.yahoo.com/quote/MAR/financials?p=MAR

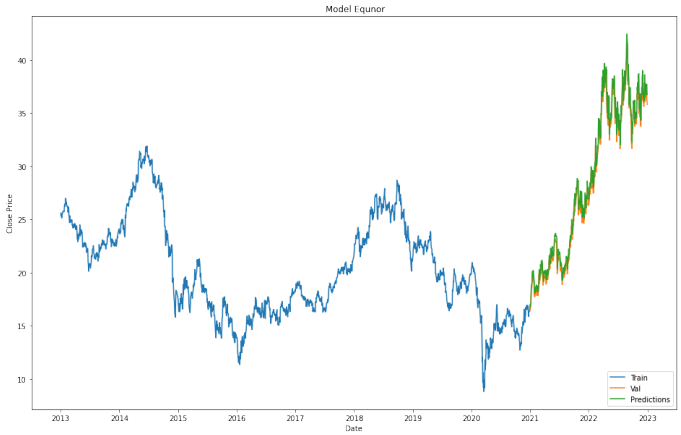

Fig. 7. Equinor LSTM Model

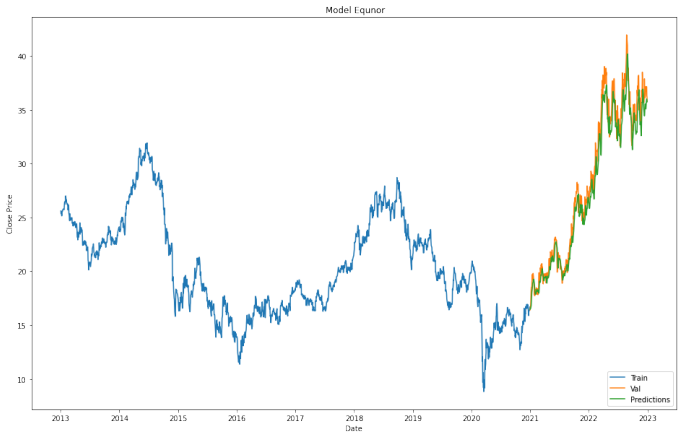

Fig. 8. Equinor CNN Model

Table 2. Root Mean Squared Error of CNN and LSTM

Company	RMSE CNN	RMSE LSTM
Equinor	2.3621428920798104	0.22169233478911934
Marriot Hotel	1.259793984400083	1.2866454085258587
Microsoft	4.854215564466503	11.873283114498609
Nestle	1.7095982695279055	0.17975368761036495
Turkish Airlines	0.944293598279561	2.301449732584496

[n. d.]. *Microsoft (MSFT) Yahoo Finance.* https://finance.yahoo.com/quote/MSFT?p=MSFT&.tsrc=fin-srch
[n. d.]. *Nestle (NESN.SW) Yahoo Finance.* https://finance.yahoo.com/quote/NESN.SW/profile?p=NESN.SW
[n. d.]. *Turkish Airlines (THYAO.IS) Yahoo Finance.* https://finance.yahoo.com/quote/THYAO.IS?p=THYAO.IS&.tsrc=fin-srch
Adam Lindberg Gustav Gerholm. 2021. Comparison of machine learning models for market predictions with different time horizons. (2021).
Qihang Ma. 2020. Comparison of ARIMA, ANN and LSTM for Stock Price Prediction. *E3S Web of Conferences* (2020). https://doi.org/10.1051/e3sconf/202021801026
Yugesh Verma. 2022. *Quick way to find p, d and q values for ARIMA.* https://analyticsindiamag.com/quick-way-to-find-p-d-and-q-values-for-arima/

Fig. 11. Marriot Hotel CNN Model

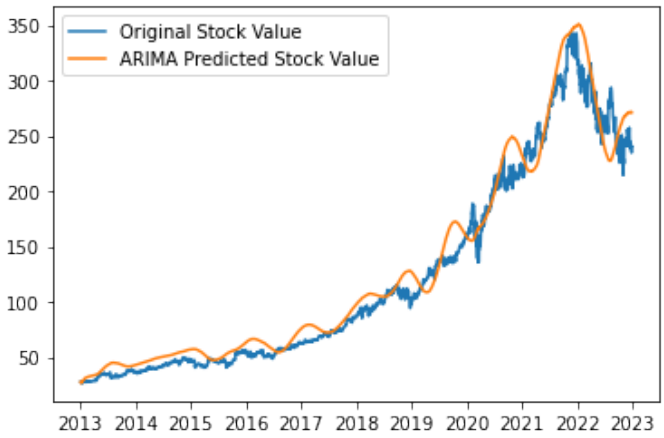

Fig. 12. Microsoft ARIMA Model

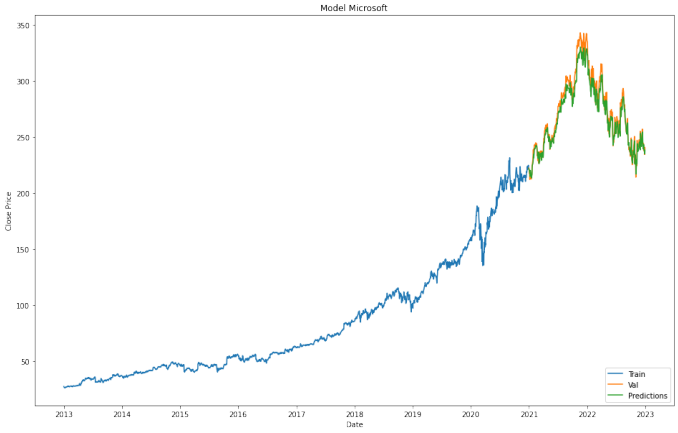

Fig. 13. Microsoft LSTM Model

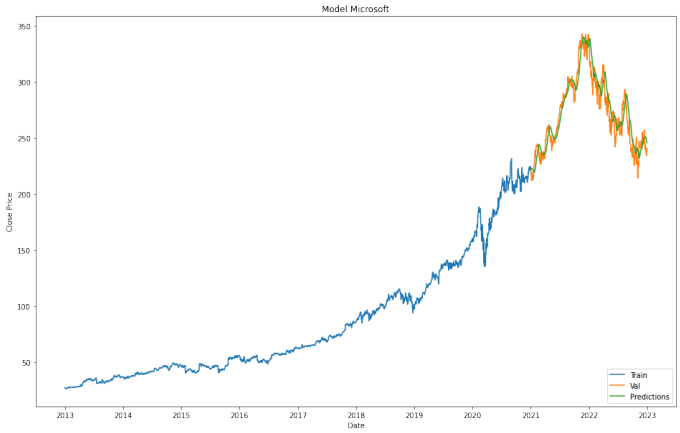

Fig. 14. Microsoft CNN Model

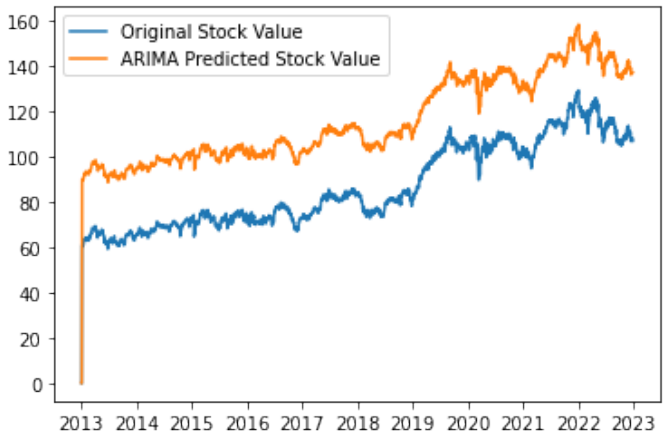

Fig. 15. Nestle ARIMA Model

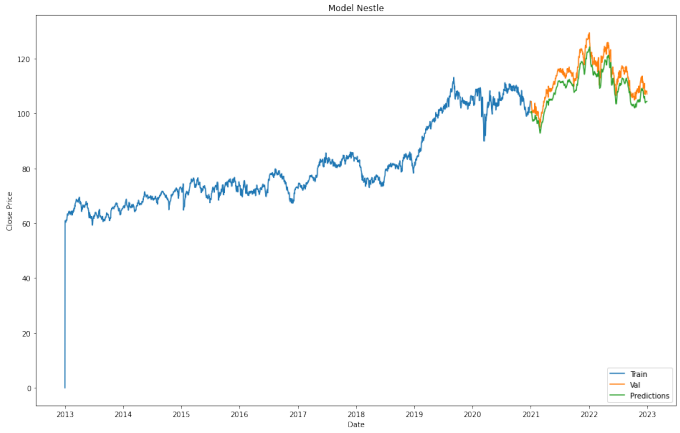

Fig. 16. Nestle LSTM Model

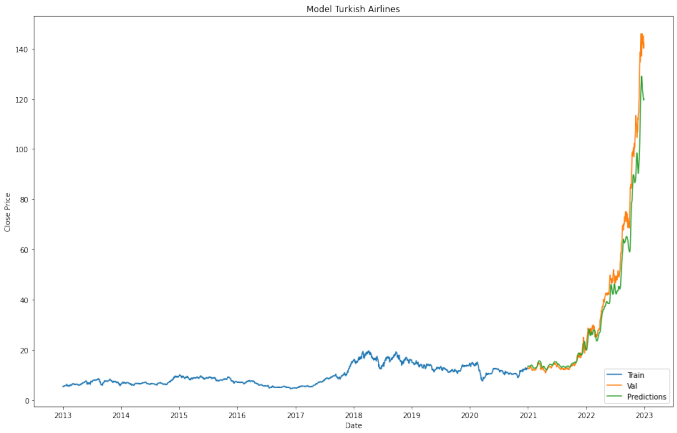

Fig. 20. Turkish Airlines CNN Model

Fig. 10. Marriot Hotel LSTM Model

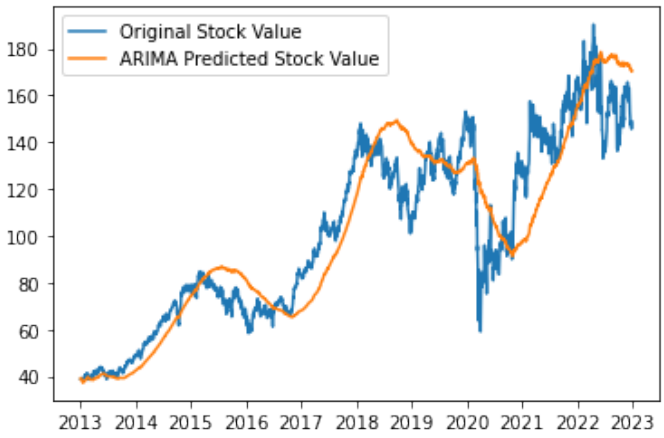

Fig. 9. Marriot Hotel ARIMA Model

Wikipedia. [n. d.]. *Autoregressive integrated moving average.* https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average

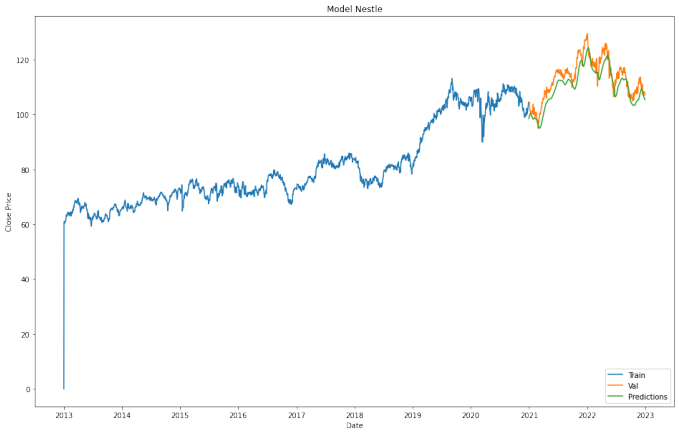

Fig. 17. Nestle CNN Model

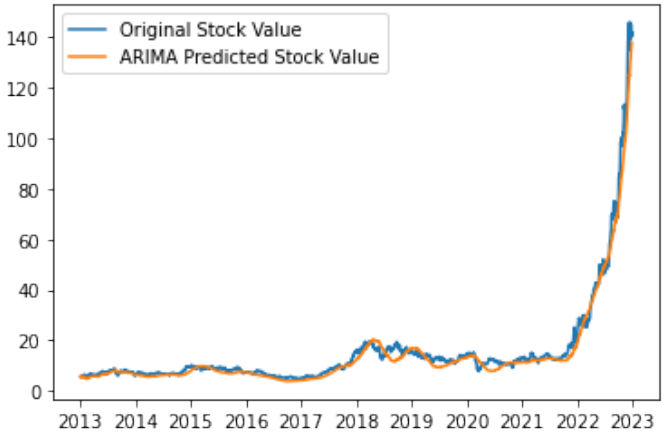

Fig. 18. Turkish Airlines ARIMA Model

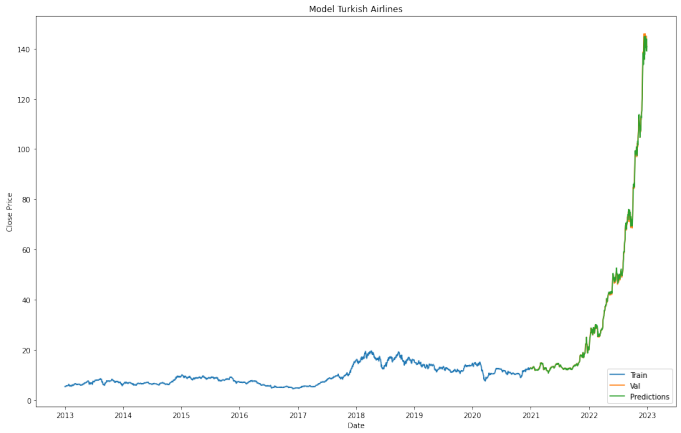

Fig. 19. Turkish Airlines LSTM Model