American Journal of
Electronics & Communication

# Automated Stock Price Prediction using LSTM Recurrent Neural Network

**Prof. Poulami Ghosh, Kushal Basak , Poushali Santra**
Dept. of CA, University of Engineering & Management, Kolkata, INDIA

***In Stock Market Prediction, the aim is to predict the future value of the financial stocks of a company. The recent trend in stock market prediction technologies is the use of machine learning which makes predictions based on the values of current stock market indices by training on their previous values. Machine learning itself employs different models to make prediction easier and authentic. The paper focuses on the use of Regression and LSTM based Machine learning to predict stock values. Factors considered are open, close, low, high and volume.***

***A correct prediction of stocks can lead to huge profits for the seller and the broker. Frequently, it is brought out that prediction is chaotic rather than random, which means it can be predicted by carefully analyzing the history of the respective stock market. Machine learning is an efficient way to represent such processes. It predicts a market value close to the tangible value, thereby increasing the accuracy. The introduction of machine learning to the area of stock prediction has appealed to many researchers because of its efficient and accurate measurements.***

***In Stock Market Prediction, the aim is to predict the future value of the financial stocks of a company. The recent trend in stock market prediction technologies is the use of machine learning which makes predictions based on the values of current stock market indices by training on their previous values. Machine learning itself employs different models to make prediction easier and more authentic. The paper focuses on the use of Regression and LSTM based Machine learning to predict stock values. Factors considered are open, close, low, high, and volume.***

***Stock market prediction is a major exertion in the field of finance and establishing businesses. The stock market is totally uncertain as the prices of stocks keep fluctuating on a daily basis because of numerous factors that influence it. One of the traditional ways of predicting stock prices was by using only historical data. But with time it was observed that other factors such as peoples' sentiments and other news events occurring in and around the country affect the stock market, for e.g. national elections, natural calamities etc. Investors in the stock market seek to maximize their profits for which they require tools to analyze the prices and trends of various stocks. Machine learning algorithms have been used to devise new techniques to build prediction models that can forecast the prices of stock and tell about the market trend with good accuracy. Many prediction models have been proposed to incorporate all the major factors affecting the price of stocks.***

***Due to the correlated nature of stock prices, conventional batch processing methods cannot be utilized efficiently for stock market analysis. We propose an online learning algorithm that utilizes a kind of recurrent neural network (RNN) called Long Short Term Memory (LSTM), where the weights are adjusted for individual data points using stochastic gradient descent. This will provide more accurate results when compared to existing stock price prediction algorithms. The network is trained and evaluated for accuracy with various sizes of data, and the results are tabulated. A comparison with respect to accuracy is then performed against an Artificial Neural Network.***

***Keyword: stock market, prediction, machine learning, neural network historical data***

## II. INTRODUCTION

Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. A correct prediction or analysis of stock can lead to huge profit for both the seller and broker and this can be achieved by carefully analyzing the history of respective stock market.

Machine learning is one of the efficient way to represent such process. It helps in predicting a tangible value which is very much close to the market value and thereby increasing the accuracy. It's accurate and efficient measurements has appealed a lot of researchers.

The vital part of any machine learning algorithm is the data on which the algorithm is supposed to function. The dataset should be as concrete as possible since a slightest mismatch can skew the data and it will have a huge impact in the outcome

In this project, we are using supervised machine learning algorithm and the dataset is obtained from Yahoo Finance. Models used: Root Mean Squared Error (RMSE) and Long Short Term Memory (LSTM) are the two models that are used in this project.

While RMSE is used to reduce the errors, LSTM is used for the remembrance of the data and the results for a long period of time. Money related authorities think about the expression, Buy low, Move high yet this does not give enough setting to settle on proper endeavor decisions. Before an investigator places assets into any stock, he should realize how money markets continues Setting assets into a wonderful stock regardless at a horrible time can have awful results, while vitality for a common stock at the fortunate time can hold up under focal points. Cash related monetary pros of today are going toward this issue of trading as they don't for the most part understand concerning which stocks to buy or which stocks to offer with the authentic objective to get impeccable focal points. RMSE and LSTM models are engaged for this conjecture separately. RMSE involves minimizing error and LSTM contributes to remembering the data and results for the long run. Graphs for Close price history and the LSTM model are plotted. RMSE graph plots fluctuation of prices with corresponding dates and LSTM model plots actual and predicted prices.

The stock market is a vast array of investors and traders who buy and sell stock, pushing the price upon down. The prices of stocks are governed by the principles of demand and supply, and the ultimate goal of buying shares is to make money by buying stocks in companies whose perceived value (i.e., share price) is expected to rise. Stock markets are closely linked with the world of economics —the rise and fall of share prices can be traced back to some Key Performance Indicators (KPI's). The five most commonly used KPI's are the opening stock price (`Open'), end-of-day price (`Close'), intraday low price (`Low'), intra-day peak price (`High'), and total volume of stocks traded during the day (`Volume'). Economics and stock prices are mainly reliant upon subjective perceptions about the stock market. It is near impossible to predict stock prices to the T, owing to the volatility of factors that play a major role in the movement of prices. However, it is possible to make an educated estimate of prices. Stock prices never vary in isolation: the movement of one tends to have an avalanche effect on several other stocks as well. This aspect of stock price movement can be used as an important tool to predict the prices of many stocks at once. Due to the sheer volume of money involved and number of transactions that take place every minute, there comes a trade-off between the accuracy and the volume of predictions made; as such, most stock prediction systems are implemented in a distributed, parallelized fashion. These are some of the considerations and challenges faced in stock market analysis.

It has never been easy to invest in a set of assets, the abnormally of financial market does not allow simple models to predict future asset values with higher accuracy. Machine learning, which consist of making computers perform tasks that normally requiring human intelligence is currently the dominant trend in scientific research. This article aims to build model using Recurrent Neural Networks (RNN) and especially Long-Short Term Memory model (LSTM) to predict future stock market values.

Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Long Short-Term Memory (LSTM) is one of many types of Recurrent Neural Network RNN, it's also capable of catching data from past stages and use it for future predictions. In general, an Artificial Neural Network (ANN) consists of three layers: 1) input layer, 2) Hidden layers, 3) output layer. In a NN that only contains one hidden layer the number of nodes in the input layer always depend on the dimension of the data, the nodes of the input layer connect to the hidden layer via links called 'synapses'. The relation between every two nodes from (input to the hidden layer), has a coefficient called weight, which is the decision maker for signals. The process of learning is naturally a continues adjustment of weights, after completing the process of learning, the Artificial NN will have optimal weights for each synapses. The hidden layer nodes apply a sigmoid or tangent hyperbolic (tanh) function on the sum of weights coming from the input layer which is called the activation function, this transformation will generate values, with a minimized error rate between the train and test data using the SoftMax function. The values obtained after this transformation constitute the output layer of our NN, these value may not be the best output, in this case a back propagation process will be applied to target the optimal value of error, the back propagation process connect the output layer to the hidden layer, sending signal conforming the best weight with the optimal error for the number of epochs decided. This process will be repeated trying to improve our predictions and minimize the prediction error. After completing this process, the model will be trained. The classes of NN that predict future value base on passed sequence of observations is called Recurrent Neural Network (RNN) this type of NN make use of earlier stages to learn of data and forecast futures trends. The earlier stages of data should be remembered to predict and guess future values, in this case the hidden layer act like a stock for the past information from the sequential data. The term recurrent is used to describe the process of using elements of earlier sequences to forecast future data. RNN can't store long time memory, so the use of the Long Short-Term Memory (LSTM) based on "memory line" proved to be very useful in forecasting cases with long time data. In a LSTM the memorization of earlier stages can be performed through gates with along memory line incorporated.

Rest of the paper is based on detail analysis of the data used and their detailed comparison of the results produced with LSTM model and the last part discusses about future scope of our project.

## III. RELATED WORKS

[1] In the early research related to stock market prediction, Fama, E. F. (1970) proposed the Efficient Market Hypothesis (EMH)

[2] Horne, J. C., & Parker, G. G. (1967) proposed the Random Walk theory. These theories proposed that market prices are affected by information other than historical prices and thus market price cannot be predicted.

[3] The EMH theory suggests that the price of a stock depends completely on market information and thus any new information will lead to a price change as a reaction of the newly released information. This theory also claimed that stocks are always traded on their fair value, where traders cannot buy nor sell stocks in a special price undervalued or inflated and therefore the only way a trader can increase her profits is by increasing her risk.

[4] EMH discusses three different variations that affect market price: Weak Form, where only historical data is considered, semi-Strong Form, which incorporates current public data in addition to historical data, and Strong Form, which goes farther to incorporate private data. EMH states that any price movement is either a result of new released information or a random move that would prevent prediction models from success.

[5] The Random Walk Hypothesis by Horne, J. C., & Parker, G. G. (1967) states that the stock prices are randomly changed and argue that past price movements are independent of current movements. This is slightly different from EMH as it focuses on short-term pattern of stock market.

[6] Based on the above two hypotheses by Horne, J. C. et al. (1967) and Fama, E. F. (1970), the stock market will follow a random move and the 17 accuracy of predicting such movement cannot exceed 50%. As opposed to these theories, many recent studies have shown that stock market price movement can be predicted to some degree.

[7] Arévalo, A. et al. (2016) used four main features as input to a Depp Neural Network (DNN) model. He formalized the input data as follows: the time feature which is included in the inputs as minutes and hours parameters, and a variable window size (n) which is used for the other inputs. Thus, the input file will include last n pseudo-log-return, last n standard deviations and last n trend indicators. The output of the model was "next one-minute pseudo-log-ret.

[8] The model was trained during 50 epochs with different window sizes and the results show that window size 3 can show the best performance of the model with accuracy 66% and 0.07 MSE.

[9] Weng, B. et al. (2017) attempted to predict one day ahead price

movement using disparate sources of data, where combining data from online sources with prices and indicators can enhance the prediction of the stock market state. This study was tested on Apple Inc. (APPL) stock information gathered over 3 years with multiple inputs and different output targets.

[10] Schumaker, R. P. et al. (2009) tried to predict direction of the price movement based on financial news. The study was done in 2009 as market prediction was and still facing difficulties due to the ill- defined parameters. In order to use the financial news articles in the prediction model, news should be represented as numerical value.

[11] AZFin Text is another system built by (Schumaker, R. P. et al 2009) that predicts price changes after 20 minutes of news release.

## IV. METHODOLOGY

In stock market we have to face a lot of problems regarding the prediction of the prices of the stocks and that's why we are doing this project on stock market prediction using machine learning algorithm. The main objective of this work is to predict the closing stock prices of an organization (Apple Inc.) using an artificial recurrent neural network called Long Short Term Memory (LSTM) . There are so many factors involved in the prediction – physical factors vs physiological, rational and irrational behavior, etc. All these aspects combine to make share prices volatile and very difficult to predict with a high degree of accuracy.

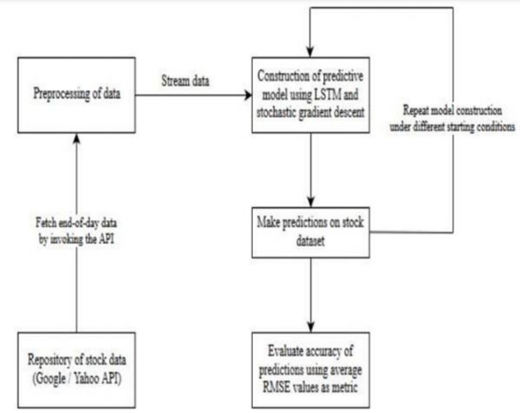

Preprocessing of data
Stream data
Construction of predictive model using LSTM and stochastic gradient descent
Repeat model construction under different starting conditions
Fetch end-of-day data by invoking the API
Make predictions on stock dataset
Repository of stock data (Google / Yahoo API)
Evaluate accuracy of predictions using average RMSE values as metric

Fig.1. LSTM-based stock price prediction system

•Technical Analysis

Technical Analysis is helpful to estimate the future economic stock movement based on stock historical movement. Technical limitations do not forecast stock price, but based on historical analysis, technical limits can forecast the stock movement on existing market condition over time. Technical examination help depositor to forecast the stock price movement (up/down) in that specific time interval. Technical examination habits a wide diversity of charts that show price over period. Then close index of every company is taken into account and put it into one data frame and try to find a connection between each company and then pre-processing the data and creating different technical parameters built on stock price, bulk and close worth and based on the movement of prices will progress technical meters that will aid set a target percentage to foretell buy, sell, hold.

•Data Manipulation

Presently that, stock estimating information of organizations is put away, unite this information in one information outline. While, the majority of the information available to us is now there, may need to really get to the information together. To do this, join the majority of the stock datasets together. Every one of the stock records right now accompany: Open, High, Low, Close, Volume. At any rate to begin, simply intrigued by the nearby for the present. Draw our recently made rundown of tickers and start with an unfilled information outline, and be prepared to peruse in each stock's information outline, for the most part intrigued by the Close segments, and the segment has been renamed to whatever the ticker name is and a common information outline begins building. Pandas outer join was used to combine the data frame and if there's nothing in the main d/f, then start with the current d/f, otherwise use pandas join.

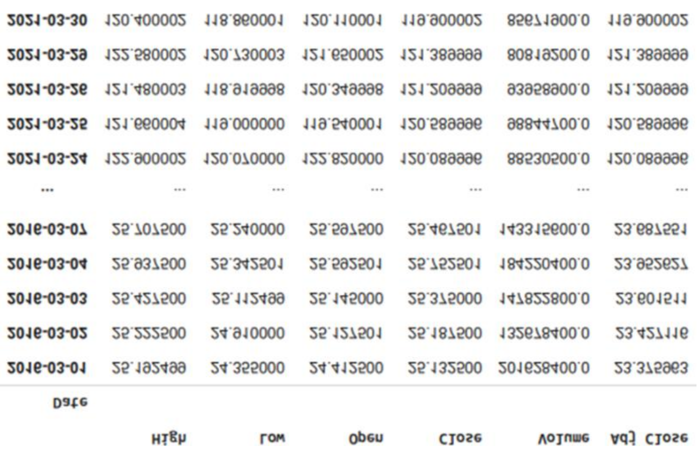
Now, we have to check if any interesting correlation data is found. To do this, we have to visualize it, since it's a lot of data. We used matplot lib with numpy to fulfill this task.

We have also used other library functions such as pandas_datareader to extract the past records of stocks from the web.

To deliver, since there will be specific space between the x's and y's. Generally, matplotlib foliage room on the extreme ends of graph since this inclines to make graphs calmer to read, but, in this case, it doesn't.

•Data Extraction from Yahoo API

Stock data of the company "Apple" was extracted from 01-03-2016 to 30-03-2021

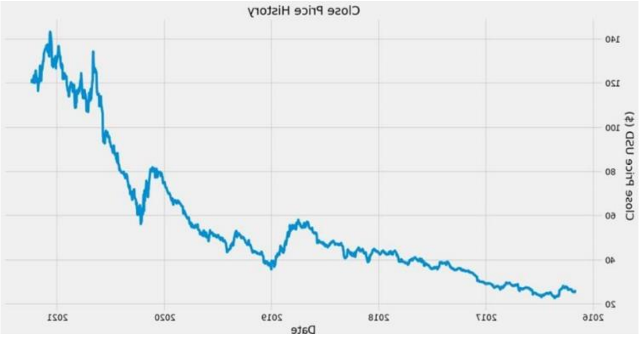


```
                  High         Low        Open       Close       Volume   Adj Close
Date
2016-03-01   25.192499   24.355000   24.412500   25.132500  201628400.0   23.375693
2016-03-02   25.222500   24.910000   25.127501   25.187500  132678400.0   23.427116
2016-03-03   25.427500   25.112499   25.145000   25.375000  147822800.0   23.601511
2016-03-04   25.937500   25.342501   25.592501   25.752501  184220400.0   23.952057
2016-03-07   25.707500   25.240000   25.597500   25.467501  143315600.0   23.687551
       ...         ...         ...         ...         ...          ...         ...
2021-03-24  122.900002  120.070000  122.820000  120.089996   88530500.0  120.089996
2021-03-25  121.660004  119.000000  119.540001  120.589996   98844700.0  120.589996
2021-03-26  121.480003  118.919998  120.349998  121.209999   93958900.0  121.209999
2021-03-29  122.580002  120.730003  121.650002  121.389999   80819200.0  121.389999
2021-03-30  120.400002  118.860001  120.110001  119.900002   85671900.0  119.900002
```

The five Key Performance Indicators (KPI's) used here are the opening stock price (`Open'), end- of-day price (`Close'), intraday low price (`Low'), intra-day peak price (`High'), and total volume of stocks traded during the day (`Volume').

•Visualization of the data

The actual closing price of the stock every year was calculated and visualized with a graph given below:



•Training and Testing of LSTM

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. An LSTM has a similar control flow as a recurrent neural network. It processes data passing on information as it propagates forward. The differences are the operations within the LSTM's cells. These operations are used to allow the LSTM to keep or forget information.

After building the model, it was trained using the tensorflow and the errors while training it were scaled using MinMax Scaler.

The loss in compilation of the model was calculated using the root mean square formula,.i.e., the error was detected with this formula. After that the data was converted into numpy array for testing.

•Evaluation of the model using Root Mean Square Error (RMSE)

Mathematics plays vital role in developing a model using RNN (Deep Machine Learning). We require measured methods to:

- Obtain the correct algorithm on the mathematical data available.
- Obtaining correct features and parameters for the model so that the future estimates are precise.
- To evaluate the model for over fitting and under fitting of the data set.
- Expecting the ambiguity of the project model.

RMSE- It is the mean of all the squared errors attained by an algorithm in logistic regression. The error is the change between the actual value and the calculated value, which is intended 28 between numerous data points.

$$\sum_{i=1}^{n}\frac{\left(w^{T}x(i)-y(i)\right)^{2}}{n}$$

Here: wT is the weight connected, x(i) is the forecast value and y(i) is the definite value

•Comparison between the predicted and actual closing stock prices A graph was plotted to compare between the predicted and actual closing stock prices using Pandas. It was also shown using the data. After training the model using the last 60 days, the predicted and the actual price was printed.

## V. ALGORITHM

Step 1 – The libraries have been imported as numpy as np, pandas as pd, matplotlib.pyplot as plt, math, pandas_datareader as web and from sklearn, MinMaxScaler was imported, from keras.models imported Sequential and from keras.layers imported Dense and LSTM.

Step 2- The data has been loaded from Yahoo and the files have been imported from the company Apple.

Step 3- The number of rows and columns have been counted in the dataset. There were 1280 rows (date) and 6 columns (features of the stock quote).

Step 4- Visualized the closing stock price history in a graphical way.

Step 5- Created a new data frame using the close column to convert it into a numpy array

Step 6- Used the number of rows to train the model using the training data using MinMaxScaler.

Step 7- Created the training data set and the scaled training data set.

Step 8- Splitted the data into x_train and y_train data sets, converted them into numpy arrays, then reshaped them.

Step 9-Built the LSTM model and compiled it.

Step 10- Calculated the models' predicted price values using the scaler inverse transform. Step 11- Evaluated the model using the root mean squared error (RMSE).

Step 12- Visualized the compared data of closing stock price(USD) predicted by LSTM and the valid value using Pandas library.

Step 13- The quote of the company Apple was taken again from the data source Yahoo starting from 01-03-2021 till 30-03-2021(the dates can be modified anytime).

Step 14- The data (feature scaling) from sklearn which was imported from Yahoo has been scaled.

Step 15- An empty list has been made to keep the predicted values to compare later on. Step 16- The models' accuracy has been printed on the training data.

Step 17- The predicted and the valid closing stock for the 61st day,i.e., 31-03-2021 was printed.

## VI: RESULTS AND DISCUSSIONS

The proposed method has been implemented using Python on Google Collaborators. After creating predictive model, efficiency can be checked. For this, the model can be tested using Root Mean Square Error method. It is the mean of all the squared errors attained by an algorithm in logistic regression. The error is the change between the actual value and the calculated value, which is intended 28 between numerous data points.

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(Predicted_i-Actual_i)^2}{N}}$$

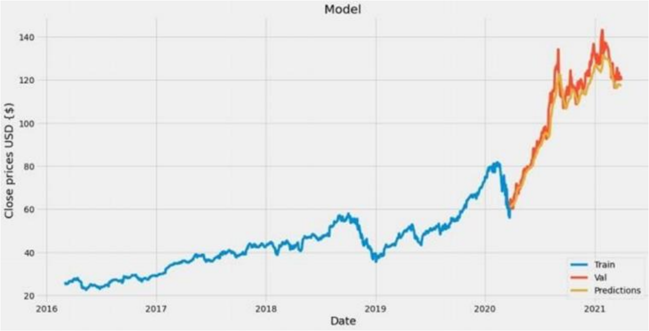
where, N is the total no. of observations

Replacing the predicted and actual values in this formula for the date 31-03-2021

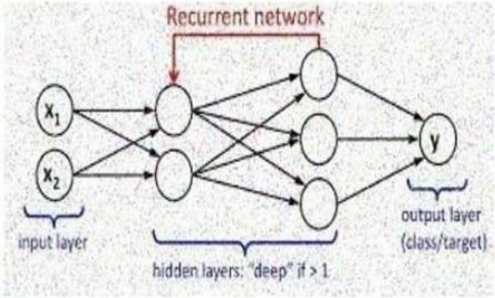
- Predicted closing stock price - 116.967766

- Actual closing stock price - 122.150002
- RMSE calculated = 3.9978585690259933
- Accuracy rate = 40% approx.

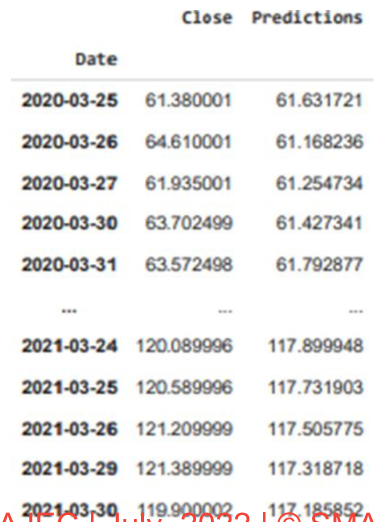
The graphical representation of the trained dataset for the years 2016-2020 (indicated by the blue line) along with the comparison of the Valid (indicated by the red line) and Predicted (indicated by the yellow line) closing stock price for the year 2020-2021 using Long Short Term Memory has been shown below:



The diagram of Recurrent Neural Network (RNN) has been shown below:



The comparison chart for the last 5 days for the month of March 2020 and March 2021 has been shown with the actual and predicted closing values for each date:



| Date | Close | Predictions |
|---|---|---|
| 2020-03-25 | 61.380001 | 61.631721 |
| 2020-03-26 | 64.610001 | 61.168236 |
| 2020-03-27 | 61.935001 | 61.254734 |
| 2020-03-30 | 63.702499 | 61.427341 |
| 2020-03-31 | 63.572498 | 61.792877 |
| ... | ... | ... |
| 2021-03-24 | 120.089996 | 117.899948 |
| 2021-03-25 | 120.589996 | 117.731903 |
| 2021-03-26 | 121.209999 | 117.505775 |
| 2021-03-29 | 121.389999 | 117.318718 |
| 2021-03-30 | 119.900002 | 117.185852 |

From the calculation and metrics above, we can say that the Long Short Term Memory model showed an accuracy score of about 40%. Hence, this model can be chosen predict the stock market closing price for any company.

So, basically the algorithm we used here depends upon the recurrent neural network where we consider long short-term memory cell. Neural Network mimics the behavior of the brain and sometimes attains the superhuman capabilities.

## VII: CONCLUSION

Hereby, it can be proposed that no trading algorithm can be 100% effective, not only 100%, it will typically never be close to 70% but to attain even an accuracy of 40% or 35% is still good sufficient to get a good forecast spread. Although extreme attained accurateness was 39%, it was still able to closely forecast the predictable outcome and have coordinated against the company graph. To make our expectation more efficient, it can be done by including bulky data sets that have millions of entries and could train the machine more powerfully. Different activities of stocks can lead to diverse raises or lows in the forecast price, use these movements to magistrate whether a company should be traded in or not. No training Data can ever be stable, hence there are always some unevenness which can be seen in the above data spread, but to still forecast close to a consequence will also lead to a good approach if it has greater than 33% accuracy.

While, developing a strategy trader should always think to always have nominal imbalance while still being above 33% accurate. This paper proposes RNN based on LSTM built to forecast future values for both GOOGL and NKE assets, the result of our model has shown some promising result. For different data set we can observe that training with less data and more epochs can improve our testing result and at the same time allow us to have beater forecasting and prediction values. The following table shows the precision of our training and testing for all the epochs for Apple Inc. The testing result conform that our model is capable of tracing the evolution of opening prices for both assets. For our future work we will try to find the best sets for bout data length and number of training epochs that beater suit our assets and maximize our predictions accuracy. While most studies are generally well constructed and reasonably well validated, certainly greater attention to experimental design and implementation appears to be warranted, especially with respect to the quantity and quality of economical data. Improvements in experimental design along with improved biological validation would no doubt enhance the overall quality, generality and reproducibility of many machine-based classifiers. Overall, we believe that if the quality of studies continues to improve, it is likely that the use of machine learning classifier will become much more commonplace in many clinical and hospital settings.

## VIII: REFERENCE

[1] Batres-Estrada, B. (2015). Deep learning for multivariate financial time series.

[2] Emerson, S., Kennedy, R., O'Shea, L., & O'Brien, J. (2019, May). Trends and Applications of Machine Learning in Quantitative Finance. In 8th International Conference on Economics and Finance Research (ICEFR 2019).

[3] Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: deep portfolios. Applied Stochastic Models in Business and Industry, 33(1), 3-12.

[4] Moritz, B., & Zimmermann, T. (2016). Tree-based conditional portfolio sorts: The relation between past and future stock returns. Available at SSRN 2740751.

[5] Olah, C. (2015). Understanding lstm networks–colah's blog. Colah. github. io.

[6] Paiva, F. D., Cardoso, R. T. N., Hanaoka, G. P., & Duarte, W. M. (2018). Decision-Making for Financial Trading: A Fusion Approach of Machine Learning and Portfolio Selection. Expert Systems with Applications.

[7] Patterson J., 2017. Deep Learning: A Practitioner's Approach, O'Reilly Media.

[8] Siami-Namini, S., & Namin, A. S. (2018). Forecasting economics and financial time series: Arima vs. lstm. arXiv preprint arXiv:1803.06386.

[9] Takeuchi, L., & Lee, Y. Y. A. (2013). Applying deep learning to enhance momentum trading strategies in stocks. In Technical Report. Stanford University.

[10] Wang, S., and Y. Luo. 2012. "Signal Processing: The Rise of the Machines." Deutsche Bank Quantitative Strategy (5 June).

[11] In the early research related to stock market prediction, Fama, E. F. (1970) proposed the Efficient Market Hypothesis (EMH)

[12] Horne, J. C., & Parker, G. G. (1967) proposed the Random Walk theory. These theories proposed that market prices are affected by information other than historical prices and thus market price cannot be predicted.

[13] The EMH theory suggests that the price of a stock depends completely on market information and thus any new information will lead to a price change as a reaction of the newly released information. This theory also claimed that stocks are always traded on their fair value, where traders cannot buy nor sell stocks in a special price undervalued or inflated and therefore the only way a trader can increase her profits is by increasing her risk.

[14] EMH discusses three different variations that affect market price: Weak Form, where only historical data is considered, semi-Strong Form, which incorporates current public data in addition to historical data, and Strong Form, which goes farther to incorporate private data. EMH states that any price movement is either a result of new released information or a random move that would prevent prediction models from success.

[15] The Random Walk Hypothesis by Horne, J. C., & Parker, G. G. (1967) states that the stock prices are randomly changed and argue that past price movements are independent of current movements. This is slightly different from EMH as it focuses on short-term pattern of stock market.