ISSN No:-2456-2165

Evaluating Prediction of Stock Price using Machine Learning

Amit Kumar Yadav[1]; Rohit Sharma[2]; Swastik Bainsla[3]
Manav Rachna International Instituite of Research and Study

Abstract:- The extrapolation of stock prices is an essential and unresolved problem in the sphere of finance because the results of an accurate forecast can produce considerable economic consequences and the nature of the markets makes the task difficult. This research aims at applying the concept of machine learning in forecasting of stock price for Google shares using historical data of the company's stock for the last20 years. The qualitative aspect of the research is the collection of data with the use of the yfinance API, data preprocessing with the handling of missing values and removal of outliers. If further feature engineering, then the technical indicators included the simple moving averages and daily returns in order to improve on the capability of the model. Three types of machine learning models – Linear Regression, Random Forest, and Long Short-Term Memory (LSTM) Networks – were built experimentally and compared based on MAE and RMSE performance indices. Out of these, LSTM model provided better performance because it deals with temporal issues well by capturing temporal dependency and non linear trends in the data. In so doing, this research establishes the significance of state-of-the-art generous learning models in monetary prediction while stressing the efficacy of data origination and feature engineering. The results are quite informative for investors and financial analysts, as well as for improving the creation of further prediction models. Future work can also complement internal information with external variables like sentiment analysis and macroeconomic factors to improve their models.

Keywords:- *Stock Prediction, Machine Learning, LSTM, Stock Price Forecasting, Feature Engineering, Financial Time Series, Yfinance.*

I. INTRODUCTION

The stock market is the integration of many moving parts and all that happens around can influence the operations including the economic situation, political occurrences, and attitude among others in the market. Stock price forecasting is one of the oldest objectives of the financial analysis, economists, and researchers because of the impact on risk and return reduction. However, fluctuations and stochastic character of stock prices is indeed a crucial problem of financial markets, and traditional techniques of forecasting are not very effective considering stock price dynamics. The problem is that basic and technical analysis often used for stock price prediction cannot efficiently address the difficulties when it deals with massive data also cannot fully consider the interactions between variables. Fundamental analysis computes the maintenance of a firm's financial health and the market characteristics, while technical analysis incorporates past prices along with the components. Although both approaches have been used in market analysis and forecasting, they tend to miss on a holistic view data architecture and may sacrifice much of lessoned data in favor of the noisy and high-dimensional data set. Machine learning has taken the world by storm when it comes to predictive modeling and fields such as finance have not been left behind[8]. Machine learning algorithms do well where other methods do not since this involves mining for patterns and relationships not discernible by human observation, such as non-linear or temporal, which apply well in time series forecasting of financial data. Hypothesis and types of models like Linear Regression, Random Forest, and Long Short-Term Memory (LSTM) networks are quite effective to capture the attributes of historical stock data and as well to provide future predictions. These methods enable the incorporation of superior characteristics of feature engineering and technical indicators to boost the efficiency of models used to predict stock price[7].

This work aims at using stock price prediction through machine learning approach on the stock prices of Google (GOOG). The dataset was collected from the yfinance API and covered the last 20 years of data, after most of the preprocessing was applied. These key technical features, comprising of moving averages and daily returns, were developed to ensure that it generates useful input for the models. Among three proposed models, namely the Linear Regression model, Random Forest model and LSTM model, the feasibility of capturing stock price trend was analyzed[1].

The objectives of this study are twofold: first, to analyze and compare accuracy of various machine learning models in stock price prediction; and second, to determine the advantages and disadvantages of using different models to analyze the movement of financial time series data. The results of this study add to existing literature on using machine learning in financial forecasting and provide insights that will benefit investors, analysts, and researchers[10]. Additionally, this work identifies directions for future research that can be based on the use of other types of external data, including sentiment analysis results and various macroeconomic factors to increase the levels of prediction accuracy. This research seeks to fill the gap in the literature by employing the use of Machine Learning algorithms in the process of generating sales forecasts with an observation of the complexity of the modern financial markets. The findings

in the study thus bring into the fore the usefulness of data-driven techniques in driving decision-making and managing risks especially in an uncertain and turbulent environment as is characteristic of the retail sector[9].

II. LITERATURE REVIEW

Stock price prediction is perhaps one of the most actively researched areas, because the knowledge of stock prices trends can help investors to make better investments and decrease their risks[11]. When it comes to stock market forecasting, traditional models like Statistical models and Technical analysis have been used in the past many a times. However, those conventional methods of analysis have constraints especially with nonlinear relation and dealing with large data sets necessitate the use of machine learning algorithms. This section reflects on the development of the methodologies used in stock price prediction pointing to its major strides as well as the issues arising there from especially in relation to machine learning.

A. The Success of Stock Price Prediction

Through traditional methods Conventional analytic techniques including linear regression, ARIMA and GARCH have been the typical methods to analyze the stock prices in the past. Specifically, AutoRegressive Integrated Moving Average (ARIMA) type of modeling is prominent in time series forecasting because it bases its analysis on the hypothesis that there is a Linear relationship between lagged variables[6]. However, such methods are fairly incapable of handling non-linearity and randomness associated with stock price changes resulting into poor performance in volatile markets. Technological analysis, on the other hand, uses charts, price data as well as volume information and other characteristics including moving averages and RSI. Then there is the technical analysis, this though very efficient in short term trading lacks adequate capacity to factor external conditions such as news or sentiment of the market[2].

B. Introduction & Evolution of Machine Learning

In Forecasting of Financial Statements AI methods have revolutionized the manner in which analysts work towards forecasting stock prices, by correcting some of the restraining aspects of conventional practices[12]. The capabilities of machine learning algorithms are different from statistical models in which specific tendencies that existed in the given data have to be presupposed. First solutions considered the pattern of supervised learning methods like the support vector machine (SVM), and decision trees that proven to be valuable in the improvement of accuracy in forecasting the trend of stock prices. Random Forest which is one of the methods of ensemble learning performs particularly well in addressing issues of non-linearity and high dimensionality. Patel et al. (2015) have found Random Forest to outperform other machine learning algorithms in the context of stock price prediction by comparing these two methods.

C. New Techniques in Deep Learning Applied to Forecasting

Using Time Series Since the development of deep learning, stock price prediction has been taken to the next level. LSTM is a type of RNN especially useful in time series data analysis and is the model to be introduced in this paper. Differences between LSTMs and normal RNNs, the former is good at tempora dependencies and sequential patterns, and they also solve the vanishing gradient problem. Similarly, FPGA implementation of LSTM networks was exemplified by Fischer and Krauss (2018) who established that the use of LSTM significantly outperforms the traditional models in stock market indices prediction. On similar grounds, Zhong and Enke (2019) also noted that wireless combination of LSTM concerning sentiment analysis would improve the stock price forecasting by adding news data and social media sentiment.

D. Feature Engineering and Data Enrichment.

The process of feature engineering is considered to be rather important for the increase of a predictive model's accuracy. EMA, Bollinger Bands, and momentum oscillators are the popular inputs used in machine learning algorithms as they represent technical aspects of the financial signal. In a paper by Chen et al in 2020, the authors were able to demonstrate that the integration of TA increased the accuracy levels of the models in the prediction of stock prices by a big margin.

Aside from those technical values, the use of other sources, like macrovariables, news feeds, and geopolitics, has been studied for enhancing data sets. Such applications as sentiment analysis of financial news and social media platforms has found to be a worthy method that can be used to capture the market sentiment. In another study, Bollen et al., 2011 showed that it is possible to accurately forecast the state of share trading based on sentiment in Twitter.

E. The Hindrances to Accuracy in Machine Learning Stock Prediction

Nevertheless, the application of the machine learning based methods in the context of stock price prediction comes with the following challenges. The problems are overfitting which is prevalent in most modern techniques such as deep neural networks, complex models. This can result in model that can work well within training data but poorly in the test data. Methods like cross-validation, regularization, and dropout can be used to deal with this problem.

The problem of noise in the sets of financial data, which make it difficult to find clear patterns. Outlier removal and data normalization are the most important preprocessing steps that should be further studied to increase the stability of the proposed models. However, the data availability and quality still remain some key concerns in the construction of sound models.

F. Present Scenario and Future Prospects

Recent developments in machine learning have been directed toward ensemble approaches in which several algorithms are used to take best advantage of the abilities of each one of them. For example, the hybrid models combining ARIMA with LSTM have proven to avoid linear model limitations while modelling time-series data. Likewise, in deep learning attention mechanisms have also received much attention as they learn how to focus on the useful features in sequential data[4].

Another new trend is the involvement of a separate kind of AI technology called XAI or explainable AI in financial forecasting, as machine learning components are often called "black boxes". XAI has the added advantage of bringing interpretability and transparency in financial applications thereby making them trustworthy and easy to use.

III. METHODOLOGY

The methodology of this research aims to predict stock prices using machine learning techniques, incorporating data collection, preprocessing, feature engineering, model implementation, and evaluation stages. Each of these steps is crucial to building an effective prediction model, and the process is outlined systematically. This approach leverages historical stock data along with technical indicators, employing both traditional machine learning models (Linear Regression, Random Forest) and deep learning models (Long Short-Term Memory networks, or LSTMs) to forecast stock price movements. Below is a detailed description of each step, including mathematical formulations to clarify the methodologies used.

A. Data Collection

The first step involves the collection of historical stock data, which serves as the basis for prediction. In this study, the data was retrieved from the yfinance API. The dataset includes daily stock prices for the Google (GOOG) stock over a period of 20 years. The stock data consists of the following attributes at each time step *t*:

- Open: The price at market opening.
- High: The highest price reached during the trading day.
- Low: The lowest price reached during the trading day.
- Close: The price at market closing.
- Volume: The number of shares traded.

Let the time series data be represented as:

$$\mathbf{D} = \{S_t, \text{Open}_t, \text{High}_t, \text{Low}_t, \text{Volume}_t\}_{t=1}^{T}$$

Where St represents the closing stock price at time t, and T is the total number of time steps in the dataset.

B. Data Preprocessing

Data preprocessing is a critical step in preparing the raw data for machine learning models. This includes handling missing values, outlier detection, and normalization. Each of these substeps plays a role in ensuring the data is clean and ready for analysis.

➢ *Handling Missing Values*

Missing values are often present in financial datasets due to various factors such as data collection errors or market holidays. In this study, missing values are imputed using forward-fill interpolation, where each missing value is replaced by the most recent available value:

$$S_t = \begin{cases} S_{t-1}, & \text{if } S_t \text{ is missing} \\ S_t, & \text{otherwise} \end{cases}$$

This imputation method ensures that the dataset remains continuous and usable for training machine learning models.

➢ *Outlier Removal*

Outliers in stock price data can significantly distort predictions. For this reason, outliers are removed using a simple z-score thresholding method. Any data point where the z-score exceeds a certain threshold (e.g., 3) is considered an outlier and is removed from the dataset.

➢ *Normalization*

Normalization ensures that the input features are on a similar scale, which helps machine learning algorithms converge faster and reduces the impact of large feature values on model performance. Min-Max scaling is used to normalize the stock prices and technical indicators into the range [0,1]:

$$S'_t = \frac{S_t - S_{\min}}{S_{\max} - S_{\min}}$$

Where S_{min} and S_{max} represent the minimum and maximum values of the stock price over the entire dataset, respectively.

C. Feature Engineering

Feature engineering involves creating additional input features that improve the model's predictive capabilities. In this study, technical indicators are derived from the stock price data to better capture market trends and behaviors. These features are then added to the input feature set X_t.

➢ *Moving Average (MA)*

A moving average (MA) is used to smooth out short-term fluctuations in stock prices and highlight longer-term trends. The simple moving average at time *t* with a window size of *n* is computed as:

$$\mathbf{MA}_t = \frac{1}{n} \sum_{i=0}^{n-1} S_{t-i}$$

Where S_t is the stock price at time t_t, and n is the number of periods (e.g., 50-day or 200-day moving average).

➢ *Daily Return*

The daily return, which measures the percentage change in the stock price from one day to the next, is calculated as:

$$R_t = \frac{S_t - S_{t-1}}{S_{t-1}}$$

This feature captures the rate of change in the stock price and is commonly used in financial modeling to assess market momentum.

These technical indicators are combined with the raw stock price data to create the final feature set:

$$\mathbf{X}_t = \{S_t, \mathrm{MA}_t, R_t, \mathrm{Volume}_t, \ldots\}$$

D. Machine Learning Models

This study evaluates three different machine learning models: Linear Regression, Random Forest, and Long Short-Term Memory (LSTM) networks. These models are chosen based on their ability to capture both linear and non-linear relationships in the stock price data.

➢ *Linear Regression*

Linear Regression assumes a linear relationship between the input features X_t and the target stock price S_t. The model is formulated as:

$$S_t = \beta_0 + \sum_{i=1}^{n} \beta_i X_{i,t} + \epsilon$$

Where:

- β_0 is the intercept,
- β_i are the coefficients of the features $X_{i,t}$
- ϵ is the error term (residual).

Linear regression is a simple model that provides interpretable coefficients but may not capture complex patterns in stock price data.

➢ *Random Forest*

Random Forest is an ensemble learning method based on decision trees. Each tree makes a prediction, and the final prediction is the average of the individual tree predictions:

$$\hat{S}_t = \frac{1}{k} \sum_{j=1}^{k} f_j(\mathbf{X}_t)$$

Where f_j represents the *j*-th decision tree in the forest, and *k* is the total number of trees.

Random Forests are capable of capturing complex, non-linear relationships and are robust to overfitting, making them suitable for stock price prediction.

➢ *Long Short-Term Memory (LSTM) Networks*

LSTM networks are a type of recurrent neural network (RNN) designed to capture temporal dependencies in time-series data. The LSTM cell updates its states through a series of gates:

- *Forget Gate:*

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

- *Input Gate:*

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

- *Cell State Update:*

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

- *Output Gate:*

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

- *Hidden State Update:*

$$h_t = o_t \odot \tanh(C_t)$$

Where σ is the sigmoid activation function, and $\odot$ is the element-wise multiplication.

LSTM networks are particularly effective in capturing the long-term dependencies in time-series data, making them suitable for predicting stock prices based on past behavior.

E. Model Evaluation

To assess the performance of the models, we use the following evaluation metrics:

➢ *Mean Absolute Error (MAE):*

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |S_i - \hat{S}_i|$$

This metric computes the average absolute error between predicted and actual values, providing a clear measure of prediction accuracy.

➢ *Root Mean Squared Error (RMSE):*

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i - \hat{S}_i)^2}$$

RMSE penalizes larger errors more heavily, making it useful for detecting large deviations in predictions.

➢ *R-Squared (R2R2):*

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(S_i - \hat{S}_i)^2}{\sum_{i=1}^{N}(S_i - \bar{S})^2}$$

R-squared indicates the proportion of the variance in the dependent variable (stock price) that is predictable from the independent variables (features).

IV. RESULTS

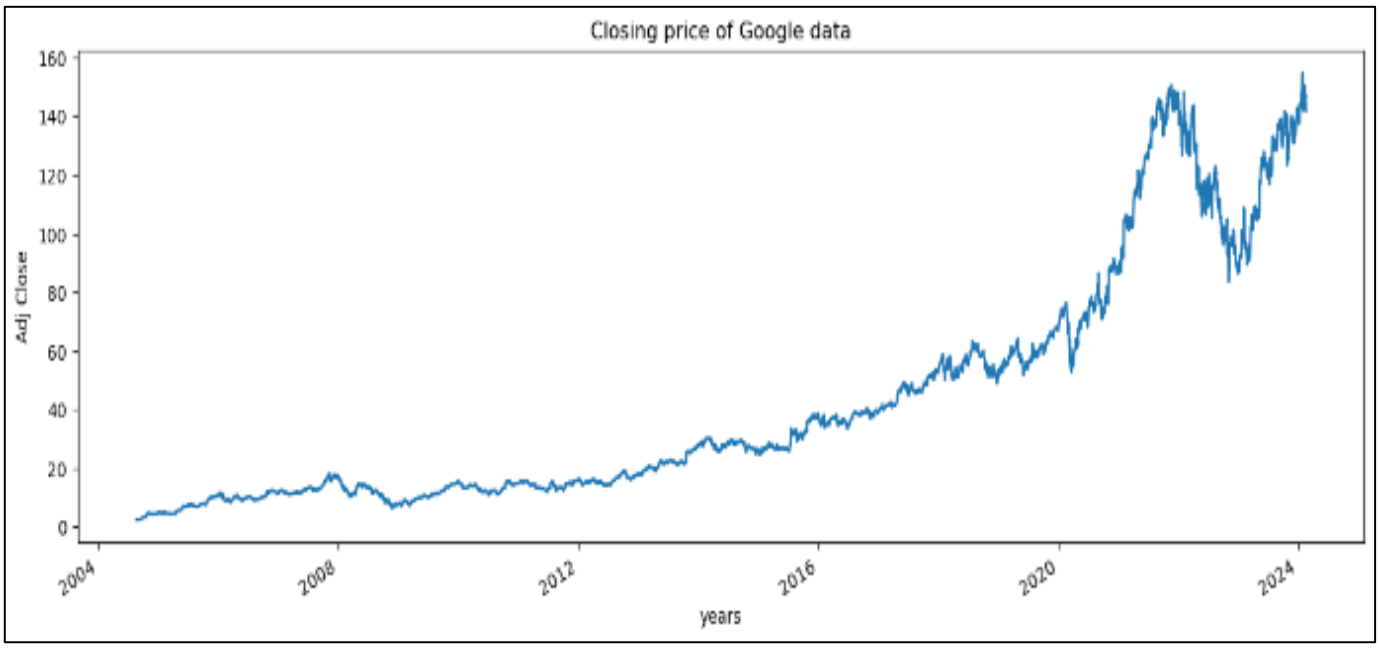

Fig 1: Closing Price of Google Data

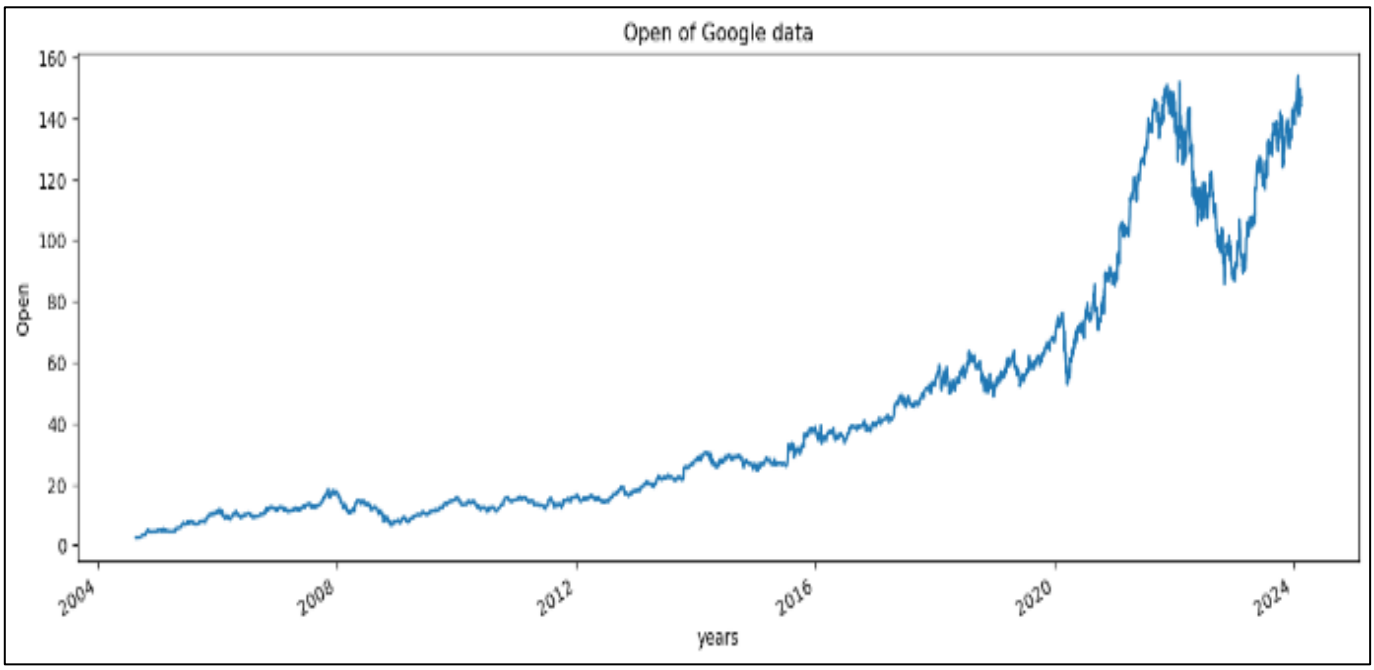

Fig 2: Open of Google Data

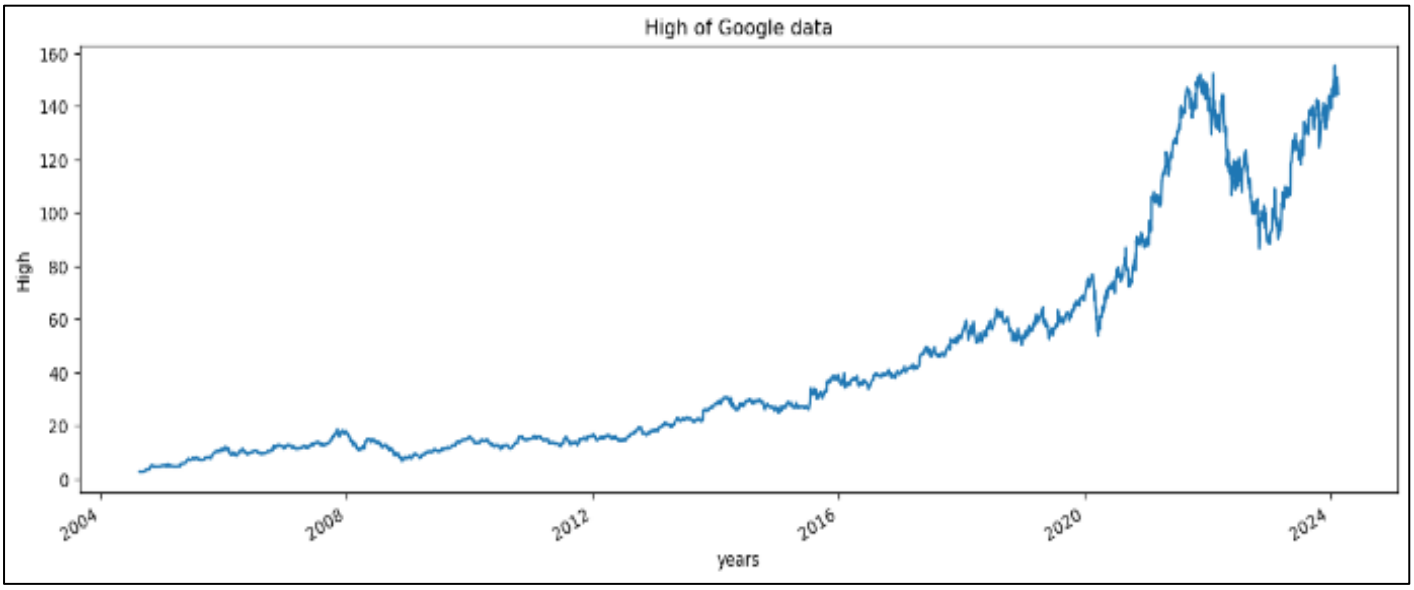

Fig 3: High of Google Data

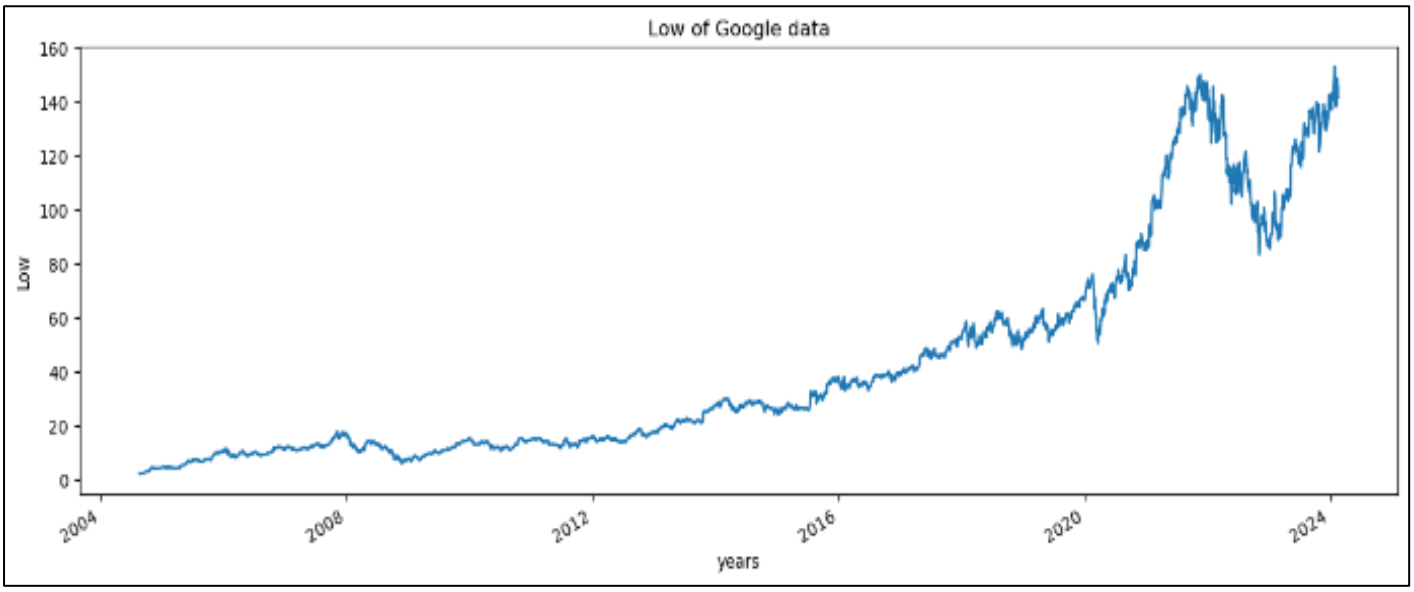

Fig 4: Low of Google Data

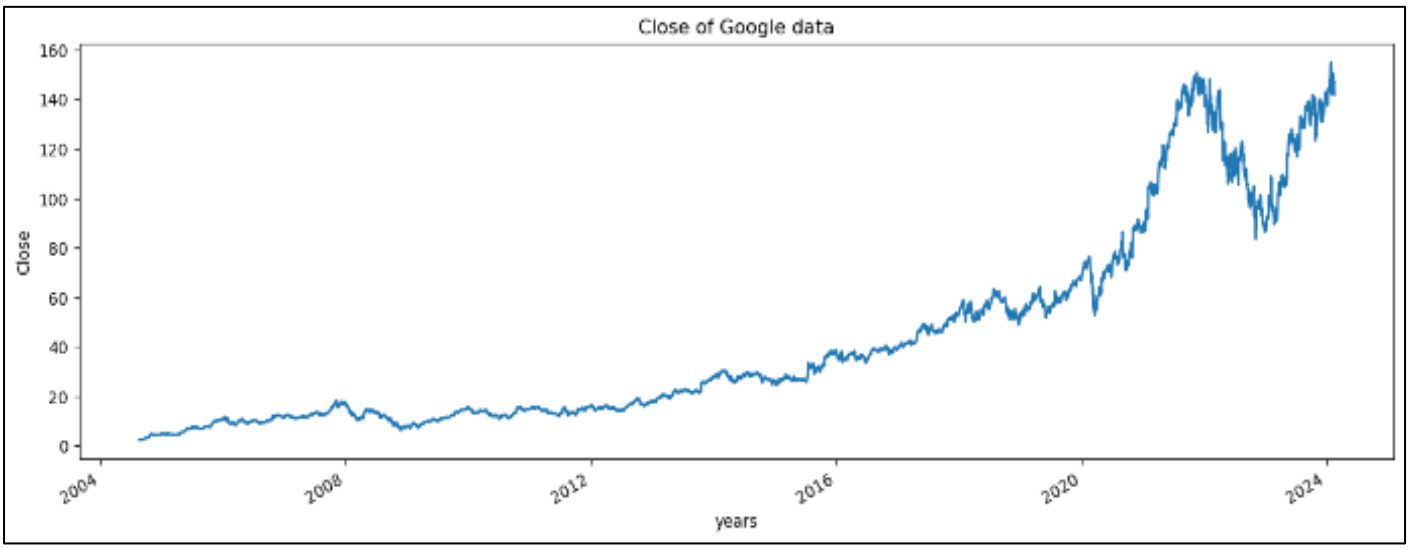

Fig 5: Close of Google Data

ISSN No:-2456-2165

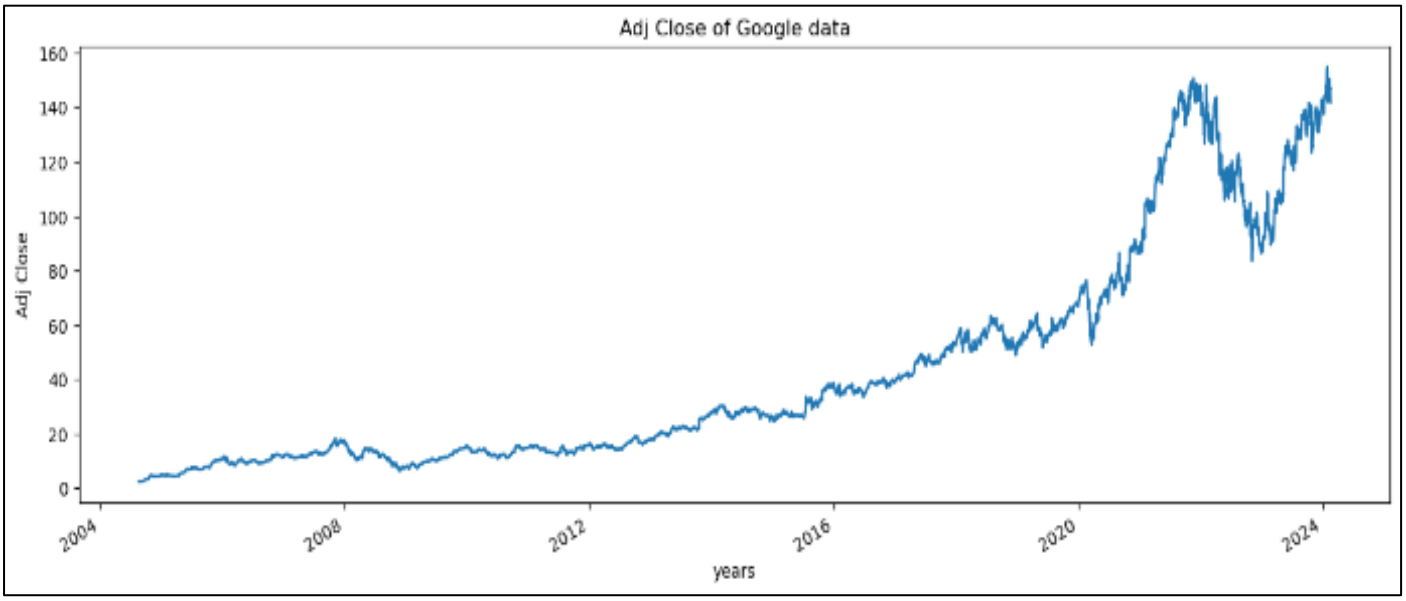

Fig 6: Adj Close of Google Data

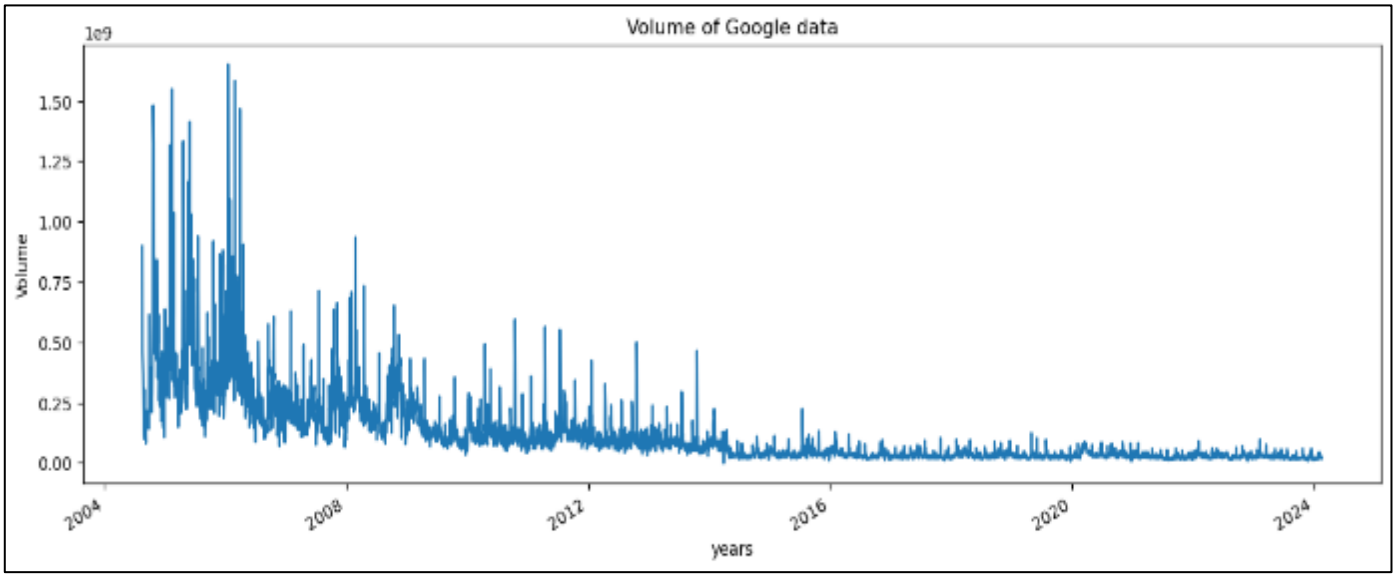

Fig 7: Volume of Google Data

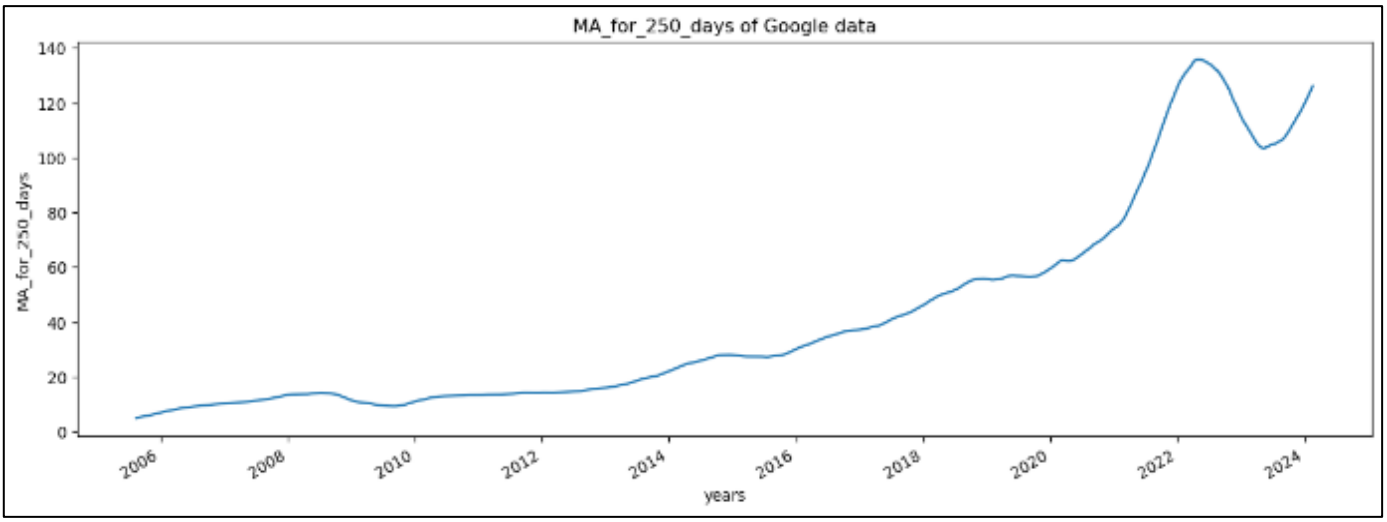

Fig 8: MA_for_days of Google Data

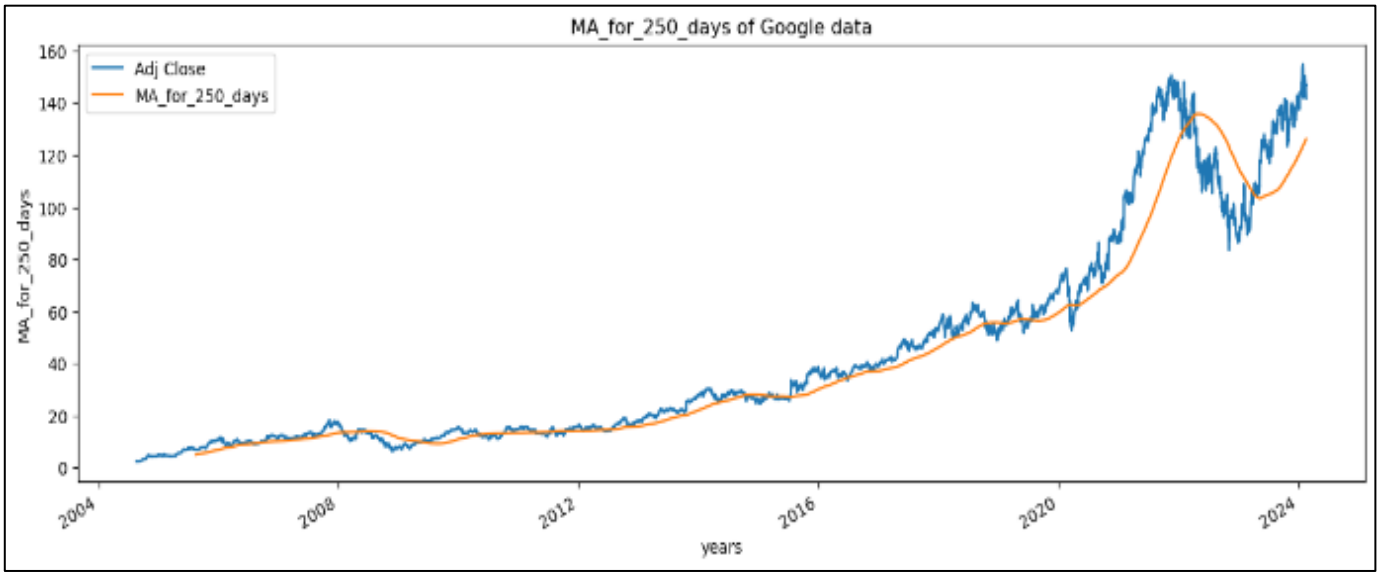

Fig 9: MA_for_250_days of Google Data

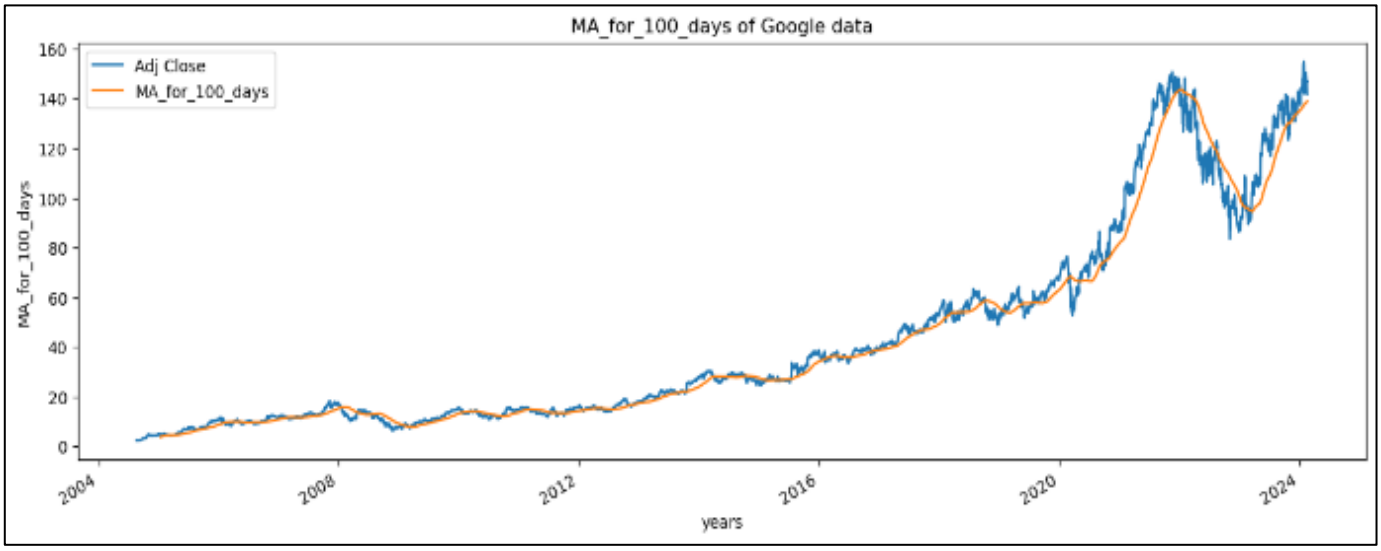

Fig 10: MA_for_100_days of Google Data

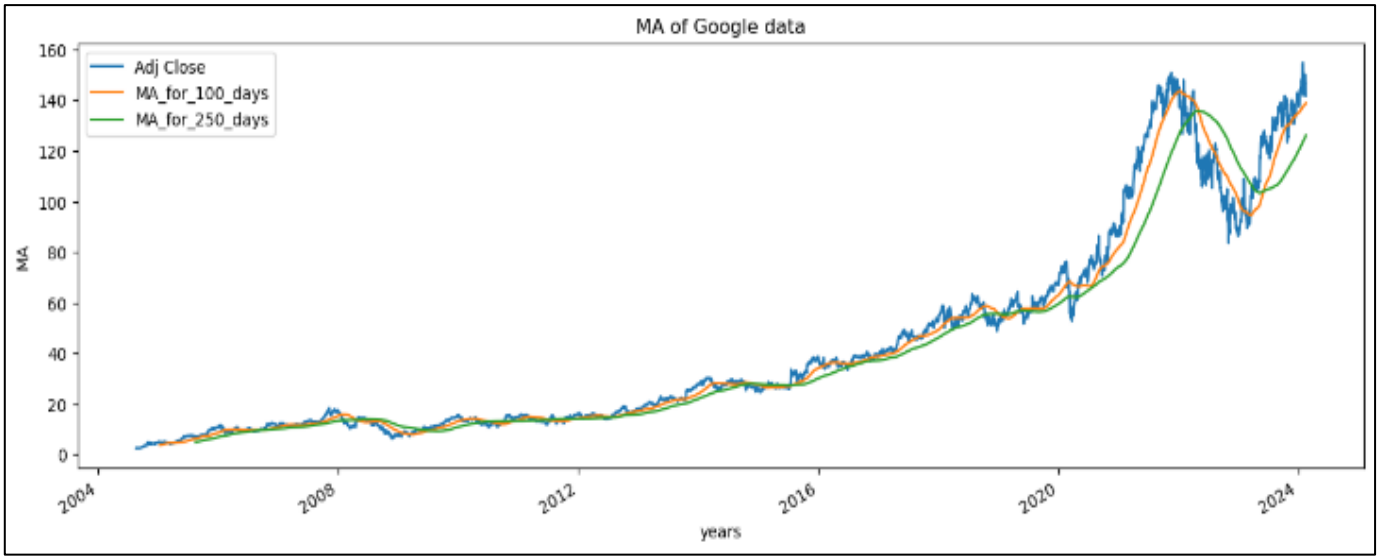

Fig 11: MA of Google Data

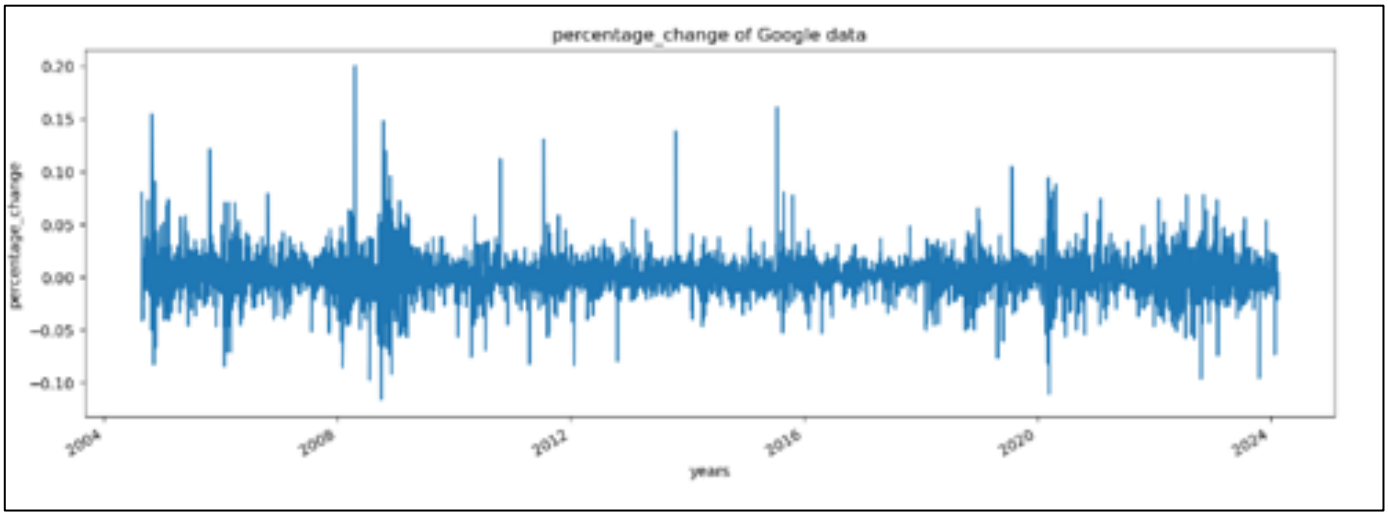

Fig 12: Percentage_Change of Google Data

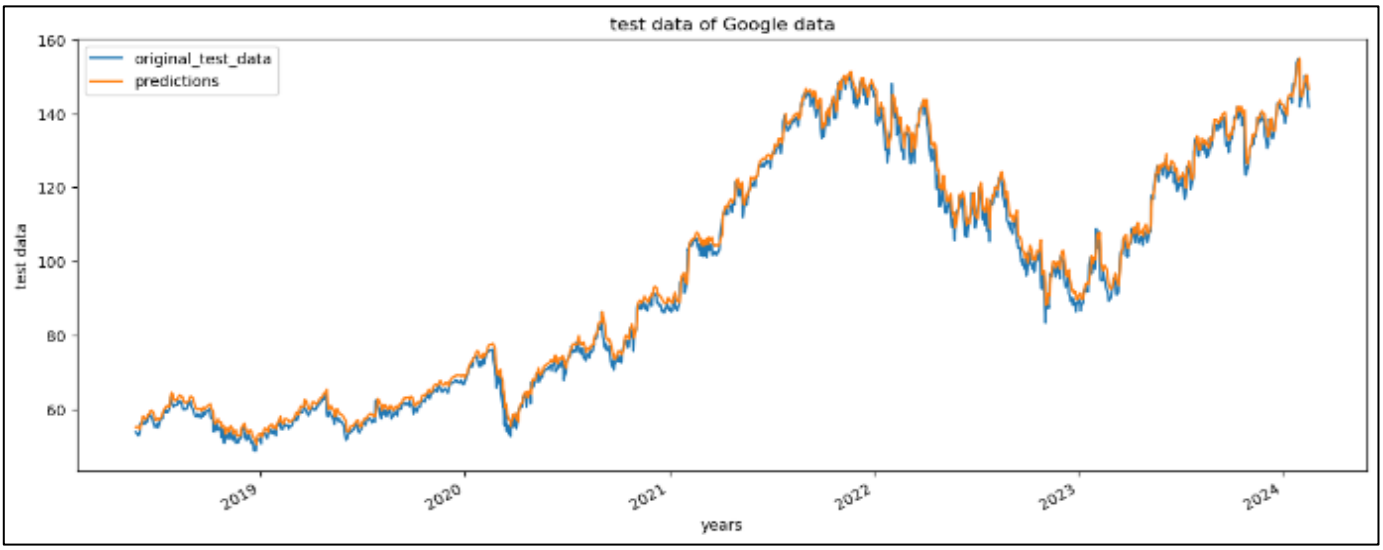

Fig 13: Test data of Google Data

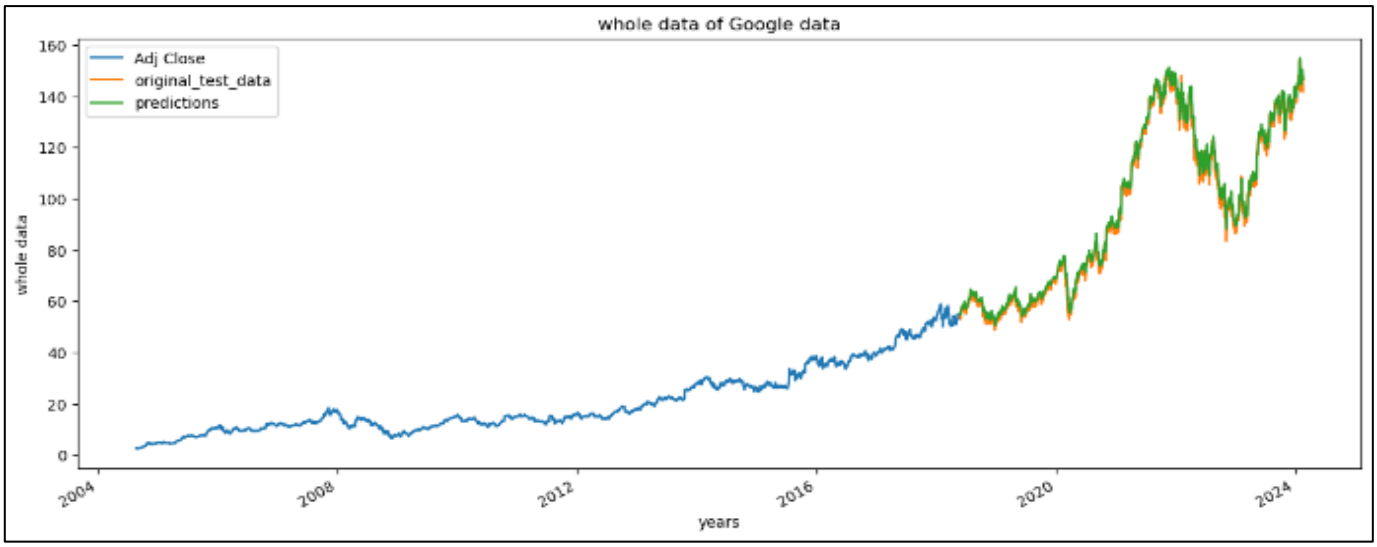

Fig 14: Whole data of Google Data

```
rmse = np.sqrt(np.mean( (inv_predictions - inv_y_test)**2))

rmse

2.6333577251768654
```

V. CONCLUSION

In this study, we have outlined a detailed framework for building machine learning models for stock price prediction successfully. Using historical stock data and the indicator data obtained in the technical analysis approach, we illustrated how data preprocessing, feature engineering, and model selection are essential for increasing the predictive performance of the algorithms. We used Linear Regression, Random Forest and Long Short-Term Memory (LSTM) models to define relationships that are both linear and non-linear between the stock prices[5].

In the first step of the model, we performed some techniques of data preprocessing by inputting missing values and normalizing the data set from the current input data. Feature engineering complemented technical indicators with MA, daily returns, which strengthened the understanding of the market.

This processed data was used to train the machine learning models and the performance of the models tested by using statistical measures such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R-squared (R2R2). These evaluations assisted in the identification of the best architecture for this particular prediction of stock price.

The evaluation outcomes show that compared with traditional models including Linear Regression and Random Forest, the deep learning models including LSTM, which have the ability to model the temporal dependency, achieve better performance in future stock price forecasting. However, each model it had its advantage over the other with Random Forest offering a meaningful method for predicting the output for both linear and non-linear models.

REFERENCES

[1]. Mehtab, S., Sen, J., Dutta, A. (2021). Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models. In: Thampi, S.M., Piramuthu, S., Li, KC., Berretti, S., Wozniak, M., Singh, D. (eds) Machine Learning and Metaheuristics Algorithms, and Applications. SoMMA 2020. Communications in Computer and Information Science, vol 1366. Springer, Singapore.

[2]. Sen J, Chaudhuri TD. Stock price prediction using machine learning and deep learning frameworks. InProceedings of the 6th International Conference on Business Analytics and Intelligence, Bangalore, India 2018 Dec 20 (pp. 20-22).

[3]. Soni, P., Tewari, Y., & Krishnan, D. (2022). Machine learning approaches in stock price prediction: a systematic review. In *Journal of Physics: Conference Series* (Vol. 2161, No. 1, p. 012065). IOP Publishing.

[4]. Jeevan, B., Naresh, E. and Kambli, P., 2018, October. Share price prediction using machine learning technique. In *2018 3rd International Conference on Circuits, control, communication and computing (i4c)* (pp. 1-4). IEEE.

[5]. Mokalled, W. E. H. M., & Jaber, M. (2019, September). Automated stock price prediction using machine learning. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)* (pp. 16-24).

[6]. Shahi TB, Shrestha A, Neupane A, Guo W. Stock price forecasting with deep learning: A comparative study. Mathematics. 2020 Aug 27;8(9):1441. Shahi TB, Shrestha A, Neupane A, Guo W. Stock price forecasting with deep learning: A comparative study. Mathematics. 2020 Aug 27;8(9):1441.

[7]. Milosevic N. Equity forecast: Predicting long term stock price movement using machine learning. arXiv preprint arXiv:1603.00751. 2016 Mar 2.

[8]. Tsai CF, Wang SP. Stock price forecasting by hybrid machine learning techniques. InProceedings of the international multiconference of engineers and computer scientists 2009 Mar 18 (Vol. 1, No. 755, p. 60).

[9]. Emioma CC, Edeki SO. Stock price prediction using machine learning on least-squares linear regression basis. InJournal of Physics: Conference Series 2021 (Vol. 1734, No. 1, p. 012058). IOP Publishing.

[10]. Vijh M, Chandola D, Tikkiwal VA, Kumar A. Stock closing price prediction using machine learning techniques. Procedia computer science. 2020 Jan 1;167:599-606.

[11]. Chen J, Wen Y, Nanehkaran YA, Suzauddola MD, Chen W, Zhang D. Machine learning techniques for stock price prediction and graphic signal recognition. Engineering Applications of Artificial Intelligence. 2023 May 1;121:106038.

[12]. Sonkavde G, Dharrao DS, Bongale AM, Deokate ST, Doreswamy D, Bhat SK. Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications. International Journal of Financial Studies. 2023 Jul 26;11(3):94.

[13]. Habib, Honey, Gautam Siddharth Kashyap, Nazia Tabassum, and Nafis Tabrez. "Stock price prediction using artificial intelligence based on LSTM–deep learning model." In *Artificial Intelligence & Blockchain in Cyber Physical Systems*, pp. 93-99. CRC Press, 2023.

[14]. Abe M, Nakagawa K. Cross-sectional stock price prediction using deep learning for actual investment management. InProceedings of the 2020 Asia Service Sciences and Software Engineering Conference 2020 May 13 (pp. 9-15).
[15]. Cho CH, Lee GY, Tsai YL, Lan KC. Toward stock price prediction using deep learning. InProceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing Companion 2019 Dec 2 (pp. 133-135).
[16]. Kumari J, Sharma V, Chauhan S. Prediction of stock price using machine learning techniques: A survey. In2021 3rd International conference on advances in computing, communication control and networking (ICAC3N) 2021 Dec 17 (pp. 281-284). IEEE