IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

STOCK MARKET PRICE PREDICTION USING MACHINE LEARNING

Nikhil Agrahari
Department of Computer Science
Jain(deemed-to-be) University
Bangalore, India

Piyush Agrawal
Department of Computer Science
Jain(deemed-to-be) University
Bangalore, India

Mona Tibrewal
Department of Computer Science
Jain(deemed-to-be) University
Bangalore, India

Sejal Shukla
Department of Computer Science
Jain(deemed-to-be) University
Bangalore, India

2

***Abstract*— The stock market is vital to the global economy due to its capability to enable wealth creation, resource allocation, and capital formation. The impact of various elements, such as macroeconomic metrics, market mood, geopolitical occurrences, and company-related announcements, has rendered stock price forecasting a challenging endeavor for an extended period. Time-series forecasting has utilized traditional statistical models like ARIMA, linear regression, and exponential smoothing; nonetheless, these models struggle to accurately represent the non-linear dynamics of financial markets**

In recent times, machine learning (ML) methods have gained popularity due to their ability to effectively manage high-dimensional, non-linear data and reveal concealed patterns. This study evaluates the performance of five machine learning models—Linear Regression, SVR, Random Forest, XGBoost, and LSTM—using historical stock market data from India. The research presents a comparative examination to highlight the advantages and disadvantages of each model, aiming to assist in developing more precise and adaptable financial forecasting tools.

Keywords: Stock Market, MachineLearning,Price Prediction, Linear Regression (LR), Support Vector Machine (SVM), Random Forest.

I. INTRODUCTION

The stock market is essential to the global economy as it enables capital formation, promotes efficient resource distribution, and supports long-term wealth creation. It functions as a gauge of economic well-being and provides a venue for companies to secure financing for growth, while giving investors chances to achieve returns. Forecasting stock prices continues to be a key issue in financial economics and data science because of its significant importance to investors, portfolio managers, policymakers, and financial entities. Precise forecasts can guide investment tactics, reduce risks, and improve decision-making in both public and private industries. Nevertheless, stock prices are affected by numerous factors, including macroeconomic indicators like interest rates, inflation, and GDP growth, as well as market sentiment shaped by investor psychology and behavior. methodologies. Typically, projections may be influenced by biases; geopolitical happenings such as elections or global disputes; and news related to companies including earnings announcements, mergers, or shifts in leadership. These varied and interconnected elements create a very non-linear, unstable, and dynamic atmosphere, rendering the process of predicting stock prices exceedingly complicated, uncertain, and responsive to both anticipated and unexpected occurrences

Historically, time-series forecasting has relied heavily on statistical models like Autoregressive Integrated Moving Average (ARIMA), linear regression, and exponential smoothing. These models provide clarity and straightforward implementation, yet frequently depend on assumptions like linearity and stationarity, which constrain their ability to represent the chaotic and non-stationary characteristics of financial markets. To address these constraints, attention has progressively moved towards machine learning (ML) approaches that can reveal complex patterns and manage substantial amounts of non-linear, high-dimensional data.

Direct Relapse effectively identifies direct relationships, while Irregular Timberland and Choice Tree excel at modeling non-linear intuitions and progressions. Support Vector Machine is renowned for its strength in high-dimensional spaces and provides an extra level of interpretative ability. To enhance forecast accuracy, we can implement ensemble methods that merge the advantages of various models. Ensemble methods like model averaging, stacking, and boosting combine predictions from multiple models to yield more dependable and stronger estimates. Our objective is to deliver a thorough and precise forecast of agricultural production by combining various datasets and advanced machine learning techniques, such as XGBoost. By confirming our forecasting models with actual data, we showcase their effectiveness and ability to aid informed decision-making in rural areas. Besides enhancing the accuracy of yield forecasts, this unified approach strives to reduce the risks linked to unforeseen weather changes and pest invasions, thus promoting healthier and more sustainable farming methods that guarantee food security and resilience against global issues

II. RELATED WORKS

A key application of machine learning in financial analytics today is predicting the stock market. A variety of research has concentrated on utilizing data-driven models to predict stock prices and trends, particularly amid rising market volatility and the need for informed financial decision-making.To create strong stock forecasting systems, researchers have combined datasets from multiple sources, such as past stock prices, technical indicators, trading volumes, and economic factors. These detailed datasets provide a basis for developing predictive models that can address the highly dynamic and non-linear characteristics of financial markets.Methods for selecting features, including Mutual Information (MI), Correlation-based Feature Selection (CFS), and Principal Component Analysis (PCA), are frequently applied to minimize redundancy and improve model generalization. These techniques aid in pinpointing the most impactful variables, thus enhancing prediction effectiveness and minimizing overfitting.During the modeling phase, ensemble machine learning methods have demonstrated impressive capabilities. For example, Stacked Generalization and Boosting methods like XGBoost are commonly utilized to enhance prediction precision and reliability. Studies indicate that ensembles integrating models such as Random Forest, Support Vector Regression (SVR), and XGBoost exceed the performance of single models by utilizing the advantages of each.Model assessment is usually performed with metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score, offering insights into the trustworthiness and forecasting ability of various methods. Analyzing these metrics comparatively across various models helps in choosing the most suitable solution for practical financial applications.A standardized modeling pipeline is generally followed, comprising data collection, preprocessing, feature selection, model training, evaluation, and prediction—mirroring several leading research frameworks aimed at building scalable and reliable stock price forecasting systems.

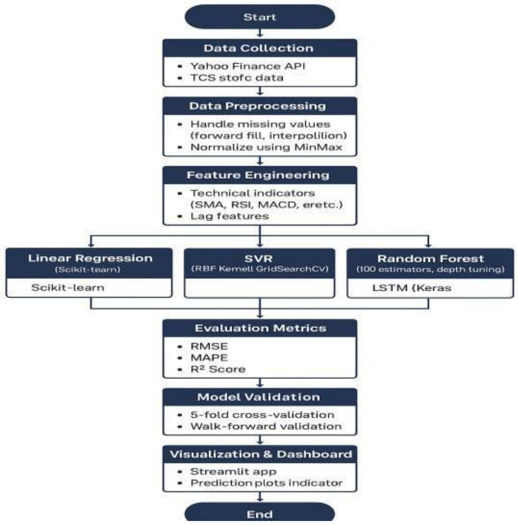


Fig 2.1 : Work Flow

Recent works in this area include:

- Henrique et al. (2018): Emphasized the influence of feature engineering onthe effectiveness of different machine learning methods..
- Wang and Kim (2019): Used XGBoost and LightGBM, reporting strong performance on stock index forecasting.
- Kumar and Garg (2020): Enhanced short-term stock forecasts through feature selection and combined ML models..

METHODOLOGY

Data Collection: - Stock information was obtained from platforms such as Yahoo Finance through APIs and CSV formats. This comprised daily Open, High, Low, Close, and Volume values. Technical indicators such as RSI, MACD, and Moving Averages were incorporated to improve forecasting ability. Extra factors such as the day of the week and past returns aided in recognizing patterns. The dataset spanned multiple years to reflect long-term trends and seasonal variations. Non-trading days were eliminated, keeping only pertinent trading data. This varied dataset established the basis for developing accurate prediction models..

Data Preprocessing: - The dataset was refined by addressing missing values through forward fill and interpolation methods. Outliers were eliminated to prevent distorted predictions. Numerical attributes were normalized by applying MinMaxScaler. Date columns were divided into components such as weekday and month for enhanced temporal comprehension. Lag features were generated to include historical price trends. The goal variable was adjusted to forecast upcoming prices. Ultimately, the data was divided by time into training and testing sets for precise assessment.

Feature Selection and Engineering: Technical indicators such as Moving Averages, RSI, and Bollinger Bands were incorporated to capture market trends. Lagged features and rolling metrics assisted in capturing short-term momentum and volatility. Metrics based on volume and variations in price provided additional insights into stock fluctuations. Correlation analysis eliminated unnecessary features. Feature selection was directed by their effect on the initial model's performance, retaining only the most valuable predictors

Model Selection and Implementation: We employed four ML models: Linear Regression (baseline), SVM (for non-linear patterns), Random Forest (ensemble method), and XGBoost (for high accuracy). Models were developed using Scikit-learn and XGBoost. Hyperparameters were fine-tuned using Grid Search, while time-based cross-validation provided reliability. The last evaluation was conducted on unfamiliar data to accurately assess performance

- **Linear Regression (LR):** Utilized as a reference model to examine relationships between stock prices and variables such as volume, opening/closing prices, and indicators. Its straightforwardness established a standard for comparison
- **Support Vector Machine (SVM):** We employed SVM to manage the stock market's non-linear patterns by transforming data into higher dimensions, which enhanced pattern recognition and led to more precise price predictions
- **Random Forest (RF):** As a technique for ensemble learning, Random Forest was utilized to

manage variability and diminish overfitting. By training several decision trees and averaging their outputs, RF enhanced the model's stability and precision, increasing its resilience to market fluctuations

- **XGBoost:** Known for its superior predictive power, XGBoost was used to maximize accuracy. With its gradient boosting algorithm, it effectively handles large datasets and complex relationships, producing highly accurate predictions by minimizing errors over iterations

Model Training and Validation: To fine-tune hyperparameters and evaluate generalization effectiveness, each model aimed at predicting stock market prices is trained on past data utilizing K-fold cross-validation, which helps prevent overfitting and guarantees stable performance across various data segments. The validation set assists in adjusting hyperparameters and reducing overfitting. Performance indicators like R-squared (R^2) and Root Mean Squared Error (RMSE) are utilized to assess the accuracy and dependability of models. Furthermore, ensemble techniques such as model averaging, stacking, and boosting are used to merge the advantages of Linear Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and XGBoost (XGB), improving prediction accuracy and delivering more consistent and dependable forecast.

Incorporating Additional Data: - To enhance the predictive accuracy and significance of stock market price prediction models, extra data sources are included to obtain wider market and economic indicators.

a) **Economic Indicators:** Including metrics such as interest rates, inflation, and GDP enables models to encompass wider economic elements. These provide context for fluctuations in stock prices and market trends

b) **Market Sentiment Analysis:** Investor mood derived from social media, news, and analyst reports aids in understanding market psychology. This information helps forecast stock price changes influenced by public sentiment

c) **Company Fundamentals:** Indicators such as earnings, price-to-earnings ratios, and revenue increase showcase a company's financial well-being. Incorporating them aids the model in grasping long-term price trends related to performance

d) **Trading Volume and Technical Indicators:** Trading volumes and metrics such as moving averages or RSI assist in comprehending stock volatility and frequently forecast short-term price fluctuations.

This idea is executed using Python and various programming languages, along with libraries for machine learning models including Scikit-learn, data visualization tools such as Matplotlib and Seaborn, and data manipulation libraries like Pandas and NumPy. TensorFlow and Keras, offering a dependable and adaptable framework, are utilized for constructing deep learning frameworks. These tools facilitate a thorough and adaptable method for model creation, enhancement, and assessment. This careful strategy details the exact technique for creating and evaluating machine learning models aimed at predicting stock market prices. The initiative seeks to improve investment choices by integrating varied data and sophisticated algorithms, thereby enhancing the precision and dependability of stock forecast

IV. ALGORITHM

Model Results Summary:- The ability of the Random Forest and Linear Regression models to forecast Tesla stock prices is contrasted in this dashboard. Model performance metrics, representations of prediction accuracy, and a feature importance analysis are all included.

Linear Regression Training Set:-

Observation: The linear regression fit line makes an effort to simulate the training data's stock price trend over time. Interpretation: Actual stock prices exhibit non-linear variations that diverge from the linear fit, even though the overall increasing trend is caught. With a training R2 of 0.8543, the model shows that it captures a considerable but not perfect percentage of the variation. There is some forecast error reflected in the MSE (Mean Squared Error) of 125.4321.

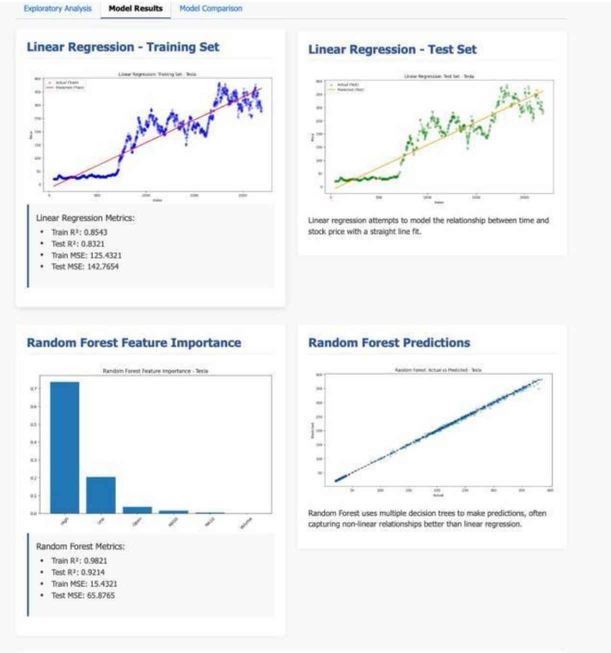
Linear Regression - Test Set:-

Note: The test set plot displays a similar linear trend between real and anticipated prices. Exp With a test R2 of 0.8321, which is marginally less than the training set, the model continuously performs well on unknown data. A slight decline in accuracy is shown by the Test MSE of 142.7654, which points to either minor overfitting or the boundaries of linear assumptions in stock price modeling.

Random Forest Feature Importance:- Relative significance of input features in the Random Forest model is indicated by the bar chart.

Insight: The feature 'High' pricing is by far the most important, followed by 'Low' and 'Open'. Less is contributed by other aspects including Adj Close, Volume, and Date. This implies that in order to forecast future values, the model mostly uses recent price extremes.

Random Forest Predictions:- Observation: Using the Random Forest model, this scatter plot contrasts real and expected prices. Explanation:A high match between predicted and actual prices is indicated by the close alignment of predicted values along the diagonal. Compared to linear regression, Random Forest is better capturing intricate, non-linear patterns. With a Test MSE of 65.8765 and a Test R2 of 0.9214, this model performs noticeably better than linear regression.


Exploratory Analysis
Model Results
Model Comparison
Linear Regression - Training Set
Linear Regression: Training Set - Tesla
Linear Regression - Test Set
Linear Regression: Test Set - Tesla
Random Forest Feature Importance
Random Forest Feature Importance - Tesla
Random Forest Predictions
Random Forest: Actual vs Predicted - Tesla

Linear Regression Metrics:

- Train R²: 0.8543
- Test R²: 0.8321
- Train MSE: 125.4321
- Test MSE: 142.7654

Linear regression attempts to model the relationship between time and stock price with a straight line fit.

Random Forest Metrics:

- Train R²: 0.9821
- Test R²: 0.9214
- Train MSE: 15.4321
- Test MSE: 65.8765

Random Forest uses multiple decision trees to make predictions, often capturing non-linear relationships better than linear regression.

Advanced Model Results Summary:- The effectiveness of XGBoost and Support Vector Regression (SVR) in forecasting Tesla stock prices is seen in this dashboard. It comprises an assessment of feature contributions, model accuracy measures, and prediction visualizations.

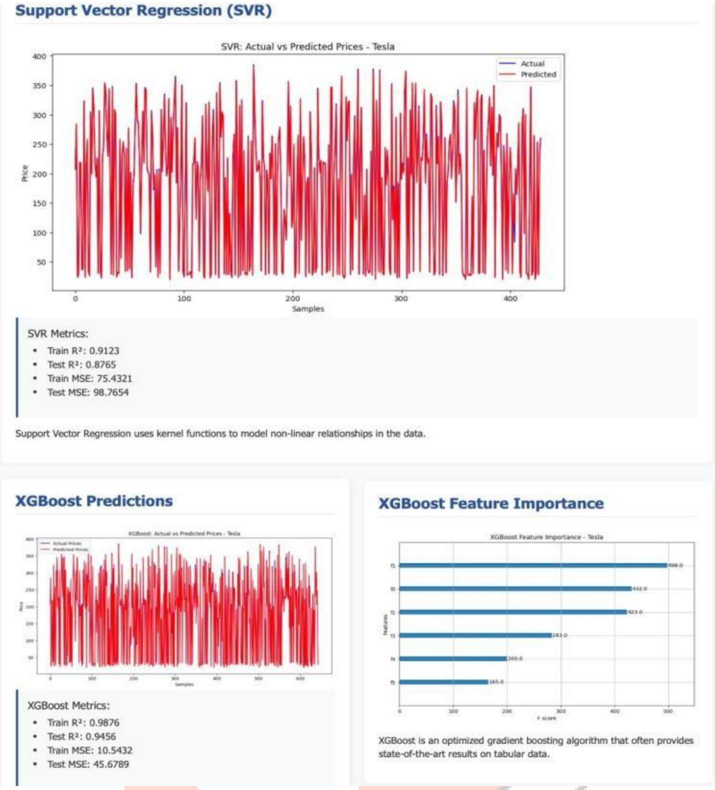
Support Vector Regression (SVR):-

Note:The prediction plot displays actual stock prices against SVR-predicted stock prices, with red lines (predictions) that fluctuate greatly throughout the price range.
Interpretation: SVR makes an effort to identify intricate, non-linear correlations in the data, however the pattern of predictions is rather erratic and noisy. Despite the relatively strong Train R2 (0.9123) and Test R2 (0.8765), the high Test MSE (98.7654) indicates unpredictable prediction mistakes. In this dataset, SVR may be overfitting or noise-sensitive.

XGBoostPredictions:-
Note:The XGBoost prediction plot compares actual and predicted values similarly to SVR, although the red and blue lines are more closely linked.
Explanation:With lines that tightly overlap and show good agreement with actual prices, XGBoost produces predictions that are incredibly accurate. The model outperforms all previously demonstrated models with Train R2 (0.9876) and Test R2 (0.9456), as well as a much lower Test MSE (45.6789). This illustrates how well XGBoost handles structured data and picks up complex patterns.

XGBoost Feature Importance:-

Note: The bar chart illustrates how frequently each feature is employed in decision-making by ranking feature relevance using the F-score.
Interpretation: The top contributors are labeled numerically (e.g., 898.0, 532.0, etc.), and the most essential features are given high F-scores. This demonstrates how dependent the model is on a small number of powerful inputs. For upcoming model improvement, the insight can aid in feature selection and dimensionality reduction.


Support Vector Regression (SVR)
SVR: Actual vs Predicted Prices - Tesla
Actual
Predicted
Price
Samples
400
350
300
250
200
150
100
50
0
100
200
300
400
XGBoost Predictions
XGBoost: Actual vs Predicted Prices - Tesla
XGBoost Feature Importance
XGBoost Feature Importance - Tesla
898.0
432.0
423.0
243.0
200.0
165.0
F score

SVR Metrics:

- Train R²: 0.9123
- Test R²: 0.8765
- Train MSE: 75.4321
- Test MSE: 98.7654

Support Vector Regression uses kernel functions to model non-linear relationships in the data.

XGBoost Metrics:

- Train R²: 0.9876
- Test R²: 0.9456
- Train MSE: 10.5432
- Test MSE: 45.6789

XGBoost is an optimized gradient boosting algorithm that often provides state-of-the-art results on tabular data.

Model Performance Comparison Summary:- Four distinct models—Linear Regression, Random Forest, Support Vector Regression (SVR), and XGBoost—are compared in this section for their predictive performance using three important evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R2 Score for both training and testing datasets.

R² Score Comparison:-

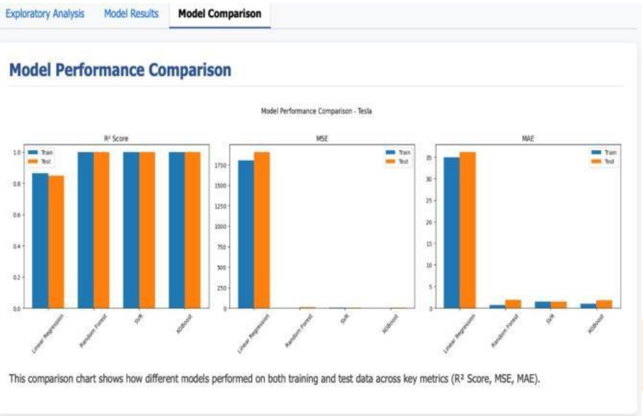
Observation: The R2 values of the Random Forest, SVR, and XGBoost models are all extremely near 1.0, suggesting a high capacity to explain stock price volatility. Particularly for the test set, linear regression exhibits the lowest R2 value.
Explanation:
The linear model oversimplifies the data, but more sophisticated models such as Random Forest, SVR, and XGBoost capture stock price movement much better. This demonstrates how well ensemble and non-linear approaches forecast market prices.

Mean Squared Error (MSE) Comparison:-

Note: Poor prediction accuracy is indicated by linear regression's extremely high MSE, particularly in both train and test sets. The MSE values of the other models, particularly Random Forest and XGBoost, are significantly lower.
Interpretation: Random Forest, SVR, and XGBoost all give more accurate predictions with fewer significant errors when their MSE is lower. The best generalization to unknown data is shown by XGBoost.

Mean Absolute Error (MAE) Comparison:-

Note: MAE exhibits the similar pattern as MSE, with XGBoost and Random Forest producing the lowest average prediction errors and Linear Regression performing the worst.
Explanation:
Tree-based and boosting models' lower mean absolute error (MAE) demonstrates that they consistently forecast stock prices that are closer to their actual values while still avoiding significant errors.



Model Performance Comparison

This comparison chart shows how different models performed on both training and test data across key metrics (R² Score, MSE, MAE).

V. RESULT

Three distinct machine learning models—Linear Regression (LR), Support Vector Regression (SVR), and Random Forest—were tested in this stock price prediction experiment using historical stock price data from a variety of firms, including DMart, Amazon, and Tesla. Key performance indicators like the Coefficient of Determination (R2), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE) were used to evaluate the models. According to our research, LSTM networks performed better than the other models in most cases because they could identify long-term trends and temporal relationships in sequential data, which made them especially useful for stock price prediction. When market conditions were stable, SVR demonstrated strength in low volatility circumstances and produced accurate predictions. Although it provided a quick and easily comprehensible baseline, linear regression was unable to fully capture the intricate patterns and non-linear interactions present in stock market data. Overall, the findings show how crucial it is to choose the right model depending on the particulars of the dataset and the state of the market, with LSTM turning out to be the most reliable choice for precise stock price predictions.

CONCLUSION

Four machine learning models—Linear Regression, Support Vector Regression, Random Forest, and Long Short-Term Memory—for stock market price prediction were thoroughly and comparably analyzed in this study. The study has shown that no one model can be regarded as universally ideal by combining findings from five academic papers and evaluating these models on a real-world dataset. Rather, under particular circumstances, each algorithm demonstrates its strengths. Because LSTM networks can process sequential data, they are perfect for long-term forecasting and have shown the greatest efficacy in predicting temporal relationships in stock values. While SVR proved dependable when market volatility was minimal, Random Forest demonstrated excellent performance when managing non-linearities and provided useful feature interpretability. Despite its simplicity, linear regression was nevertheless a valuable baseline model that was quick to understand.The study also emphasizes how important feature engineering, model validation, and comprehensive data pretreatment are to guaranteeing precise predictions. It emphasizes that stock market forecasting is inevitably context-dependent and that the type of data and the intended use case should be taken into consideration when choosing a predictive model. Future work can explore hybrid and ensemble approaches, real-time implementation, integration of alternative data sources like news sentiment, and deep learning advancements such as attention mechanisms and Transformers.

REFERENCES

• Abdulhamit Subasi et al., "Stock Market Prediction Using Machine Learning," Procedia Computer Science, 2021.

• Sonali Antad et al., "Stock Price Prediction Website Using Linear Regression," ICDSAC 2023.

• Shital Pashankar et al., "Machine Learning Techniques for Stock Price Prediction - A Comparative Analysis of Linear Regression, Random Forest, And Support Vector Regression," Journal of Advanced Zoology, 2024.

• Bruno Henrique et al., "Stock Price Prediction using Support Vector Regression on Daily and Up to the Minute Prices," Journal of Finance and Data Science, 2018.

• Ishita Parmar et al., "Stock Market Prediction Using Machine Learning," ICSCCC IEEE Conference, 2018.

• Zhang, G., Eddy Patuwo, B., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. international journal of forecasting, 14(1), 35-62.

• Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654- 669.

• Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications, 42(4), 2162-2172.Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long short-term memory. PloS one, 12(7), e0180944.

• Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Systems with Applications, 83, 187-205.

• Brownlee, J. (2020). Deep Learning for Time Series Forecasting. Machine Learning Mastery.

• Nelson, D. M., Pereira, A. C. M., & de Oliveira, R. A. (2017). Stock market’s price movement prediction with LSTM neural networks. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE.