Proceedings of the Second International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS-2024)
IEEE Xplore Part Number: CFP24DN7-ART; ISBN: 979-8-3503-6841-3

2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) | 979-8-3503-6841-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSSAS64001.2024.10760490

A Comparative Study of Machine Learning Models for Stock Price Prediction

Rashmi Ashtagi
Department of Computer Engineering
Vishwakarma Institute of Technology,
Pune, India.
rashmi.astagi@vit.edu

Amogh Deshmukh
Department of Computer Engineering
Vishwakarma Institute of Technology,
Pune, India.
amogh.deshmukh22@vit.edu

Siddhesh Dhomse
Department of Computer Engineering
Vishwakarma Institute of Technology,
Pune, India.
siddhesh.dhomse221@vit.edu

Gayatri Dhumal
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India.
gayatri.dhumal221@vit.edu

Siddhartha Dhurve
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India
siddhartha.dhurve22@vit.edu

***Abstract* - The stock market has always been evolutionary. It has been an area of interest for researchers to forecasting the stock prices too guide investment decisions and avoiding financial risk since many years. This paper highlights the significance of precise stock price prediction and investigates the use of several machine learning techniques to strengthen the predictive accuracy. The proposed system makes uses two different approaches: Regression, which predicts the company's stock closing price, and Classification predicts whether the closing price of the stocks will increase or decrease the next day. Historical stock data is examined carefully and preprocessed to ensure high quality of data. Using feature engineering appropriate features are extracted to identify the intrinsic patterns and trends within the stock market. Algorithms like Linear Regression, ARIMA, LSTM and Random Forest are implemented and thoroughly evaluated for their efficiency in stock price prediction. After conducting a comparative analysis of the two algorithms, it becomes evident that while LSTM performs exceptionally in regression tasks, identifying the temporal associations and patterns that guide the stock market data in its development, Random Forest provides more firm predictions of the stock price movement. This study provides the extensive analysis of the real-life application of machine learning for discussions on the stock market, thus allowing investors to derive a more reliable resource to refer to when dealing with the challenges of the stock market**.

Keywords— Stock price, LSTM, ARIMA, Random Forest, Linear Regression, Financial markets, Predictive accuracy, Comparative analysis.

I. INTRODUCTION

Predicting the prices of stocks has always been challenging yet significant task in financial markets. Forecasting the stock prices accurately, give investors essential awareness enables them to make proper choices and manage risks efficiently. Due to the abundant availability of historical stock data and recent advancement in machine learning methodologies, the fascination with predicting stock price movements has been fueled. Both the BSE Sensex and NSE Nifty 50 prove this point as good representations of how well our Indian stock market has grown over time. For example: BSE Sensex, which was somewhere around 25000 points in the year of 2014 but it crossed the levels of Rs.5000 marks approximately in INR terms depicting a CAGR (return) at an average rate for almost ten percent (10%) yearly compounding growth from that period till now. Similarly, NSE Nifty 50 has shown strong performance aligning with growth of Sensex, brought by the major sectors such as pharmaceuticals, IT and banking. These changes yield investment opportunities and challenges, which demand a more accurate model to forecast the stock prices. In this study, we will implement different methods to predict share prices by employing machine learning algorithms. This study attempts to explore an efficient framework using the power of machine learning that could help investors in discovering open-door investment opportunities and also boost their risk management. Hence the prediction of stock prices become a crucial factor in financial markets. It can help in providing better investment strategies, decrease financial risks and has a higher decision-making capability. This project aims to predict future stock prices and analyses which of the several algorithms — Random Forest, ARIMA (Auto Regressive Integrated Moving Average), LSTM (Long Short-Term Memory), Linear Regression are most efficient in doing so. This study has been conducted to explore the most efficient method of predicting price ever. The analysis demonstrates the benefits of deploying machine learning algorithms for prediction of stock prices and highlights the importance of developing this practice. The main purpose is to give investors a reliable source, where they will be more informed and feel at ease in facing the different complexities of financial market. The result of this research not only increased the field in financial prediction but also gave valuable advice to stakeholders (e.g. users, investors) at Indian counter and institutional level.

II. LITERATURE REVIEW

[1] The need to develop accurate stock price prediction models has long been the focus of financial markets research. The ability to precisely predict fluctuations in stock prices is particularly significant because it is a crucial component of predictive models. Theoretical perspectives which imply that an accurate description of stock price movement cannot be accomplished, describe empirical observations by noting that well designed models can make robust predictions against historical data and average analysis even without perfect knowledge. Deep learning methods, particularly the Long Short-Term Memory (LSTM) Algorithm have proven to be fruitful in this domain. In this study, the LSTM Algorithm is considered as indicator to predict stock price movements with ten years historical data of NIFTY 50 Index across India. The dataset that was selected for the analysis ran from 2011 to December 10 2021, and normalization was carried out before the model was trained and validated. From the preliminary results of the model, the findings were promising with accuracy of 83.88 percent. This study reaffirms the suitability and precision of DL-based LSTM Algorithm to forecast the stock price movement, thus its integration and adoption in financial forecasting.

[2] The aim of this study is to minimise the risk in prediction of stock markets by utilizing machine learning and deep learning algorithms. It concentrates on numerous sectors such as, non-metallic minerals, basic metals, and petroleum within the Tehran Stock Exchange. The research compares nine machine learning algorithms SVC, Naïve Bayes, KNN, Random Forest, Logistic Regression, Artificial Neural Network, Adaptive Boosting, eXtreme Gradient Boosting combined deep learning techniques, specifically Long Short-Term Memory (LSTM)and Recurrent Neural Network (RNN). This study utilizes two approaches as binary transformation of indicators and continuous calculations using stock data from past decade. Evaluation highlights RNN and LSTM as exceptional performers for continuous data analysis, while deep learning methods maintain efficacy in binary data analysis. These conclusions contribute to enhancing forecasting techniques in financial markets, hence advancing risk mitigation strategies for stakeholders.

[3] This analysis inspects the complexity of stock market price prediction, a progressively essential aspect of trading activities. Understanding the intricacy of this task, the research leverages machine learning methodologies, particularly focusing on ARIMA model and LSTM networks, known for their precision in dynamic environments. The study evaluates the performance of ARIMA for short-term forecasting using data from Tata Global Beverages, demonstrating its substantial potential and ability to with other widely recognized techniques. Also, LSTM networks are appreciated for their potential to predict future stock movements. The careful testing demonstrates that these perform much well with a remarkable accuracy rate of 77%. These perception highlights the crucial importance of machine learning methods in forecasting stock market trends, with LSTM networks standing out as especially effective tools for precise forecasting.

[4] This study offers a detailed approach for predicting stock price momentum by employing machine learning methodologies utilizing raw data from Yahoo Finance, the system extracts feature such as stock movement, index volatility, and sector momentum. The dataset is divided in training and test sets, with the model trained on the training data. The following matrices are used to evaluate the prediction performance of models: computation time, confusion matrix, and accuracy. Also, the study provides a comprehensive analysis of technical terms related to finance and stock markets, such as stock market index, outstanding shares, market capitalization, and the S&P 500 index. The significance of the S&P 500 index as a benchmark for representing market volatility and economic performance is emphasized in the discussion of dataset selection and preprocessing. The study focuses on the crucial role of machine learning in increasing the accuracy of stock price predictions and provides knowledge of practical uses for investors and fund managers.

[5] Stock market prediction poses challenges due to nonstationary and noisy stock prices influenced by various factors. Simple Exponential Smoothing and ARIMA are some of commonly used techniques for forecasting. However, few studies have focused on the (CSE) Colombo Stock Exchange for predictive analysis. This article explores the effectiveness of deep learning algorithms, such as Back Propagation Neural Network (BPNN), in contrast to traditional methods. The results indicate that BPNN performs better than ARIMA in terms of Mean Absolute Error (MAE) and Mean Squared Error (MSE), suggesting its superiority for forecasting CSE stock prices.

[6]The researchers wrestle with the challenge of predicting future price movements amid the debate over the efficient market hypothesis (EMH). While some believe accurate predictions are unattainable, key literature suggests otherwise, showcasing the potential for precise forecasts with suitable methodologies. This paper introduces a robust framework for stock price prediction by integrating statistical, machine learning, and deep learning models. Using daily data from the National Stock Exchange of India, collected at five-minute intervals and aggregated into three daily slots, the study constructs eight classification models and eight regression models, also two deep learning-based regression models utilizing LSTM and CNN architectures. This comprehensive model-building approach aims to effectively capture the volatile and random patterns in stock prices. Nevertheless, the study mentions the potential limitations regarding the generalizability of the results and the scalability of the proposed framework, indicating the need for further research.

[7] In the field of stock prediction, the accuracy of forecasts holds paramount importance, particularly with the increasing interest in stock investment. Employing SARIMA as the statistical method, the study enhances prediction outcomes by

utilizing five deep Learning models: Single-Layer LSTM, Three-Layer LSTM, Bidirectional LSTM, CNN LSTM, and Convolutional LSTM. Each model is designed to utilize specific features, such as LSTM layers for enhanced output, Bidirectional LSTM for incorporating future inputs, and CNN LSTM for capturing abstract features. The performance of each Deep Learning model is assesed with help of metrics such as RMS, RMAE, MAE, and R^2-Score. However, there may be limitations in the generalizability of the results and their applicability to other company stocks, highlighting the need for further research and validation.

[8] This study addresses the formidable challenge of accurately predicting stock market returns by leveraging artificial intelligence and advanced computational techniques. Employing ANN and Random Forest methodologies, we forecast the closing price for next day for five companies spanning various sectors. Using financial data attributes such as Open, High, Low, and Close prices, new variables are generated as inputs for the models. Evaluation using the standard metrics like RMSE and MAPE indicates the efficiency of our models in predicting stock closing prices. However, a limitation to consider is the reliance on a sparse historical dataset, which may impact prediction accuracy.

[9] For predictions of stock price movements, various methodologies have been explored, including traditional Linear Regression and Support Vector Regression (SVR), which proved inadequate due to non-linear and complex nature of stock price movements. Researchers turned to ARIMA for improvement, but the high variations in stock prices necessitated the application of deep learning techniques. While ANN showed promise, RNN was employed to better handle time-series data. However, RNN's limitations led to the adoption of LSTM networks. This paper compares LSTM and SVR using data from various stock indices including NASDAQ, NSE, NYSE, BSE, and Dow Jones illustrating LSTM's higher accuracy compared to SVR. SVR's reliance on linearity and sensitivity to hyperparameters impedes accurate stock price prediction.

[10] The prediction of stock market data captivates global investors, with speculation regarding the relationship between past and future stock returns. Understanding stock market trends is vital for the economy, attracting both trading communities and computer enthusiasts. This system makes the predictions of stock prices for the next day and certain future dates by employing the Moving Average technique to improve accuracy. This study seeks to evaluate the intrinsic value of Moving Average indicators for investors and analysts. Additionally, sentiment analysis of real-time data allows users to gauge public opinion about specific companies, aiding in trend identification. The prediction accuracy is evaluated and presented to users, enabling them to assess stock trends with known accuracy and check future prices of individual companies.

III. METHODOLOGY

A. Flowchart:

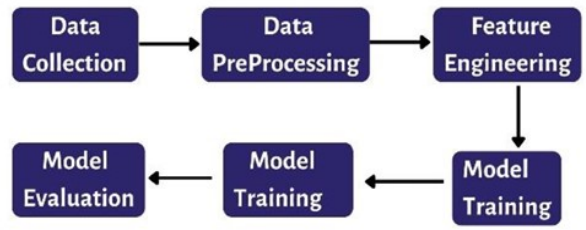

Fig 1. Flowchart

B. Implementation:

1. Dataset:

The selected dataset for this research consists of stock data for **Nestle India**, a leading multinational food and beverage company. This dataset provides a comprehensive view of Nestle India's stock market dynamics, including key variables like stock prices (open, close, high, low), trading volumes, and other financial indicators. The data was collected from publicly available sources such as stock exchanges or financial data platforms (e.g., Yahoo Finance, NSE, or BSE). These sources ensure the reliability and accuracy of the stock data, which covers a significant time frame for predictive analysis and exploration of market trends.

2. Data Preprocessing:

The dataset, containing key stock market metrics such as daily high and low prices, opening and closing prices, and trading volumes, requires thorough preprocessing before it can be used for machine learning. Initially, handling missing values is essential; this can be done by either removing incomplete records or using techniques like interpolation to estimate missing data. Extracting relevant features, including stock prices and trading volumes, is the next step, ensuring that the model can learn from historical patterns. Normalization of these features is crucial, especially since stock prices and volumes vary widely. Scaling values to a common range, typically between 0 and 1, prevents discrepancies that could skew the model's predictions. This normalization standardizes the data, allowing the algorithm to treat all features equally. Additionally, feature engineering involves deriving new financial indicators, such as moving averages, to enhance model performance. Data transformation, including converting the dataset into sequences for models like LSTM, further prepares it for accurate predictions. This comprehensive preprocessing ensures the dataset is clean, consistent, and ready for effective machine learning analysis.

3. Algorithms:

In this study, we used four distinct predictive models to predict stock closing prices: LSTM, ARIMA, Linear Regression and Random Forest. Each model offers unique strengths in capturing different aspects of stock price movements.

Long Short-Term Memory (LSTM): It is an RNN (recurrent neural network) that is able to of learn long-term dependencies in the sequential data. RNN seems especially promising for time-series prediction problems such as turning stock price predictions into a hit or miss game because it relies on memory cells that store information over extended time periods. From a stock market perspective, these are the sets of stock prices (prices can be open, close high and low), trading volumes or other financial performance metrics recorded into such series through time at regular intervals. LSTMs enable models to learn different temporal structures or patterns in the past stock price data and use it for prediction of future prices. LSTMs, on the other hand uses a more sophisticated architecture to traditional RNNs which contains Cell States which retains information across time steps as Forget Gate, Input Gate and Output Gates. The forget gate determines which information from the previous time step to retain or discard, while the input gate decides what new information should be added to the memory. The output gate then controls the final output, filtering the updated memory cell's information. Together, these gates enable LSTMs to selectively store, update, and recall information over long sequences, making them highly effective at capturing long-term dependencies in data. This ability to handle both short-term fluctuations and long-term trends makes LSTMs well-suited for stock price prediction. In this research, the LSTM model used historical stock data to predict future closing prices, successfully capturing the nonlinear relationships and temporal dependencies in stock prices.

Auto Regressive Integrated Moving Average (ARIMA): The ARIMA method, is a statistical technique for time series forecasting. It is the combination of three components: Autoregression (AR), differencing (I), and moving average (MA). One of the best-in-practice methods in Time series forecasting is working on stock price prediction. The ARIMA model most suited as it can handle linear trends and seasonal patterns which are not present or minimalistic under typical market conditions. To build a successful ARIMA model it is crucially need to look at Autocorrelation Function (ACF) and Partial Autoregression functions. ACF shows correlation between a variable and its previous values, helping to identify the MA order, whereas the PACF identifies the AR order by reflecting the direct effects of past values. The Auto ARIMA function streamlines the process and improves prediction accuracy along with saving time by choosing the most suitable parameters automatically on the basis of factors like AIC or BIC. In this study, the parameters (6,0,7) were chosen for the ARIMA model.

Linear Regression: This is a statistical technique employed to model the correlation between a dependent variable and one or multiple independent variables fitting in a linear equation to the observed data. Linear regression is often employed for forecasting tasks when the relation between the input variables and the outcome is approximately linear, the model calculates the coefficients (weights) for each feature and an intercept term to establish a linear function for predicting the simple and efficient approach. Despite its simplicity, linear regression has drawbacks, such as the assumption of linearity, which may not apply to more complicated datasets which have non-linear relationships. Also, this method can be susceptible to outliers and issues of multicollinearity among the features.

Random Forest: Random Forest is an ensemble learning technique which constructs numerous decision trees during training and combines their outputs to generate predictions. It collects the mode of the classes that each individual tree predicts, for classification tasks and for regression, it calculates the average prediction from all trees. Random Forest is a robust and versatile model for modelling complex nonlinear relationships in data. In this implementation, an ensemble of 800 decision trees is employed, controlled by the n_estimators' parameter, which helps the model to capture various data patterns. The min_samples_split parameter is set to 200 to ensure that a minimum of 200 samples are required for splitting an internal node, which helps to avoid overfitting and supports generalization. The randomized nature of model is managed by setting random_state to 1, which ensures consistent outcomes across runs. These parameter settings are designed to balance model complexity with predictive accuracy and computational efficiency, making the Random Forest Classifier a robust tool for predicting stock prices.

Following training and validation of these four models, the closing point values predicted by each algorithm are obtained. Overall, this methodology integrates diverse predictive models, leveraging their unique capabilities to forecast stock prices accurately.

IV. RESULTS AND DISCUSSION

To ensure a robust evaluation process for our stock price prediction models, we implemented several key steps. We defined and utilized metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to gauge model accuracy. The dataset was divided into training and test sets while preserving chronological order to maintain the integrity of time series data. We employed Time Series Cross-Validation to assess performance across various periods. Model comparisons were made using these metrics, and we analyzed residuals to identify and address systematic errors. Results were visualized through prediction versus actual plots and error histograms to gain deeper insights into model performance. We fine-tuned hyperparameters and refined models based on these evaluations. Additionally, we validated the models on real-time or out-of-sample data to confirm their robustness and generalization capabilities. This comprehensive evaluation approach ensured the accuracy and reliability of our stock price predictions.

After studying and analyzing the four models on the stock price prediction dataset, it was found that LSTM and ARIMA models have performed better. LSTM and ARIMA precisely captured the complex patterns and dynamics inherent in stock price data, resulting good proficiency in handling temporal dependencies and nonlinear relationships. The RMSE values obtained for LSTM was 0.27 and for ARIMA was 0.35, highlighting their strong performance.

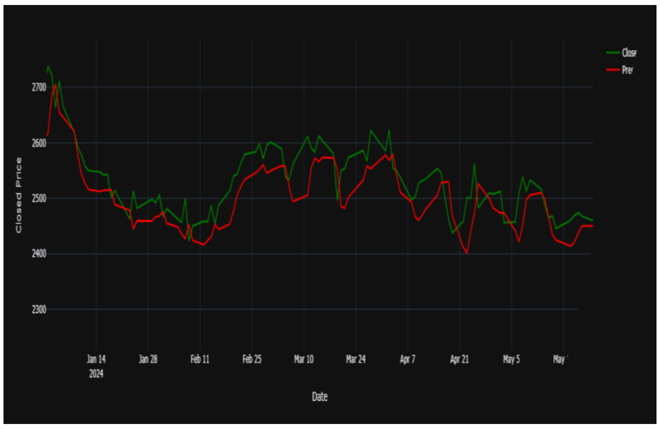

Fig 2. Results obtained from LSTM

Figure 2 shows a comparison of the closing prices predicted by LSTM, which shows the highest performance among the models evaluated.

Linear Regression displayed signs of overfitting, indicating its limitations in determining the underlying trends in stock price prediction. In spite of its simplicity and understandability, Linear Regression struggled to predict unseen data.

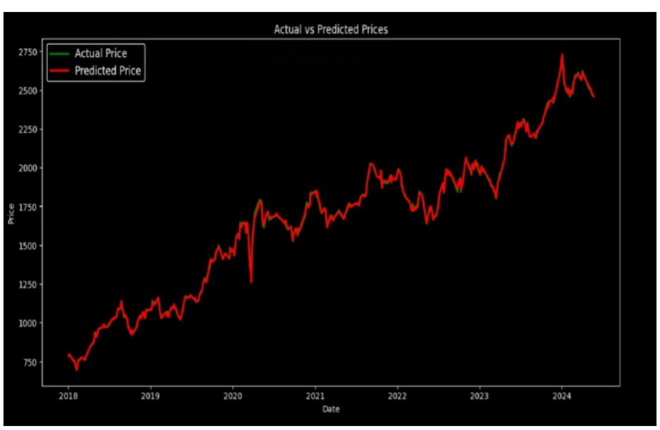

Fig 3. Results obtained from Linear Regression

Figure 3 shows the occurrence of overfitting in Linear Regression by comparing the actual closing prices with the predicted closing prices.

Random Forest Classifier, while not as effective as LSTM and ARIMA, it has shown the results with an accuracy of 65% on the test data set. It accurately predicted 73% of the time when the stock price decreases and 56% of the time when the price increases. Figure 4, illustrates these outcomes.

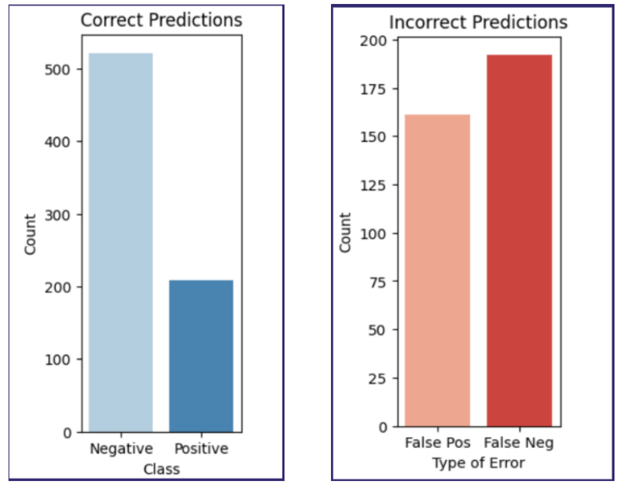

Fig 4. Results obtained from Random Forest Classification

In summary, LSTM and ARIMA emerge as preferred choices for stock price prediction tasks, while Random Forest Classifier offers a viable alternative for classification tasks. To enhance the accuracy, techniques like hyperparameter tuning (e.g., adjusting the number of layers, learning rate), grid search, early stopping, and dropout can be employed to prevent overfitting. Additionally, incorporating feature engineering, such as using technical indicators or adding relevant financial variables, could further enhance model performance.

VI. CONCLUSION

In conclusion, the results of our research highlight the importance of choosing appropriate machine learning algorithms for predicting stock prices. Our comparative analysis revealed significant performance differences between the ARIMA and LSTM models, with ARIMA showing superior accuracy and lower RMSE values compared to LSTM. Although LSTM is better at learning the long-term temporal dependency in stock feature data, ARIMA is more suitable for detecting linear or near-linear connection of the time sequence among stocks prices. Our results demonstrate the importance of using accuracy in addition to RMSE and suggest a multi-metric evaluation approach is essential for complete understanding on model performance. In addition, our findings contribute to demonstrate the advantages of utilizing ensemble methods including LSTM plus ARIMA as well as combining these tools with techniques such like KNN in order taking benefit from all models and consequently increasing prediction robustness and accuracy. Looking forward, future research ventures could focus on further enhancing model performance through additional model improvements, feature engineering techniques, and the inclusion of external variables. To improve model performance, advanced approaches like hybrid models (e.g., combining LSTM with ARIMA for trend-seasonal decomposition) can be explored. Ensemble learning or meta-learning techniques can also be used to blend predictions from different models, leading to more robust and accurate results. By enhancing the predictive ability of these algorithms, especially in real-world stock market scenarios, researchers and practitioners can more effectively utilize machine learning techniques for better decision-making and risk management in financial markets.

REFERENCES

[1] Sisodia, Pushpendra Singh, et al. " [1]." *2022 2nd international conference on innovative practices in technology and management (ICIPTM)*. Vol. 2. IEEE, 2022.

[2] Nabipour, Mojtaba, et al. "Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis." *Ieee Access* 8 (2020): 150199-150212.

[3 Ashok, Asha, and C. P. Prathibhamol. "Improved analysis of stock market prediction:(ARIMA-LSTM-SMP)." 2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE). IEEE, 2021.

[4] S. Ravikumar and P. Saraf, "Prediction of Stock Prices using Machine Learning (Regression, Classification) Algorithms," *2020 International Conference for Emerging Technology (INCET)*, Belgaum, India, 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9154061.

[5] G. W. R. I. Wijesinghe and R. M. K. T. Rathnayaka, "Stock Market Price Forecasting using ARIMA vs ANN; A Case study from CSE," 2020 2nd International Conference on Advancements in Computing (ICAC), Malabe, Sri Lanka, 2020, pp. 269-274, doi: 10.1109/ICAC51239.2020.9357288.

[6] Mehtab, Sidra, and Jaydip Sen. "A time series analysis-based stock price prediction using machine learning and deep learning models." *International Journal of Business Forecasting and Marketing Intelligence* 6.4 (2020): 272-335.

[7] Srilakshmi and C. Sai Sruthi, "Prediction of TCS Stock Prices Using Deep Learning Models," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 1448-1455, doi: 10.1109/ICACCS51430.2021.9441850.

[8] Vijh, Mehar, et al. "Stock closing price prediction using machine learning techniques." Procedia computer science 167 (2020): 599-606.

[9] Bathla, Gourav. "Stock Price prediction using LSTM and SVR." 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE, 2020.

[10] Shaikh, Aryan, and Sandip Shinde. "Prediction of Stock Market Prices using Prediction algorithm and Sentiment Analysis." *Available at SSRN 4111874* (2022).

[11] Singh, Bhupinder, et al. "Auto-regressive integrated moving average threshold influence techniques for stock data analysis." International Journal of Advanced Computer Science and Applications 14.6 (2023).

[12] A. Mahadik, D. Vaghela and A. Mhaisgawali, "Stock Price Prediction using LSTM and ARIMA," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2021, pp. 1594-1601, doi: 10.1109/ICESC51422.2021.9532655.

[13] Yadav, Anita, C. K. Jha, and Aditi Sharan. "Optimizing LSTM for time series prediction in Indian stock market." Procedia Computer Science 167 (2020): 2091-2100.

[14] Bathla, Gourav. "Stock Price prediction using LSTM and SVR." 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE, 2020.

[15] Mehtab, Sidra, and Jaydip Sen. "A time series analysis-based stock price prediction using machine learning and deep learning models." *International Journal of Business Forecasting and Marketing Intelligence* 6.4 (2020): 272-335.