DLINE JOURNALS

Journal of Digital Information Management

JDIM

ISSN 0972-7272

Published online: 27 July 2024

Predictive Modeling of Stock Price Trends Using Machine Learning and Deep Learning Techniques

K. Kiruthika, E.S. Samundeeswari
Department of Computer Science
Vellalar College for Women
Erode
Tamil Nadu
India
{kkiruthika3108@gmail.com}
{essamundeeswari@gmail.com}

ABSTRACT: *Predicting stock price movements has been challenging yet crucial for investors and financial analysts. Fluctuations in stock prices are valuable economic indicators, providing insights into overall economic well-being, consumer confidence, and market sentiment. In this study, we evaluate the efficacy of three different machine and deep learning algorithms in anticipating stock price trends. We assess the performance of Logistic Regression, Random Forest Regression, and Long Short-Term Memory (LSTM) algorithms in forecasting whether a stock's price will rise or fall in the upcoming period, utilising historical stock price data as input features. Our findings demonstrate that while each algorithm exhibits varying degrees of predictive accuracy, LSTM networks stand out as they generally outperform Logistic Regression and Random Forest Regression in capturing the complex temporal dependencies inherent in stock price data. This suggests that LSTM networks, with their superior performance, hold significant promise as effective tools for stock price trend prediction, particularly in volatile and non-linear financial markets. This could be a game-changer in stock price prediction, instilling optimism about the future of stock market analysis.*

Subject Categories and Descriptors: **[J. Computer Applications]:** Financial; **[H.5 INFORMATION INTERFACES AND PRESENTATION]**: Artificial, augmented, and virtual realities; **[I.2.11 Distributed Artificial Intelligence]:** Intelligent agents

General Terms: Stock Price Prediction, Logistic Regression, Random Forest Regression, Long Short-Term Memory

Received: 14 March 2024, Revised 3 May 2024, Accepted 28 May 2024, Published online 27 July 2024

Keywords: Stock Price Prediction, Logistic Regression, Random Forest Regression, Long Short-Term Memory

Review Metrics: Review Scale: 0/6, Review Score: 5.12, Inter-reviewer consistency: 87.2%

DOI: https://10.6025/jdim/2024/22/3/83-90

1. Introduction

Stock Price prediction is pivotal for stakeholders across the financial ecosystem, including individual investors,

institutional traders, policymakers, and researchers. The role of researchers in this ecosystem is not just crucial but integral to the success of our research. Stock Price prediction supports informed decision-making, enhances market efficiency, and fosters economic growth by aligning capital allocation with market expectations. Stock Price forecasts play a crucial role across various domains for several reasons: Investors rely on predictions to optimise their investment decisions, whether buying, holding, or selling stocks, aiming to maximise returns and manage risks effectively. Our research, and by extension, the researchers, provide accurate predictions and can directly impact investment strategies and outcomes. Moreover, accurate predictions contribute to market efficiency by ensuring stock prices reflect all available information, thereby promoting fair valuation grounded in fundamental and technical factors. The significance of market efficiency underscores the importance of our work in stock price prediction.

Corporations utilise stock price predictions to gauge market sentiment and investor expectations, guiding strategic decisions related to financing, acquisitions, and capital investments. Additionally, stock price movements serve as economic indicators, offering insights into broader economic health, consumer confidence, and market sentiment trends.

In algorithmic trading, predictive models are essential for executing trades efficiently, allowing automated systems to capitalise on fleeting market opportunities. Academic research in stock price prediction drives advancements in machine learning, statistical modelling, and financial theory, continually shaping investment strategies and enhancing market understanding.

In financial markets, machine learning and deep learning algorithms are extensively employed for prediction tasks because they can learn patterns from historical data and subsequently make predictions based on those patterns. Our previous study analysed the correlation between news headlines mentioning DJIA and stock price trends using five machine learning algorithms: Linear SVC, Logistic Regression, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Random Forest. Our findings showed that Logistic Regression and Random Forest Classifier achieved higher accuracy in predicting stock price trends.

This study assesses the effectiveness of machine learning algorithms (Logistic Regression and Random Forest Regression) that demonstrated higher accuracy in our previous research alongside a deep learning algorithm (LSTM) to identify the most accurate approach.

2. Early Related Work

Stock Price prediction is pivotal for stakeholders across the financial ecosystem, including individual investors, institutional traders, policymakers, and researchers. The researchers' role in this ecosystem is crucial as well as integral to the success of our research. Stock price prediction supports informed decision-making and enhances market stock market price prediction, which is a significant area of research among business, machine learning, and database researchers. Many studies have recently applied deep learning and machine algorithms to predict stock prices. Understanding these studies is beneficial before we undertake another work.

For the past twenty years, many studies have frequently employed the random forest (RF) algorithm for regression and classification (1). Since its initial applications, studies have been organised to compare the performance of Random Forests with three versions of logistic regression, particularly for predicting imbalanced datasets (2).

Random Forest regression has enormous scope in application areas. This is reinforced in application work for predicting the bearing capacity of strips (3). A loan default prediction model is framed using the Random Forest algorithm, where the RF algorithm performs better than other machine learning algorithms. (4) In another work, neural Networks, Support Vector Machines, and Decision Trees are used to forecast Corporate Credit Ratings, where the decision tree-based models performed well. (5). A Long Short-Term Memory (LSTM) neural network-based model is more effective than other neural models during the testing in Oil production forecasting [6].

Many studies favoured the LSTM models in many applications; however, Convolutional neural network models and LSTM equally yielded convincing results (7). The logistic regression models obtained by stepwise selection outperform the other models in financial distress prediction. (8).

In a study to predict the most portable sectors in stocks, DAORI et al. (9) used many machine learning algorithms and recorded higher accuracy in the random forest algorithm. Integrating random forest framework and long short-term memory yielded better results. The multi-tasking model has been proven to predict stock returns and movements together. (10).

Kummari Vikas (11) et al. compared various machine learning algorithms and documented the better performance of Recurrent Neural Networks (RNN) and Long-Short-Term Memory (LSTM). Chung et al. (12) created a new approach to metric analysis, improving the traditional moving average by adding a deep-learning training model and documenting enhanced performance. Using multiple feature engineering techniques combined with a customised deep learning-based system yields more acceptable results. (13) Deep learning applications in stock market prediction have increased due to their better accuracy in financial prediction. (14, 15). Deep learning models have replaced linear and Machine learning models in the last few years because of their elegance and effective outcomes. (16). Recent studies on stock forecasting using enhanced deep learning features reinforce the promise of early re-

search on it. (17,18,19, 20)

The research in the above directions offers insights into selecting particular or combined deep learning and machine learning models to predict the stock market trend. These studies further proved that more research is required to find the best and optimum model, which involves not just the model selection but also improving the features and refining training models.

3. Methodology

3.1. Dataset

Stock market prediction studies use heterogeneous datasets, and the parameters, mainly using two or more models, are complex. The primary dataset used in the studies is the stock exchange. Some research considers a particular or a few stocks, but the use of variables and models varies. We gain deeper insights into complex stock data using multi-model patterns.

The dataset for this work is the historical stock price for Apple Inc. (AAPL) spanning 2015 to 2020, accessed through Tiingo API. This dataset includes several critical components for each trading day within the specified period:

• **Date:** The date of each trading day.

• **Open:** The initial price of AAPL stock at the beginning of the trading day.

• **High:** The highest price reached by AAPL stock during the trading day.

• **Low:** The lowest price reached by AAPL stock during the trading day.

• **Close:** The price of AAPL stock at the end of the trading day.

The AAPL dataset from 2015 to 2020 is a rich and comprehensive source of information. It is an ideal resource for training and evaluating Machine Learning models such as Logistic Regression, Random Forest Regression, and LSTM networks.

3.2. Feature Selection

Data cleaning is crucial to guarantee that the dataset is consistent, accurate, and prepared for analysis. Consequently, missing values are addressed using statistical techniques such as mean, median, and mode, while any duplicate rows are removed to facilitate further study.

Feature selection entails identifying and selecting the most pertinent features that substantially contribute to the prediction or analysis. The features that impact stock prices (open, close, high, low, adjClose, volume) are considered independent variables, with 'close' as the dependent variable.

3.3. Train-Test Split

The dataset is partitioned into training (80%) and testing (20%) sets to assess the model's performance on unseen data. Since the stock price data is time-series, the dataset is split sequentially to preserve temporal order.

3.4. Normalisation

Normalisation ensures all features are brought to a comparable scale, preventing any single feature from dominating the model simply because of its larger numerical values.

• **Min-Max Scaling:** Scale numerical features (e.g., open, high, low, volume, adjClose) to a range typically [0, 1] using:

$$X_{\text{scaled}} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (1)$$

Where X is the original feature value.

3.5. Description of Methods

Logistic Regression

Logistic regression is a commonly employed statistical model designed for binary classification tasks. When predicting stock prices, the objective usually involves forecasting whether a stock's price will rise or fall using historical data and pertinent features

Logistic Regression can be adapted for stock price direction prediction. In this binary classification task, the model predicts whether the stock price will go up or down based on historical data and relevant features.

Logistic regression models probability $P(y = 1 | x)$. X represents the input features (*open, high, low, volume, adjClose*), and y is the binary outcome (price increase or decrease). The logistic regression algorithm employs the logistic function, also known as the sigmoid function, to model and predict probabilities.

```
[16] df.head(2)
```

	Unnamed: 0	symbol	date	close	high	low	open	volume	adjClose	adjHigh	adjLow	adjOpen	adjVolume	divCash	splitFactor
0	0	AAPL	2015-05-27 00:00:00+00:00	132.045	132.26	130.05	130.34	45833246	121.682558	121.880685	119.844118	120.111360	45833246	0.0	1.0
1	1	AAPL	2015-05-28 00:00:00+00:00	131.780	131.95	131.10	131.86	30733309	121.438354	121.595013	120.811718	121.512076	30733309	0.0	1.0

Table 1. AAPL dataset

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \quad (2)$$

```
[20] df.head(1)
```

date	close	high	low	open	adjClose	volume
2015-05-27 00:00:00+00:00	132.045	132.26	130.05	130.34	121.682558	45833246

Table 2. Independent Variables

```
[24] print(features)
['high', 'low', 'open', 'adjClose', 'volume']

[25] print(target)
close
```

Table 3. Dependent Variable

```
print(X_train)

[[0.89842419 0.83804184 0.90935698 0.84244118 0.45309608]
 [0.03913969 0.0373086  0.03882242 0.0333952  0.14519385]
 [0.72155026 0.7193444  0.71408281 0.72702    0.05880427]
 ...
 [0.74510221 0.750022   0.75540157 0.74399119 0.04925366]
 [0.55451448 0.53137805 0.53129035 0.54634225 0.18009389]
 [0.73138842 0.73137805 0.72626323 0.73643922 0.09407538]]
```

Table 4

Random forest Regression

Random Forest Regression is an ensemble learning technique that builds multiple decision trees during training and derives predictions by averaging the outputs of these individual trees. In contrast to classification problems where each tree in the forest votes for a class, regression problems like stock price prediction predict a continuous value, and the final prediction is the average (or sometimes median) of these values.

Random Forest Regression forecasts future stock prices using historical data and a specified set of input features. It combines predictions from numerous decision trees, each trained on random subsets of features and data samples. By averaging the predictions of these trees, it provides a robust estimate of the stock price. The predicted stock price $\hat{y}$ is the average prediction across all trees.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \quad (3)$$

where $f_i(x)$ is the prediction of the i^{th} tree.

Random Forests determine feature importance by evaluating how significantly each feature reduces impurity across all trees within the forest. This helps in understanding which features are most influential in predicting stock prices.

LSTM (Long Short-Term Memory)

LSTM, a variant of recurrent neural networks (RNNs), excels in capturing temporal dependencies and long-term patterns within sequential data, which is ideal for tasks like forecasting time series data such as stock prices. LSTM cells maintain an internal state (cell state) to remember important information over long sequences. The model predicts the next stock price $\hat{y}_{t+1}$ based on the previous prices and other relevant features.

$$\hat{y}_{t+1} = f_{LSTM}(X_{1:t}) \quad (4)$$

where $X_{1:t}$: represents the sequence of historical prices up to time *t*.

LSTMs have memory cells that maintain a cell state to retain information over long sequences.

They selectively update and forget information through a series of gates, allowing them to learn and remember patterns over time.

LSTMs have three gates

Forget Gate: Determines which information to remove from the cell state.

Input Gate: Integrates new information into the cell state.

Output Gate: Generates the output based on the updated cell state.

LSTM networks are trained using backpropagation through time (BPTT) to optimise the network parameters (weights and biases). The loss function (for example, mean squared error) measures minimised to improve the model's ability to predict future stock prices accurately.

4. Results and Discussion

The dependent variable (Open, High, Low, adjClose, Volume) and independent variable (Close) in the AAPL company's historical stock price are extracted from the Yahoo finance website through Tiingo and split into training and testing datasets. The split dataset is trained and tested to compare the performance of Random Forest, Logistic Regression, and LSTM algorithms. Accuracy in predicting the stock price trend with these algorithms is evaluated using a confusion matrix and evaluation metrics. A confusion matrix allows visualisation of the performance of the algorithms in predicting stock price trends. For stock price predictions, if we consider a binary classification problem where:

- **Positive class (P):** Indicates a prediction that the price will increase (or a buy signal).

- **Negative class (N):** Indicates a prediction that the price will decrease (or a sell signal).

Algorithm	TP	TN	FP	FN
Logistic Regression	7	131	4	110
Random Forest Regression	61	75	60	56
Long Short-Term Memory	0	141	0	110

Further, leveraging the confusion matrix enables the calculation of critical quantitative evaluation metrics such as Accuracy, Precision, Recall, and F1-Score, which are determined to predict stock price trends.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

$$\Pr ecision = \frac{TP}{TP + FN} \quad (6)$$

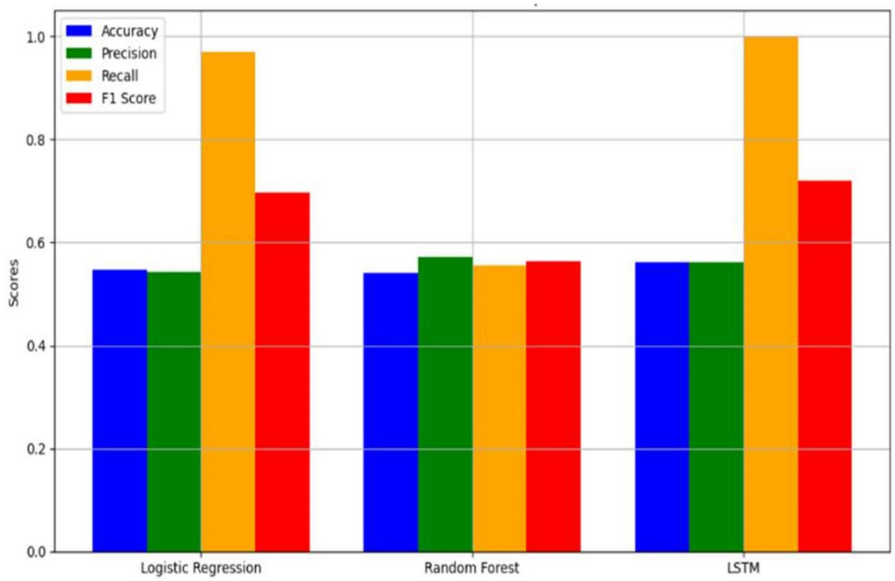

Figure 1. Comparison of Evaluation Metrics

$$\text{Re}call = \frac{TP}{TP + FN} \quad (7)$$

$$F1 = 2 \times \frac{\text{Pr}ecision \times \text{Re}call}{\text{Pr}ecision + \text{Re}call} \quad (8)$$

Following the ROC Curve of Logistic Regression, Random Forest Regression, and LSTM algorithms help assess how well the model can distinguish between price movements and determine the optimal threshold for making buy/sell decisions.

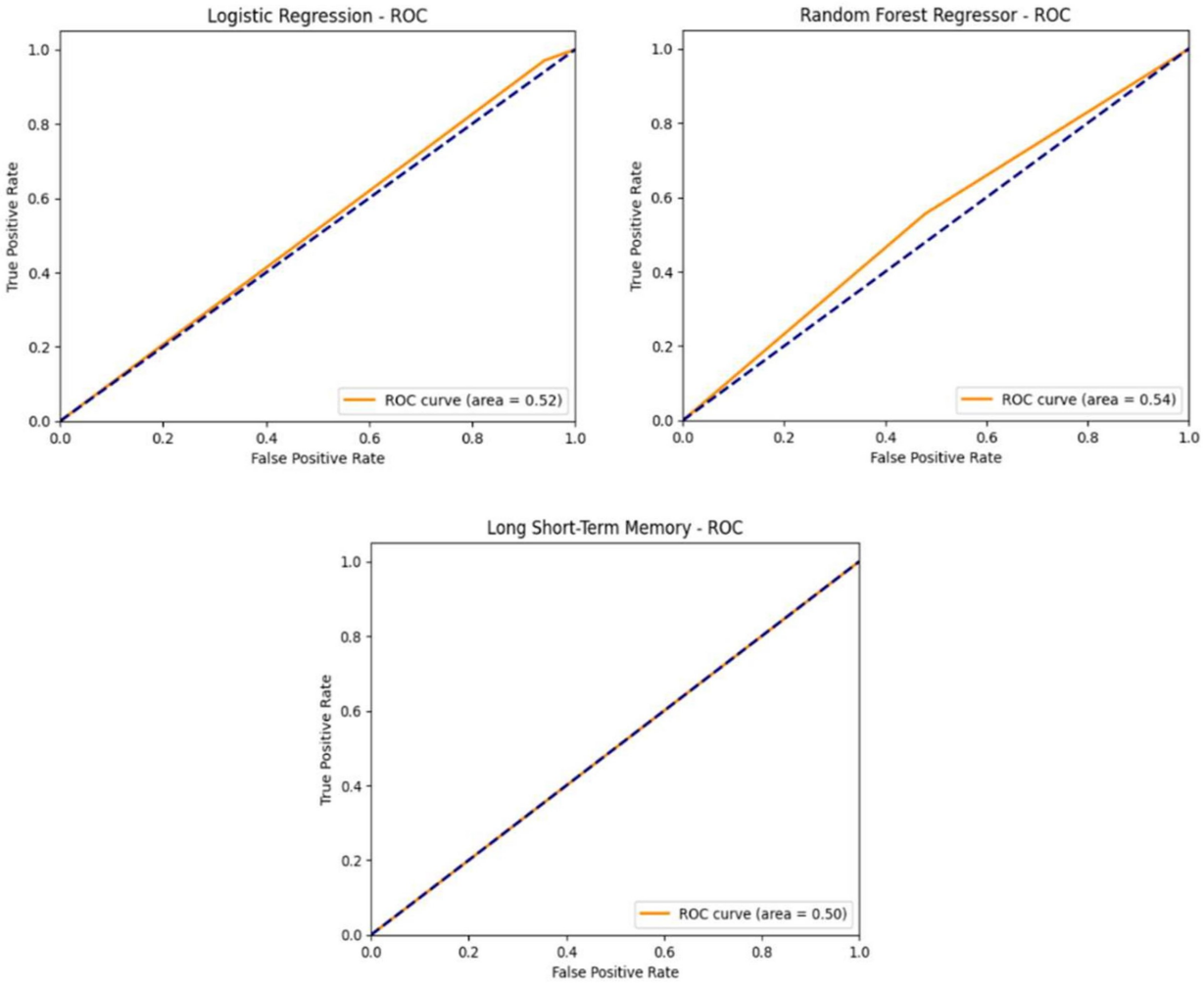

Figure 2. ROC Curve for Logistic Regression, Random Forest Regression and LSTM

Logistic Regression shows a relatively balanced distribution between correctly predicted positive and negative instances, with seven true positives and 131 true negatives. However, it also predicts four false positives and 110 false negatives. This indicates that while it correctly identifies many non-events (true negatives), it struggles with identifying actual events (true positives) and has many false negatives, which are events incorrectly classified as non-events.

The Random Forest Regression performs better than Logistic Regression regarding true positives (61) and negatives (75), indicating it effectively predicts positive and negative instances. However, it also has more false positives (60) and false negatives (56) than Logistic Regression. This suggests that while Random Forests capture more true positives, they also make more errors in classification, leading to a trade-off between sensitivity and specificity.

LSTM shows many true negatives (141), indicating it correctly identifies non-events. However, it fails to predict any true positives (TP = 0), resulting in many false negatives (FN = 110). This suggests that while LSTM is highly

specific in identifying non-events, it struggles significantly with sensitivity to detect actual events, resulting in poor performance for positive predictions.

5. Performance Comparison

Current deep and machine learning technology surpasses humans in predicting stock trends and prices. Still, it falls short of understanding and perceiving socio-political changes that impact the financial markets at any time. The success of prediction models based on deep or machine learning depends on how elegantly the proposed models capture insights and translate them to forecasts. To assess this kind of model, we can diagnose the inclusion of features in the models. Infusing multimodal and human synthesis features will take it to more volume of success. Secondly, the accuracy and success rate determine the efficiency of the models.

Our work used logistic regression, random forest regression, and LSTM to predict stock and assess performance using variables such as evaluation metrics and ROC curves. The outcome documents the accuracy and success rates leading to acceptance of the proposed model.

6. Conclusion

In our previous work [21], top news headlines with DJIA stock price shifts were taken, and five machine learning algorithms were used to analyse which algorithms predicted the stock price fluctuation more accurately. Among the five, Logistic Regression and Random Forest Classifiers predict better than others, with little difference in accuracy. As an extension, the best-performing top two algorithms in that work and the deep learning algorithm (LSTM) are implemented in this paper for Continuous data.

In evaluating these algorithms for predicting stock price trends, Logistic Regression demonstrates reasonable specificity (TN) but lacks sensitivity (TP), resulting in many false negatives. Random Forest Regression offers a better balance between sensitivity and specificity than Logistic Regression, yet it still shows notable errors in classification (FP and FN). On the other hand, LSTM excels in specificity (TN) but fails to predict positive outcomes (TP = 0), revealing substantial shortcomings in sensitivity.

Logistic Regression might be preferred for its simplicity and interpretability, while Random Forests could be chosen for their balance between sensitivity and specificity. LSTM, despite its shortcomings in this scenario, remains powerful for tasks requiring sequence modelling and long-term dependencies but may require additional tuning and data preprocessing to improve performance in predicting positive events.

Further, this work can be extended by including technical indicators and an improved LSTM model to achieve higher accuracy in predicting stock price trends.

References

[1] Couronné, Raphael., Probst, Philipp., Boulesteix, Anne-Laure. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 19, 270.

[2] Luo, Hanwu., Pan, Xiubao., Wang, Qingshun., Ye, Shasha., Qian, Ying. (2019). Logistic regression and random forest for effective imbalanced classification. *In IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*.

[3] Dutta, R. K., Gnananandarao, T., Sharma, A. (2019). Application of random forest regression in the prediction of ultimate bearing capacity of strip footing resting on dense sand overlying loose sand deposit. *Journal of Soft Computing in Civil Engineering*, 3(4), 28-40.

[4] Zhu, Lin., Qiu, Dafeng., Ergu, Daji., Ying, Cai., Liu, Kuiyi. (2019). A study on predicting loan default based on the random forest algorithm. Elsevier, 503-513.

[5] Golbayani, Parisa., Florescu, Ionut., Chatterjee, Rupak. (2020). *A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees*. Elsevier Ltd.

[6] Song, Xuanyi., Liu, Yuetian., Wang, Jun. (2020). *Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model*. Elsevier Ltd.

[7] Madaeni, Fatemehalsadat., Chokmani, Karem., Lhissou, Rachid., Homayouni, Saeid., Gauthier, Yves., Tolszczuk-Leclerc, Simon. (2021). Convolutional neural network and extended short-term memory models for ice-jam predictions. *Copernicus Publications*, 1447-1468.

[8] Zizi, Youssef., Jamali-Alaoui, Amine., El Goumi, Badreddine., Oudgou, Mohamed., El Moudden, Abdeslam. (2021). An optimal model of financial distress prediction: A comparative study between neural networks and logistic regression.

[9] Daori, Hind., Alharthi, Manar., Alanazi, Alanoud., Alzahrani, Ghaida. (2022). Predicting stock prices using the random forest classifier. https://doi.org/10.21203/rs.3.rs-2266733/v1

[10] Park, Hyun Jun., Kim, Youngjun., Kim, Ha Young. (2022). Stock market forecasting using a multi-task approach integrating long short-term memory and the random forest framework. *Applied Soft Computing*, 114.

[11] Kummari, Vikas., Kotgire, Kamal., Bayyapu, Sai., Reddy, Prudhvi., Teja, Akula Sai Krishna. (2022). An integrated approach towards stock price prediction using LSTM algorithm. *In: International Conference on Edge Computing and Applications* (ICECAA).

[12] Chung, Yao-Liang., Wu, Zheng-Lin., Pichappan, Pit. (2024). Application of deep learning and statistical methods in predicting Taiwanese stock trends. *Journal of Computational Methods in Sciences and Engineering*, 24(3), 2017-2035.

[13] Shen, Jingyi., Shafiq, M. Omair. (2020). Short-term stock market price trend prediction using a comprehensive deep learning system. *Research*, 7, 66.

[14] Hu, Zexin., Zhao, Yiqi., Khushi, Matloob. (2021). A survey of Forex and stock price prediction using deep learning. *Appl. Syst. Innov*., 4(1), 9.

[15] Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A., Salwana, E. S. (2020). *Deep learning for stock market prediction. Entropy*, 22, 840.

[16] Jiang, Weiwei. (2021). Applications of deep learning in stock market prediction: Recent progress. *Expert Systems with Applications*, 184, 115537.

[17] Muhammad, Tashreef., Aftab, Anika Bintee., Ibrahim, Muhammad., Ahsan, Md. Mainul., Muhu, Maishameem Meherin., Khan, Shahidul Islam., Alam, Mohammad Shafiul. (2023). Transformer-based deep learning model for stock price prediction: *A case study on Bangladesh* stock market. *International Journal of Computational Intelligence and Applications,* 22(3), 2350013.

[18] Mukherjee, Somenath., Sadhukhan, Bikash., Sarkar, Nairita., Roy, Debajyoti., De, Soumil. (2021). Stock market prediction using deep learning algorithms. *CAAI Transactions on Intelligence Technology*, 82-94.

[19] Sonkavde, G., Dharrao, D. S., Bongale, A. M., Deokate, S. T., Doreswamy, D., Bhat, S. K. (2023). 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*, 11, 94.

[20] Habib, Honey., Kashyap, Gautam Siddharth., Tabassum, Nazia., Tabrez, Nafis. (2023). Stock price prediction using artificial intelligence based on LSTM-deep learning model. *In Artificial Intelligence & Blockchain in Cyber Physical Systems*. CRC Press.

[21] Kiruthika, K., Samundeeswari, E. S. (2023). Sentiment analysis to predict stock price fluctuations using multiple machine learning algorithms. *In International Conference on Recent Trends in Computer Science and Data Analytics (ICRTCSDA'23)*.