Measurement: Sensors 31 (2024) 100950

Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

Financial stock market forecast using evaluated linear regression based machine learning technique

J. Margaret Sangeetha [a,*], K. Joy Alfia [b]

[a] Department of Computer Science, St. Xavier's College, Palayamkottai, Tamil Nadu, India
[b] Department of Computer Applications, St. Xavier's College, Palayamkottai, Tamil Nadu, India

ARTICLE INFO

Keywords:
Stock market forecasting (SMF)
Prognosis
Evaluated linear regression (ELR) technique
based machine learning (ML)
Standard and poor's 500 (s & p 500) index
Open
Close
Low
High and volume

ABSTRACT

The objective of Stock Market Forecasting (SMF) is to forecast the future value of a company's financial stocks. The availability of Machine Learning (ML), particularly obtains forecasts regarding specific prices of current stock market indexes through training on actual high economic value, constitutes a significant improvement in stock market prediction automation. ML itself utilizes a wide range of models to improve as well as evaluate prediction. Regarding the nonlinearities as well as discontinuities of the factors that are anticipated to effect stock markets, the selection of a modest price of global financial data is frequently acknowledged as a significant basic phase in any stock market prediction model. The fundamental emphasis of the study is the application of Evaluated Linear Regression based Machine Learning (ELR-ML) technique to forecast stock financial values of the Standard and Poor's 500 (s & p 500) index with Open, close, low, high, and volume factors.

1. Introduction

Predicting stock prices has proven substantially more problematic in recent years because they are currently impacted by a variation of factors including the socioeconomic situation of such nation, the political climate, natural disasters as well as other factors in addition to a company's financial performance. The financial markets are one of today's most fascinating innovations. These financial markets have a big impact on a lot of things, like business and employment technologies. In order to invest their money and increase returns while minimizing risk, investors have mostly adopted two tactics. Among expert financial analysts, such advancement of stock market forecasting has taken on important implication. According to the chaotic environment in the market, it is exceedingly complicated to analyze price actions including stock market movements.

[1] provides a summary of automation along with artificial intelligence as stock market predictive analytics capabilities [2]. was accomplished using LSTM, a Recurrent Neural Network (RNN) variation, as opposed to the latter, which makes use of Convolutional Neural Network. The Generative Adversarial Network based Hybrid Prediction Algorithm (GAN-HPA) is approved for more usage [3]. Suggested using LSTM to effectively predict stock values by combining sentiment and historical data. According to a review of the sentiment analysis study, there is a significant correlation between the movement of stock prices and the publication of news reports [4]. Asserts a deep learning approach incorporating genetic algorithms to predict the overnight return direction of a given stock market index, using worldwide stock market indices as an instructional source. In order to forecast stock market prices, multiple regression as well as support vector machine methods will be used [5].

The following is the list of paper's main contribution.

(1) The input data is collected from the data source of yahoo finance dataset based on their price index.
(2) After collecting the input the data is preprocessed for noise removal of parameters.
(3) Then extraction and selection of features from the huge amount of input data's.
(4) The two subcategories of current and forecast data's are analyzed and estimated.
(5) This estimated results make a strong decision and notification is sent to the investors about the price index through an ELR-ML technique.

The following is the representation of this paper's configuration: Section 1 offers a comprehensive introduction to forecasting the stock

* Corresponding author.
E-mail address: margaret.msu@gmail.com (J.M. Sangeetha).

https://doi.org/10.1016/j.measen.2023.100950
Received 8 April 2023; Received in revised form 4 October 2023; Accepted 12 November 2023
Available online 15 December 2023

market. The literature on the stock market forecasting approach is summarized in Part 2. The procedure for identifying research subjects is stated in section 3. An assessment of some experimental findings including discussions is covered under Paragraph 4, while Section 5 clearly includes the conclusion.

2. Literature survey

Stock analysts Christy Jeba Malar, others et al. [6] are continually focusing towards this study to forecast future share prices, which helps investors decide whether to buy or sell stocks for a profit. Koukaras et al. [7] developed a model for predicting stock movement using information from Stock Twits as well as SA on Twitter. This strategy's effectiveness was examined through sentiment but also stock movement data, and then validated using Microsoft stock. For a fuzzy nonlinear forecast, Hussain et al. [8] is proposed model can take into account the weighted average, IOWA operator as well as relevance degree of each concept together in specific situation. A thorough analysis of all the many learning models utilized to forecast the stock market during the previous 50 years (1970–2020) is provided by Sarangi et al. [9]. According to Sing et al. [10], the Nifty 50 Index is predicted utilizing Eight Supervised Machine Learning Models. AdaBoost, Linear Regression, Random Forest, Stochastic Gradient Descent, Artificial Neural Networks, k-Nearest Neighbors, Support Vector Machines, and Decision Trees are the methods employed in the empirical investigation (DT).

A machine learning model-based stock market forecast utilizing the K Nearest Neighbor (KNN) technique is suggested to overcome these issues. KNN can process the relationship between the numeric data by Latha et al. [11], making it particularly effective in numeric assumption issues for predicting next day change in stock value. The impenetrable shield of shareholders has a positive effect on short-term predictability, especially for small equities, according to Leippold et al. [12]. Soni et al. [13] analyze the many approaches used in the prediction of share prices, ranging from traditional machine learning and deep learning techniques to neural networks and graph-based approaches. According to Luo et al. [14], SFLA can mimic the frog lifestyle when it comes to creating a shared search in a global arena. To improve training and SFLA effectiveness, a correction technique based on mutation and crossover is developed. Albahli et al. [15] focused on the prediction of closing stock prices using ten years of Yahoo Finance data of ten notable stocks or STIs. They used 1D DenseNet followed by an autoencoder.

In order to predict stock prices using Twitter data, Swathi et al. [16] introduce a brand-new TLBO model, which stands for Teaching and Learning Based Optimization with LSTM-based sentiment analysis. A time-efficient hybrid stock trends prediction framework (HSTPF) is suggested by Bhanja et al. [17] in particular for accurately forecast stock market trends there in future, even in the presence of Black Swan events. Here, Black Swan events analysis and feature selection operations are carried out in order to increase the prediction accuracy of HSTPF, as well as the effectiveness of several machine learning classifiers is also examined. According to Kaczmarek et al. [18], portfolio optimization methods such as Markowitz mean-variance as well as Hierarchical Risk Parity (HRP) optimizers boost the risk-adjusted return of portfolios that are constructed using stocks which have been preselected using a machine learning tool. Houssein et al. [19] develop a hybridised approach that combines the equilibrium optimizer (EO) with support vector regression (SVR) methods for predict the closing prices of the Egyptian Exchange (EGX). Patel et al. [20] created a method to predict movements by combining statistical indicators and trading strategies like Auto Regression Integrated Moving Average (ARIMA), Prophet, Momentum Trading, as well as Pairwise Trading with machine learning forecasting models like Attention Integrated Long Short Term Memory (LSTM) Model as well as a Reinforcement Learning agent.

3. Proposed methodology

Stock market seems like a complex problem due to the numerous factors that need to be considered as well as the prediction initially doesn't seem quantitative. However, by employing the relevant machine learning techniques, it's indeed feasible can link outdated data to relevant data, teach the server can make inferences from it, even design itself to generate accurate predictions. The data set under evaluation were obtained from Yahoo Finance. The collection, consisting comprised all crucial stock prices and perhaps other significant statistics, contained almost 2 lakh pieces. For every day of the year, stock prices were shown in the data at specified intervals. Some of the subparts included date, symbol, open, close, low, high, as well as volume. It uses a single potential piece of data.

According to Yahoo Finance's price index, initial input data is first acquired. A feature of the stock market named the price index enables investors to perceive performance by comparing the current price with previous market prices. Data is first preprocessed to remove noise but instead other parameters after data collection. Preprocessed data can thus be important in predicting the stock market. From a huge amount of data, the feature selection algorithms choose out a few features. The dataset is split into current as well as prediction details subcategories by several data analyzer functions or user-friendly approval. Developing enhanced stock market selections is done simpler as a result of these details. Investors being notified about the price index after a final decision has been taken. Because it informs investors of the price index's current profit or loss, this message is quite beneficial. If the clearance results in a profit, the investor can utilize the shares for high sales, and when the price index is low, expansion will receive more attention, resulting in improved decision-making.

3.1. Datasets

Indeed, a crucial component of machine learning is the dataset used. The dataset should be as precise as possible because even minor alterations to the data might create large differences in the final outcome. Supervised machine learning is employed in this suggested ELR on a dataset obtained from Yahoo Finance. This dataset has five variables, including open, close, low, high, as well as volume. The open, close, low, as well as high bid prices are different bid prices for the stock at specific times with essentially straight names. The volume is determined by the number of shares were transferred from one owner to another within the time frame. Utilizing test data, the model is then evaluated.

3.2. Evaluated linear regression (ELR) technique

A methodology named linear regression comprises searching for the straight line that represents the input data points the best. A regression line is the line that fits the data the best. Data values that depart from the regression line after it has been modelled for a batch of data are referred to as outliers. Incorrect data are represented as outliers, which may also point to a poorly fitted regression line. The accuracy of the prognosis can be significantly reduced with such incorrect data values. Outliers can significantly alter the results of a ELR. A straight line representing a linear relationship between the output variable and also the input variable or variables seems to be exactly linear regression always seeks for. Though this may not be usually the case, it is always assumed in linear regression that the input and output variables have a linear relation. For nonlinear data, linear regression does not offer the best fit. It provides appropriate to take into effect polynomial regression for stock movement forecasting.(see Fig. 1)

Using a collection of independent values, ELR-ML is mostly used to predict continuous values. Using a specified linear function, regression forecasts continuous data:

$$V = c + dI + E \tag{1}$$

Where, I signifies for known independent values, c or even d are coefficients, while V is a continuous parameter. Fig. 3 represents the flow chart for ELR-ML model.

As shown in Fig. 2, by minimizing the error function, the gradient descent linear regression algorithm generates accurate predictions of values.

The formula for calculating the line of best fit is,

$$c = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \tag{2}$$

$$d = \frac{\sum Y - c\sum X}{n} \tag{3}$$

Where, the chosen set of data points is n.

The point on the y-axis where the graph crosses or intercepts the y-axis is known as the y-intercept, and it is where the slope, which measures how steep the line is, is located. The equation of a straight line is expressed slightly differently in linear regression analysis using the model,

$$\overline{Y^{\wedge}} = c + dX \tag{4}$$

The slope of the line is denoted by d, while the y-intercept is denoted by c in this formulation. The notation $\overline{Y}$, which denotes the expected value of the dependent variable Y for a specific value of the independent variable X, is known as the "y-hat".

Two (x, y) points were required to formulate the equation for a line while constructing it in algebra.

$$Y = mX + d \tag{5}$$

Consequently, in regression analysis, the linear regression model will be generated considering some of the (x, y) points in the data set.

The evaluated linear regression analysis results in a scatter plot that resembles one from the early stages. After that, the correlation coefficient r will be determined, and its significance will be examined. Given a value for the independent variable, X, a regression line can be used to predict the value of the dependent variable, Y.

This data point's residual could be estimated as follows:

$$\text{Residual} = (\text{measured Y value}) - (\text{forecasted Y value})$$

$$\text{Residual} = Y - \overline{Y^{\wedge}} \tag{6}$$

By identifying the best-fit regression line, the analysis attempts to predict y values in a method that minimizes the error difference between the predicted value as well as the true value.

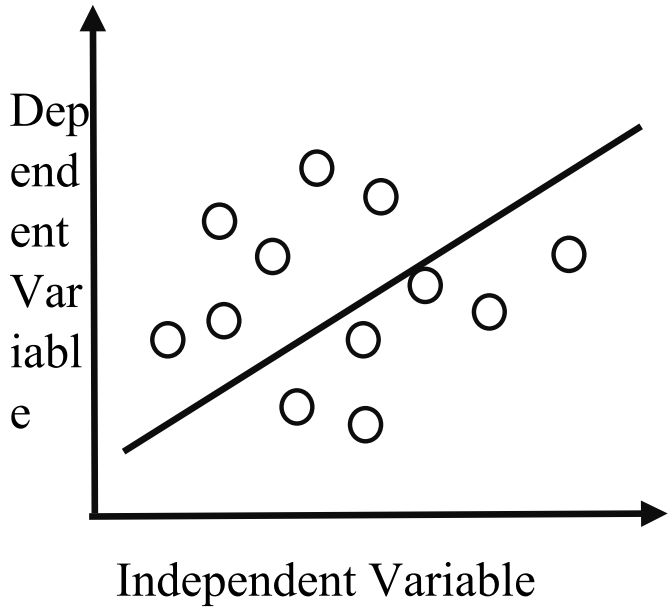

Fig. 2. Linear Regression model for stock market.

$$J = \frac{1}{n}\sum_{m=1}^{n}(\text{estimated}_m - Y_m)^2 \tag{7}$$

The Root Mean Squared Error (RMSE) between the predicted y value (approximated) and also the true y value seems to be the cost function (J) of linear regression (y).

It is imperative to determine the relationships between input as well as output as a set of weights, and then a supervised training set is available. The subsequent steps comprise locating these relationships. In order to calculate costs, the sum square error (SSE) is used.

$$SSE = \sum_{m=1}^{n}(f_m - o_m)^2 \tag{8}$$

Where, f_m seems to be approximately expensive as the desired result, or even o_m produced by the weighted sigmoidal function.

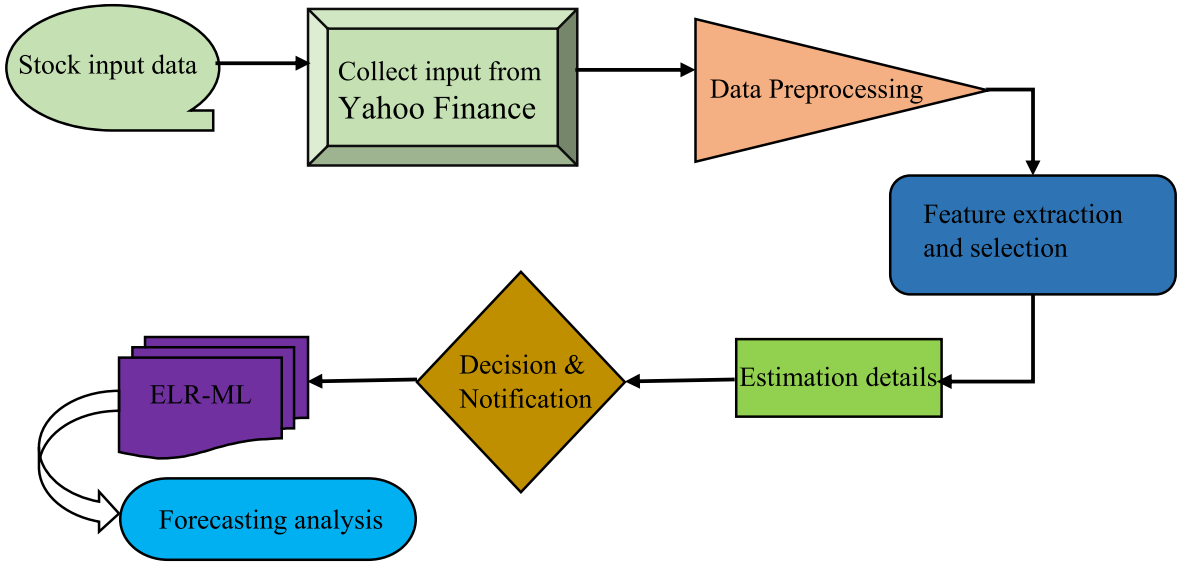

Fig. 1. Block diagram for stock market forecasting.

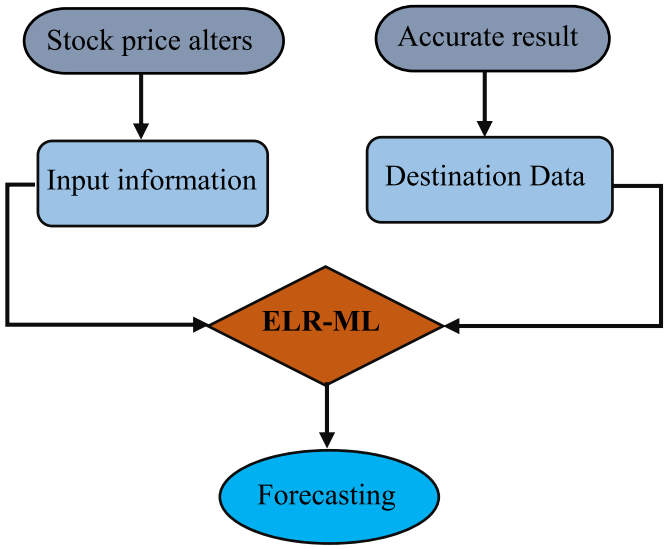

Fig. 3. Flow diagram for Evaluated Linear Regression based Machine Learning model.

All features that are provided seem to be accessible. Below is a description of the attributes.

1. Opening price:

The starting price is the price of each share at the opening of trading on the S&P 500 stock index. The opening price provides a reliable forecast of the stock's daily movement. The beginning price need not match the closing price from the previous day because the stock market is comparable to an auction when buyers and sellers meet to negotiate with the highest bidder.

2. Highest/lowest price of the day:

The previous of a day's high as well as low prices are recorded, providing information of how much the shares generally change over the course of a trading day and how this will ultimately effect the closing price. An adjusted closing price is the closing price of a property that has been modified to account for distributions as well as corporate actions that have occurred prior to the beginning of the following trading day.

3. When analyzing historical returns in detail or examining at recent returns, the adjusted closing price is frequently employed.
4. Volume: One of the most fundamental as well as essential concepts to acquire when trading stocks is volume. The quantity of shares or contracts exchanged in a security or the entire market during a specific time period is known to as the volume.

4. Experimental results and discussion

The suggested approach was trained and tested using data from Yahoo Finance. A numerical value is predicted through regression. The stock market, a particularly risky area of investing, has a large number of buyers, sellers, and investors. A stock typically indicates ownership of a company by a certain person or group of people. With a minimal initial investment as well as low risk compared to starting a new business or finding a rewarding profession, this market has given investors the opportunity to make money as well as spend a pleasant existence. Finding the stock price pattern and making an investment at the right time and place are the only requirements for potentially making money. So, if someone makes an accurate prediction at the right time, they will profit greatly from the limited cash flow that is accessible.

Table 1
Relationship between the independent parameters.

	Low	Close	Open	High
Low	–	–	0.962	–
Lose	–	0.9630	0.993	–
High	0.983	0.986	0.983	–
Volume	−0.421	−0.412	−0.392	−0.396

Table 2
Achieved regression statistics by ELR-ML technique analysis.

Evaluated Linear Regression based Machine Learning statistics (ELR-ML)	
Adjusted R Square	0.352
R Square	0.428
Multiple R	0.612
Observations	68

As can be seen from the relationship chart in Table 1, there is a close association between the four S&P 500 index values, and the dependent variable for volume has a low degree of independence from the other 4 prices.

Data analysis, a feature of Excel that is used for financial analysis, can be done by applying ELR to the data as well as creating predicted patterns. Table 2 displays the summary results of applying the regression analysis. Multiple R has a value of 0.612. The regression line and also least squares value are adequate and well-adjusted to the data when this value is near to 1.

Given that the (Average) parameter is the mean of the prices of the Open, Low, High as well as close, Fig. 4 shows a scatter plot of the data for the remaining 20 %.

Relationship between the average share price (Average) and trading volume (Volume), the dependent as well as independent variables in the regression equation ($y = 8843.5X + 463683$), as well as the R-squared ($R^2 = 0.358$), which was calculated in (Table 2) and displayed with a blue trend line in (Fig. 4). The two variables that were utilized to choose the direction of certain trend line are 35.8 % connected, according to R-squared analysis. Using a scatter plot of the data, this number is utilized for analysis (Fig. 4). The correlation between stock prices according to its Open, Low, High, Close, and Volume status is established as the initial stage in obtaining correlations or associations between desirable independent metrics for a certain situation (Table 1). The interaction between the coefficients have been estimated to three decimal places (80 % of the total values are related).

The standxard error which equals 285,577 but also represents the difference between actual data and indeed the estimated value of volume, was calculated by adding up all residual values, degree of freedom, sum, and mean of squares as given in Table 3.

Table 4 represents the coefficients obtained from applying Linear Regression. Utilizing the coefficients from Fig. 4, the correlation of linear regression produced the following results:

$$\mathrm{ELR} = 4675513 - 106938 * \mathrm{AVERAGE} \quad (8)$$

The proximity of these two lines suggests (Fig. 5) that, after a significant amount of time has passed, a forecast typically approximates the real trend. The model generated a Train Score of 0.00106 MSE (0.03 RMSE) and a Test Score of 0.00875 MSE. The accuracy that is gained will increase when the algorithm is trained more and a larger dataset is used. Compared to the Regression-based Model, the ELR-ML Model was more accurate.

Fig. 6's results show a simulation that produced mixed outcomes, but there are still many crucial lessons to learn from it. The probability of achieving a perfect prediction is almost zero, similar to any financial forecast. With this in consideration, the new volatility estimates' simulation performance can be deemed tolerable. The financial crises of 2000 and 2022 are when the simulation and real S&P 500 index most.

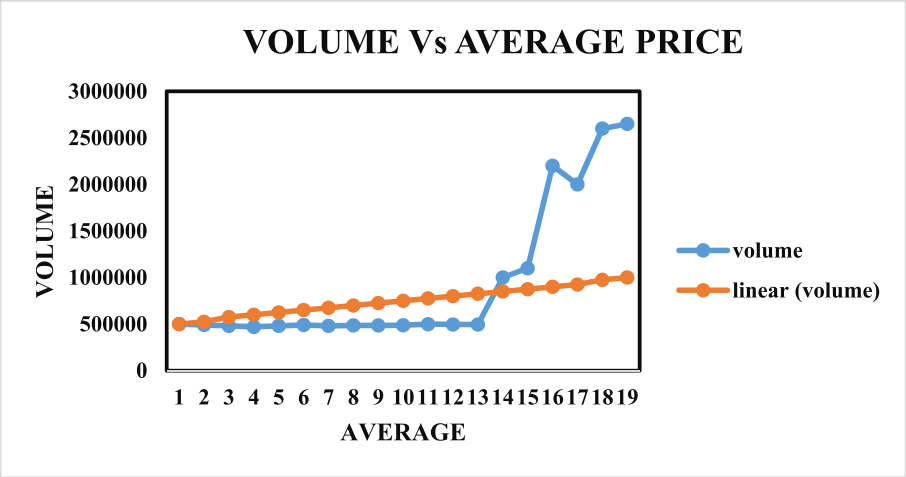

Fig. 4. Visualization of stock prices in scatter.

Table 3
Exploration of deviations of Linear Regression.

S & P 500	df	SSE	MSE
Residual	66	6E+13	9E+12
Regression	4	2E+13	4E+13
Total	70	8E+13	

Table 4
Variables determined through the utilization of ELR.

	Coefficients	Standard error
Intercept	4782610	702,146
Average	−115,402	19,104

Fig. 7 illustrates the effectiveness of linear regression for predicting S&P 500 stock market movement using standard deviation as well as mean square error.

Fig. 8 shows the plot of the predictive performance MSE analysis of financial S & P 500 stock market forecast.

Table 5 represent the reference frequency of some top listed of stock market. It is clear that the only stocks covered repeatedly were those from North America and Asia.

4.1. Prediction metrics

An assessment metric is used to assess the efficiency for a predictive model. A model may first be developed on a dataset, after which it must be used to predict values on a test dataset that was not utilized during training. The predicted values in the test dataset must then be compared to the model's predictions.

(1) R-squared method: A statistical technique that assesses the convergent validity is R-squared. On a scale of 0–100 %, it assesses the strength of the correlation between the dependent and independent variables. A good model is one in which the R-square value is high since it indicates that there is little variation between the predicted and actual values. It can be calculated using the formula below:

$$\text{R squared} = \frac{\text{Stated variation}}{\text{Total variation}} \tag{9}$$

(2) Use the Mean Squared Error (MSE) cost function for linear regression, which measures the average squared error between the predicted as well as actual values. It is spelled as follows:

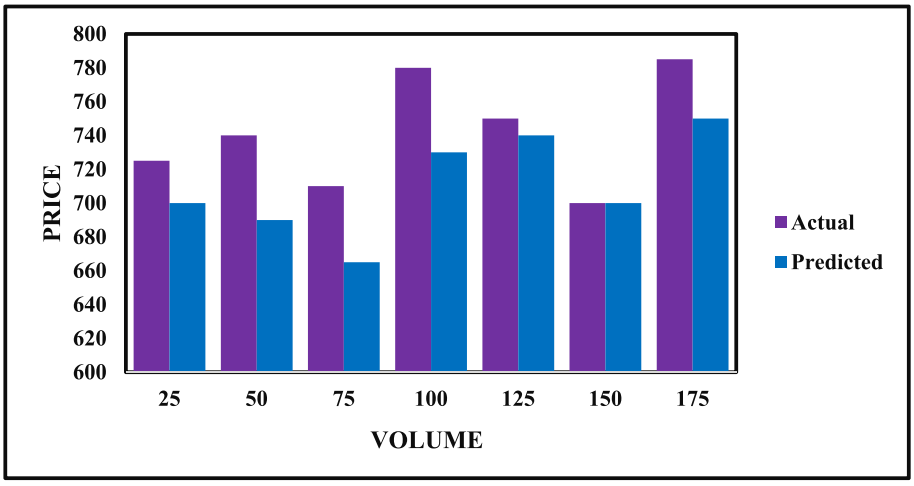

Fig. 5. Visualization comparing the ELR trend's predicted as well as actual characteristics.

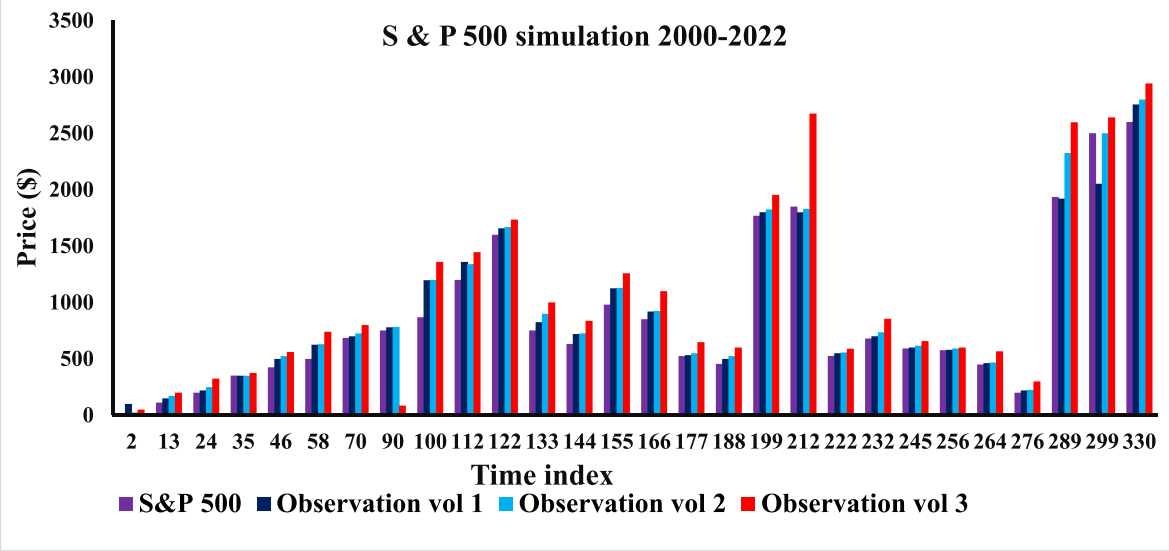

Fig. 6. Visualization of certain simulation.

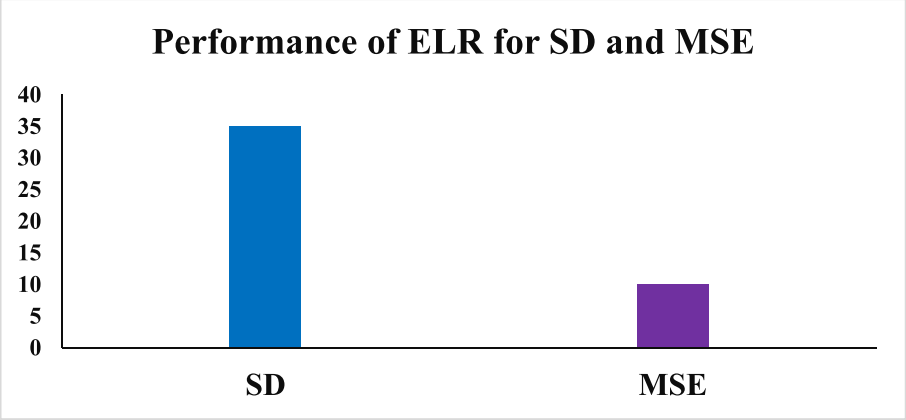

Fig. 7. Performance of linear regression using SD and MSE.

$$MSE = \frac{1}{N}\sum_{m=1}^{n}\left(Y_m(c_1X_m + c_0)\right)^2 \quad (10)$$

Where, N = Total no. of observation, Y_m = Actual Value, $c_1X_m + c_0$ = Observed Value.

(2) **Standard Deviation (SD):** A set of values' variation or dispersion is measured by the standard deviation.

$$SD = \sigma = \sqrt{\frac{(\text{total} - \text{error})^2}{(n-1)}} \quad (11)$$

5. Conclusion

For the development of successful market trading methods, forecasting the movements of the stock market index is crucial. By selecting an efficient forecasting model, traders can decide whether to purchase or sell an object. Accurate Stock Market Index Movement Prediction may be profitable for investors. Predicting changes in the Stock Market Index is an extremely challenging and intricate task. This study used machine learning techniques to forecast stock prices for a corporation with a higher degree of accuracy and reliability. The experts' main contribution was the addition of the ELR-ML Model as a mechanism for calculating stock prices.

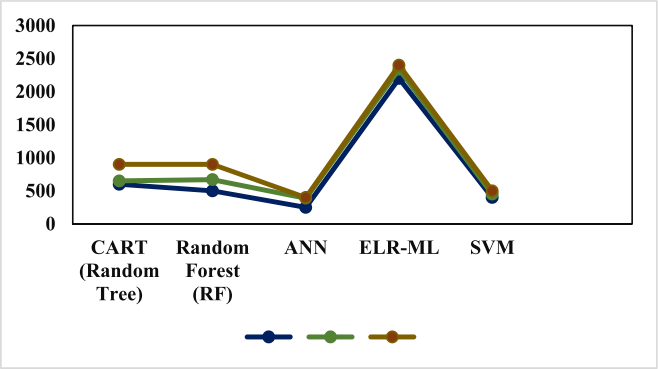

Fig. 8. Plot of the predictive performance of financial S & P 500 stock market forecast.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table 5
Frequency of some top stocks list.

Stocks	Count
Alphabet (Google)	5
Abbott Lab.	5
Amazon	7
Altria	4
Amgen	3
Dell	3
CoCa Cola	2
Facebook	3
IBM	10
LG Corp	3

Data availability

No data was used for the research described in the article.

References

[1] Parshv Chhajer, Manan Shah, Ameya Kshirsagar, The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction, Decision Analytics Journal 2 (2022) 100015.

[2] Subba Rao Polamuri, Kudipudi Srinivas, A. Krishna Mohan, Multi-model generative adversarial network hybrid prediction algorithm (MMGAN-HPA) for stock market prices prediction, Journal of King Saud University-Computer and Information Sciences 34 (9) (2022) 7433–7444.

[3] Ishu Gupta, Tarun Kumar Madan, Sukhman Singh, Ashutosh Kumar Singh, HiSA-SMFM: historical and sentiment analysis based stock market forecasting model, arXiv preprint arXiv:2203.08143 (2022).

[4] Ruize Gao, Xin Zhang, Hongwu Zhang, Quanwu Zhao, Yu Wang, Forecasting the overnight return direction of stock market index combining global market indices: a multiple-branch deep learning approach, Expert Syst. Appl. 194 (2022) 116506.

[5] J. Kalaivani, Ronak Singhania, Shlok Garg, Effect of COVID-19 on stock market prediction using machine learning, in: Biologically Inspired Techniques in Many Criteria Decision Making, Springer, Singapore, 2022, pp. 649–656.

[6] A. Christy Jeba Malar, M. Deva Priya, M. Kavin Kumar, S. Mangala Arunsankar, K. V. Bilal, S. Karthik, Deep learning-based stock market prediction, in: Proceedings of International Conference on Recent Trends in Computing, Springer, Singapore, 2022, pp. 709–716.

[7] Paraskevas Koukaras, Christina Nousi, Christos Tjortjis, Stock market prediction using microblogging sentiment analysis and machine learning, in: Telecom, vol. 3, MDPI, 2022, pp. 358–378, 2.

[8] Walayat Hussain, José M. Merigó, Muhammad Raheel Raza, Predictive intelligence using ANFIS-induced OWAWA for complex stock market prediction, Int. J. Intell. Syst. 37 (8) (2022) 4586–4611.

[9] Sarangi, Pradeepta Kumar, Sunny Singh, Ashok Kumar Sahoo, A study on stock market forecasting and machine learning models: 1970–2020, in: Soft Computing: Theories and Applications, Springer, Singapore, 2022, pp. 515–522.

[10] Gurjeet Singh, Machine learning models in stock market prediction, arXiv preprint arXiv:2202.09359 (2022).

[11] R.S. Latha, G.R. Sreekanth, R.C. Suganthe, M. Geetha, R. Esakki Selvaraj, S. Balaji, K.R. Harini, P. Priya Ponnusamy, Stock movement prediction using KNN machine learning algorithm, in: 2022 International Conference on Computer Communication and Informatics (ICCCI), IEEE, 2022, pp. 1–5.

[12] Markus Leippold, Qian Wang, Wenyu Zhou, Machine learning in the Chinese stock market, J. Financ. Econ. 145 (2) (2022) 64–82.

[13] Payal Soni, Yogya Tewari, Deepa Krishnan, Machine Learning approaches in stock price prediction: a systematic review, in: Journal of Physics: Conference Series, vol. 2161, IOP Publishing, 2022 012065, 1.

[14] Jia Luo, Ge Zhu, Hui Xiang, Artificial Intelligent based day-ahead stock market profit forecasting, Comput. Electr. Eng. 99 (2022) 107837.

[15] Saleh Albahli, Tahira Nazir, Awais Mehmood, Aun Irtaza, Alkhalifah Ali, Waleed Albattah, AEI-DNET: a novel densenet model with an autoencoder for the stock market predictions using stock technical indicators, Electronics 11 (4) (2022) 611.

[16] T. Swathi, N. Kasiviswanath, A. Ananda Rao, An Optimal Deep Learning-Based LSTM for Stock Price Prediction Using Twitter Sentiment Analysis, Applied Intelligence, 2022, pp. 1–14.

[17] Samit Bhanja, Abhishek Das, A Black Swan event-based hybrid model for Indian stock markets' trends prediction, Innovat. Syst. Software Eng. (2022) 1–15.

[18] Tomasz Kaczmarek, Katarzyna Perez, Building portfolios based on machine learning predictions, Economic Research-Ekonomska Istraživanja 35 (1) (2022) 19–37.

[19] Essam H. Houssein, Mahmoud Dirar, Laith Abualigah, Waleed M. Mohamed, An efficient equilibrium optimizer with support vector regression for stock market prediction, Neural Comput. Appl. 34 (4) (2022) 3165–3200.

[20] Harshita Patel, R. Anunay, Bagga, Stock market forecasting using ensemble learning and statistical indicators, Journal of Engineering Research (2022).