2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS) | 979-8-3503-0345-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/AICERA/ICIS59538.2023.10420270

An LSTM-Based Model for Stock Price Prediction

V. Sarika
Dept. of Computer Science
B V Raju Institute of Technology
Narsapur, India
21211a05x5@bvrit.ac.in

G. V. Siva Kamal
Dept. of Computer Science
B V Raju Institute of Technology
Narsapur, India
21211a05x3@bvrit.ac.in

S. V. Pratham
Dept. of Computer Science
B V Raju Institute of Technology
Narsapur, India
21211a05u2@bvrit.ac.in

N. V. S. S. Deepak
Dept. of Computer Science
B V Raju Institute of Technology
Narsapur, India
21211a05y5@bvrit.ac.in

T. Veneela
Dept. of Computer Science
B V Raju Institute of Technology
Narsapur, India
veneela.t@bvrit.ac.in

***Abstract*—The endeavor to predict the future price of an organization's stock is referred to as stock market prediction. It is challenging to predict future trends accurately because the stock market is a dynamic system continually evolving. An impressive profit might be made by correctly predicting the price of a stock in the future. Understanding a company's stock price pattern and forecasting its future development and financial growth will be quite advantageous. The use of machine learning algorithms to forecast stock values has gained popularity in recent years. The objective of this research endeavor is to develop an artificially intelligent model capable of estimating a certain company's stock prices. The application of a sort of machine learning technique known as Long Short Term Memory, which is based on recurrent neural networks (RNNs), is the main topic of this study.**

***Index Terms*—Long short-term memory(LSTM), Recurrent Neural Network(RNN), prediction, stock price.**

I. Introduction

In the world of finance and investing, stock price prediction is a difficult but crucial endeavor. The tendency of stock price fluctuation has long been one of the most serious challenges in the financial sector. For many years, investors and scholars have been interested in stock market prediction due to its volatile, complex, and always-changing character, which makes trustworthy predictions challenging. Making correct predictions in the stock market can be tremendously advantageous for investors, traders, and financial experts. A successful forecast of a stock's future value could result in a substantial profit. Estimating the performance of the stock market is one of the most difficult tasks. It is now an exciting area for research and development since artificial intelligence and machine learning improvements have made forecasts more accurate and trustworthy. Using historical data and a variety of algorithms, a stock price prediction project seeks to reliably anticipate future prices. This entails analyzing massive amounts of data, recognizing trends, and using statistical models to generate predictions. By successfully projecting future stock values, investors can make more educated judgments about when to buy or sell a stock, perhaps resulting in big profits. The characteristic that defines a prediction model depends on variables that can affect market performance. Daily studies on stock price forecasts have been undertaken using a range of data sources with many built-in models, such as articles from newspapers, and data collected from Twitter, Google data, and Wikipedia data. All of these external factors, together with stock prices and stock technology indicators, have impacted the spread of conditional possibilities as a function of foresight disclosed when the correlation between past and prospective recognition is uncertain in many circumstances.

II. Literature Survey

An article [1] presented a single-layer RNN model to estimate the closing value and the subsequent day's peak price. Along with making use of all the time series data, this suggested method also makes use of the following six input attributes: high price, adj-close price, close price, volume, low price, and open price. In this system, RNN, CNN, and LSTM use the information provided to estimate the ending and peak prices for the next day. The suggested single-layer RNN model performs better than other current models when compared to experimental results, according to the analysis. The proposed model has the least MAE and RMSE when compared to other models. The one-layer RNN model proposed is the best for predicting the price of stocks, and it will assist the shareholder in picking the correct stock to buy at the appropriate time to maximize profits.

An article [2] reviewed studies incorporating a standard SMP design. The research was analyzed and contrasted by considering the types of information that were used as feed and significant preliminary processing procedures such as selection of features, order reduction, and visualization of features. The machine learning approaches that were used for prediction are SVM, Deep Neural Networks (DNN), and Regression Algorithms. According to a thorough comparative analysis that was conducted, SVM is the most often utilized method for SMP. Approaches such as ANN and DNN, on the other hand, are commonly utilized because they produce more accurate and faster predictions. Furthermore, including both financial information as well as information acquired from internet resources improves the precision of forecasting.

In [3], a multi-value related network system using an LSTM-based deep recurrent neural network is designed to predict many stock prices at the same time. The related net's feasibility and accuracy are proven by comparing it to the LSTM network model and the LSTM deep recurrent neural network model. Several data sets were used to test the relevance of the Related Network system. Research demonstrates that the Associated Net model's average accuracy is not only higher than that of the other two models. Additionally, it is capable of making many predictions at once, with an average precision of above 95 percent. Although the model produces good results, there are still certain areas that can be strengthened.

[4] proposed a model of forecasting stock markets using a neural network with recurrence, with the data composed of the regular starting prices of two firms on the stock exchange of the NYSE (GOOGL and NKE) gathered from the financial website Yahoo Finance. This model was created using the LSTM and RNN, 80 percent of the gathered data is used for training and 20 percent for evaluation. To optimize this model during training, they employed mean squared error (MSE). Additionally, they used different Epochs for training. Both the number of intervals and the quantity of the input information had a significant impact on the outcomes of validation after training the Neural Network. The results of the tests show that the model can track the growth of starting prices for both resources.

[5] The behavioral response towards web news is taken into account in a method for stock price prediction utilizing machine learning to close the gap and increase the prediction's accuracy for which real-time stock prices were collected. Sentiment scores were computed after preprocessing each tweet. In this study, sentiment analysis is a key component. A recurrent Neural Network (RNN) is employed as stock prediction requires a time series. It could be stated that the influence of social networks and web news data sources is fading for forecasts for the future, while Twitter is useful for short-term prediction.

The stock price of Apple was used as a data set in the neural network methodology proposed in [6] for forecasting stock price. The TensorFlow deep learning framework was used. To validate the influence of the LSTM network on stock time series prediction, this experiment employs two strategies for training the model. Using past stock closing price data, forecasts the next day's stock closing price, 2. Multi-factor input characteristics are based on the starting price, closing price, trading volume, and other parameters in the historical stock data, estimating the closing price of the stock for the following day. As the loss function, the mean square error is used. The mean absolute error is used to analyze the forecasts' outcomes. The model has improved prediction performance on multivariate feature input and fulfills actual demand.

A comparative study [7] was conducted between the standard statistical approach and the machine learning approach. The main intention of this study is to assess the performance of predictive modeling methodologies and traditional statistical methodologies. This method forecasts data by utilizing the statistical mean of past closing prices. The Statistical Prediction using the Simple Moving Average method, weighted moving average method, exponential smoothing method, and naive approach were compared with machine learning techniques such as Simple linear regression, K-Nearest neighbor algorithm, Random Forest algorithm, Vector machine, SLP model, MLP model, and LSTM Predictions. In terms of prediction accuracy and performance, a comparison of statistical and machine learning methodologies has been conducted. After examining each approach separately, Artificial intelligence approaches, specifically MLP and LSTM, are particularly successful in predicting stock prices, with the lowest MSE and mean absolute percentage error values.

A random forest-based model was implemented for stock market prediction in [8]. Noise in the data is eliminated by smoothing to get the information ready to use the Random Forest method. The technique's effectiveness was evaluated by determining how well it predicts stock prices and by calculating the variance score it produces. The four regression values were used in all calculations: variance score, mean absolute error, mean squared error, and mean squared log error. The polarity score was calculated using sentiment analysis and then used to identify the types of articles that had either a favorable or adverse effect on the stock, which could be utilized for future analyses. Finally, the random forest technique was used, and evaluated its effectiveness in comparison to logistic regression. Random forest modeling is substantially more effective than logistic regression for making forecasts for stock markets using analysis of sentiment.

In the methodology implemented in [9], the dataset to be evaluated was collected from Yahoo Finance. It was categorized into numerous sections, including date, symbol, open, close, low, high, and volume. This study focuses on two of the most important models and employs them to produce predictions: The regression-based Model and the LSTM model. The accuracy that can be achieved will increase with the amount of training the system receives and the size of the dataset used. Compared to the Regression-based Model, the LSTM Model provided greater accuracy. The researchers' key contribution was the use of the innovative LSTM Model to determine stock prices.

In an approach proposed in [10], two organizations from the IT sector and the pharmaceutical sector were chosen for the research. Following the training phase, each of these models was tested, and the model with the lowest RMSE was chosen as the ultimate forecasting model. RNN, LSTM, and CNN deep-learning Designs were used. This study exhibits the proposed method's capacity to uncover some interrelationships in the collected information. Furthermore, the results show that CNN architecture is capable of identifying variations in patterns. CNN is chosen as the ideal model for the suggested methodology.

III. Existing Work

In statistical analysis [3], historical market data is examined to find patterns and trends that might predict future price movements. Statistical models frequently make assumptions and simplify the data and correlations between variables. These presumptions could not hold in reality and can result in incorrect forecasts. ANN, an artificial intelligence program [9] based on the structure and operation of the human brain, is frequently used to accomplish the latter. Aside from historical price and volume data, ANN models can also be trained using other types of data, including sentiment from social media and other sources like financial statements, news articles, and financial statements. Although ANN is a flexible machine learning algorithm that may uncover complicated patterns in the data, it can experience overfitting and may not function effectively in a changing market environment. Prior knowledge of forecasting stock prices [1] can be utilized to analyze news stories. social media messages, and other sources of textual data to identify how positive or negative the sentiment is towards a specific stock or company. Sentiment analysis algorithms may be inaccurate and fail to capture the genuine sentiment of the text. Because of the complexities of human language, the setting in which the work was written, and the existence of sarcasm, irony, or other grammatical ambiguities, this may have occurred. For greater precision Deep learning methods frequently use convolutional neural networks[4] used in image and signal processing applications. CNNs can be used to examine historical price data and detect trends to forecast future stock prices. CNNs can be used to analyze historical price data and find patterns and trends that may hint at potential future price changes in the context of stock price prediction. The internal workings and understanding of CNNs are challenging due to their complexity and number of layers. CNNs are susceptible to overfitting, a condition in which the model performs well on training data but does not generalize well to new data. Predictions could be wrong, and model performance could suffer as a result.

IV. Proposed Work

The proposed work intends to use a Long short-term memory algorithm to forecast stock prices. Long Short-Term Memory (LSTM), a sort of recurrent neural network (RNN) technology, has been used to anticipate stock values. LSTM networks have gained popularity in this business due to their ability to deal with dependencies for a long time and successfully describe sequential data. The basic idea behind utilizing LSTM for stock price prediction is to train a model on a historical dataset of stock prices and then use it to forecast future prices. An input series of previous stock prices is fed into the LSTM network, and it uses its memory cells to learn the long-term dependencies in the given sequence. To predict the future stock price, the network's output is then utilized. To produce more precise predictions, LSTM can capture the non-linear correlations between these elements. Another advantage of using LSTM to predict stock prices is its capacity to deal with noisy and non-stationary data. The unpredictable fluctuations and abrupt adjustments in market sentiment that can affect stock prices make it challenging to estimate future prices using conventional statistical techniques. LSTM's effectiveness in handling noisy and non-stationary data comes at the expense of requiring meticulous feature engineering and rigorous data pre-processing, as visualized in Figure 1, showcasing the intricate architecture involved in these preparatory steps for robust predictive modeling.

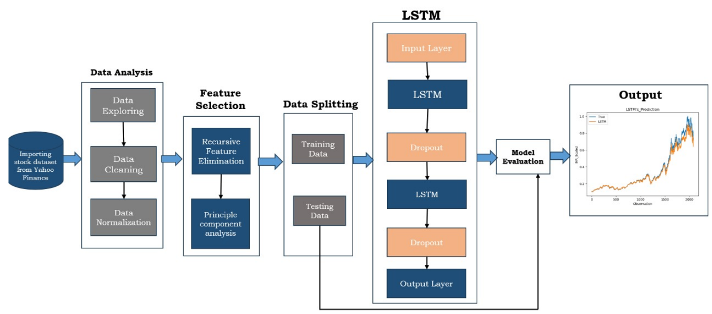

Fig. 1. LSTM Architecture

V. Methods

A. Dataset description

The dataset for this research was acquired from Yahoo Finance, as well as other pertinent values. This dataset spans the years 1980 to 2021. The data includes the prices of stocks at specific periods for each day of the year. It has been partitioned into multiple categories, including open, volume, date, low, high, close, and adj close. For simulation and analysis, just one company's data is used. The complete data set is available in a text file called CSV, which is accessed and turned into a data frame using Python's Pandas library. Following that, the data were normalized using the Python sklearn program and the data. After normalizing the data, it was split into testing and training samples by employing the Python sklearn module. The test set was limited to 20 percent of the total dataset.

B. Data Normalization

Stock price datasets frequently contain features with varying sizes and ranges. All the features are scaled to a common value using data normalization techniques like min-max scaling or standardization. Normalization assists in preventing

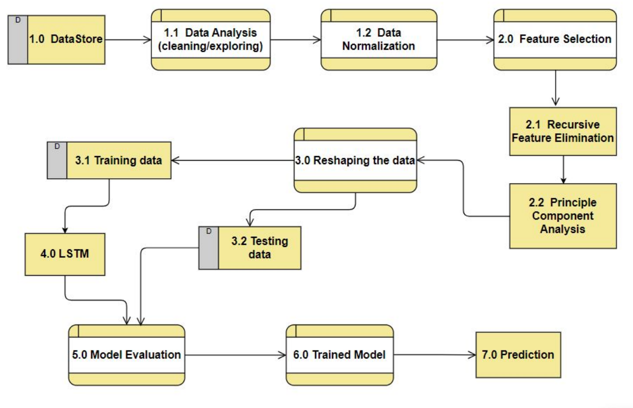

Fig. 2. Data Flow Diagram

the model training process from being dominated by features with greater magnitudes. It ensures that each feature makes an equal contribution to the model and enhances the training process' convergence and stability.

C. Separate features and target variable

By separating the features and the target variable, we can clearly distinguish between the data used for prediction and the data that has to be forecasted. This separation allows us to train the LSTM model by feeding it historical feature data and matching goal values. The model can identify patterns in the features during training and comprehend how the features relate to the target variable, allowing it to predict future feature data that has not yet been observed.

D. Sequence Generation

Create input-output pairs for LSTM training from the pre-processed data. The amount of prior time steps the model will take into account when predicting the upcoming time step is represented by the sequence length, which you should choose. For instance, if the sequence length is set to 30, the model will forecast the stock price for the subsequent day by considering the previous 30 days of data.

E. Feature Selection

Recursive Feature Elimination is a feature selection approach for identifying the most relevant characteristics from a given feature set. RFE systematically removes less significant features from the set until the required number of features is left. This method helps enhance model performance, decreasing overfitting, and concentrating on the key characteristics that lead to reliable predictions. The problem's dimensionality can be decreased by using RFE to find and choose a subset of features that are most important for predicting stock prices. This enables a more concise comprehension of the features' relationship to the target variable.

Principal Component Analysis is a dimensionality-reduction method for reducing the number of dimensions in a dataset while maintaining the most crucial data and reducing information loss. The principle component analysis (PCA) accomplishes this by locating the main components, linear combinations of the original attributes, and capturing the most variance in the data. Stock price prediction datasets frequently contain noisy or redundant information that can degrade the model's performance. PCA can assist reduce the influence of noise by removing the less significant components that are more likely to capture noise or irrelevant variations.

F. Data Splitting

Figure 2, portraying a Data flow diagram, elucidates the training and testing process of the LSTM model.

Training Set: The LSTM model is trained using a training set. The majority of the dataset, or a subset of the historical information, is often included. The data used for the training set includes both the input characteristics and the target variable.

Test Set: The test set is used to determine the outcome of the trained LSTM model on unseen data. The input features are included in the test data, but the target variable is hidden from the model during evaluation. This enables us to evaluate the model's ability to generalize and generate accurate predictions about future stock values.

G. Model Building and Training

For sequence data processing, such as stock price prediction, the recurrent neural network (RNN) architecture known as Long Short-Term Memory (LSTM) is widely utilized. Sequential data, such as historical stock price time series, are highly suited for LSTM modeling. It can effectively capture temporal dependencies and patterns in data, considering both short-term and long-term information. An LSTM model can learn to recognize patterns, trends, and seasonality by examining the historical price sequence, which may help it make more precise stock price predictions. The associations between the input features and the target variable can be accurately captured by LSTM models. The relationship between previous prices and future price movements in stock price prediction can be quite non-linear and impacted by a variety of factors. For simulating the dynamic dynamics of stock markets, LSTM is a good option since it can represent long-term interdependence and non-linear patterns.

H. Model Evaluation

Evaluating stock price prediction with LSTM includes evaluating the LSTM model's effectiveness and precision in predicting stock prices. This evaluation procedure aids in determining how well the model is operating and whether It is capable of making accurate forecasts. The effectiveness of LSTM models for stock price prediction is typically evaluated using the following metrics and methods:

Mean Squared Error (MSE): MSE is a popular statistic for assessing regression models, including LSTM. It calculates the avg. squared difference between the stock price

TABLE I
DATASET SCHEMA

S No.	Column	Data Type
1	Date	Object
2	Open	float64
3	High	float64
4	Low	float64
5	Close	float64
6	Adj Close	float64
7	Volume	int64

predictions and actual prices. Since predicted values are more closely matched to the actual values, a lower MSE indicates better prediction accuracy.

The square root of Mean Squared Error (MSE) is Root Mean Squared Error (RMSE). It calculates the average magnitude of the prediction error. Similar to MSE, a lower RMSE indicates better prediction efficiency.

The average absolute difference between expected and actual stock prices is determined by the Mean Absolute Error (MAE). It provides insight into the average magnitude of the errors. Similar to MSE and RMSE, a lower MAE denotes more effective accuracy in prediction.

R2 score, also known as the coefficient of determination, is a statistic used to assess the effectiveness of regression models, especially LSTM models, in stock price prediction. It determines the fraction of the variance in the selected variable that the predictive algorithm can explain. The R2 score goes from 0 to 1, with a greater value suggesting a better fit between the model and the data.

VI. RESULTS

The dataset schema, as shown in Table I, delineates the structure and data types of the financial dataset utilized for analysis. Derived from sources such as Yahoo Finance, this dataset spans from 1980 to 2021, encompassing crucial attributes like Date, Open, High, Low, Close, Adj Close, and Volume.

The dataset from Yahoo Finance is used to train and test the suggested method. It is divided into training and testing sets. To evaluate the estimated results with the practical outcomes, we converted the analysis into a graphical representation using an LSTM.

The actual trend, represented by a blue line, is juxtaposed with the predictive outcomes illustrated by an orange line in Figure 3, showcasing the LSTM model's forecasted stock price movements. The closeness of these two lines demonstrates the efficiency of the LSTM-based model. After a significant amount of time has passed, the prediction

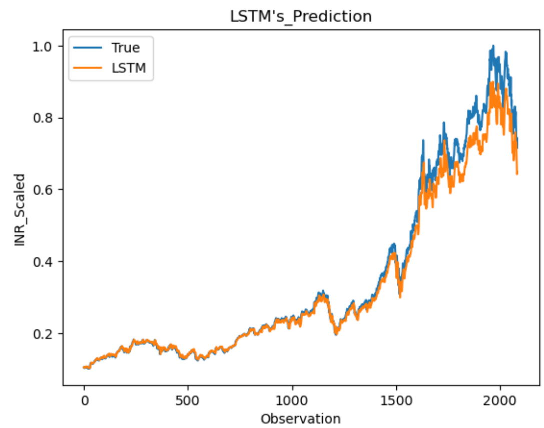

Fig. 3. LSTM Predictions

TABLE II
PERFORMANCE ANALYSIS

Proposed Model	LSTM with RFE and PCA
RMSE	0.06
MSE	0.0039
MAE	0.037
R2 Score (Train Data)	0.87
R2 Score (Test Data)	0.97

approximates the true pattern. The collection of data from Yahoo Finance is used to train and evaluate the proposed approach. It is segregated into sets of training and testing. After applying the LSTM algorithm to generate a prediction, we transformed it into a visual representation to demonstrate the prediction's results with the actual result. By computing the accuracy score and overfitting loss of our model, we can assess its performance. An improved fit between the model and the data is shown by a higher R2 score, which ranges from 0 to 1. The model resulted in an R2 Score of 0.87 on Train data and 0.97 on Test data. The RMSE calculates the mean value of the error in predictions. The Root Mean Squared Error (RMSE) quantifies how well the data fits the model. With a total of 6% Root Mean Square Error in prediction, the accuracy of our proposed work is 94%, implying that our estimated outcomes are 94% equivalent to real stock values. The model accurately anticipated the outcomes, with a mean absolute error of 0.037. and a root mean squared error of 0.0039.

VII. CONCLUSIONS

The research attempted to predict the future prices of a company's stocks using predictive machine-learning strategies with better reliability and precision. The proposed approach forecasts stock prices using a machine-learning LSTM. When compared to LSTM, the model gets the greatest accuracy by combining LSTM with RFE and PCA. The highest accuracy obtained is 94% This synergistic integration significantly improved the predictive accuracy, outperforming the standalone LSTM model across various performance metrics, as outlined in Table II above.

REFERENCES

[1] Shahzad Zaheer, Nadeem Anjum, Saddam Hussain, Abeer D. Algarni, Jawaid Iqbal Sami Bourouis, and Syed Sajid Ullah, A Multi-Parameter Forecasting for Stock Time Series Data Using LSTM and Deep Learning Model", Mathematics 2023. Forecasting for Stock Time Series Data Using LSTM and Deep Learning Model", Mathematics 2023.

[2] Nusrat Rouf, Majid Bashir Malik, Tasleem Arif, Sparsh Sharma, Saurabh Singh, Satyabrata Aich and Hee-Cheol Kim, "Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions". Electronics 2021.

[3] Guangyu Ding, Liangxi Qin, International Journal of Machine Learning and Cybernetics (2020) 11:1307–1317, Springer.

[4] ADIL MOGHARA, MOHAMED HAMICHEB "Stock Market Prediction Using LSTM Recurrent Neural Network", in International Workshop on Statistical Methods and Artificial Intelligence. Elsevier 2020.

[5] Naadun Sirimevan; I.G. U. H. Mamalgaha; Chandira Jayasekara; Y. S. Mayuran; Chandimal Jayawardena," Stock Market Prediction Using Machine Learning Techniques ", in International Conference on Advancements in Computing. IEEE 2019.

[6] Jingyi Du, Qingli Liu, Kang Chen, Jiacheng Wang, "Forecasting stock prices in two ways based on LSTM neural network", 3rd Information Technology, Networking, Electronic, and Automation Control Conference. IEEE 2019

[7] Indrani Bhattacharjee, Pryonti Bhattacharja," Stock Price Prediction: A Comparative Study between Traditional Statistical Approach and Machine Learning Approach" in 4th International Conference on Electrical Information and Communication Technology (ICEICT). IEEE 2019

[8] Subhadra Kompella, Kalyana Chakravarthy Chilukuri, " STOCK MARKET PREDICTION USING MACHINE LEARNING METHODS", in International Journal of Computer Engineering and Technology (IJCET). IJCET, 2019.

[9] Ishita Parma, Navanshu Agarwal, Sheirsh Saxena, Ridam Arora, Shikhin Gupta, Himanshu Dhiman, Lokesh Chouhan, "Stock Market Prediction Using Machine Learning," First International Conference on Secure Cyber Computing and Communication(ICSCCC). IEEE 2018.

[10] Sreelekshmy Selvin, Vinaya Kumar R, Gopalakrishnan E.A, Vijay Krishna Menon, Soman K.P, "STOCK PRICE PREDICTION USING LSTM, RNN, AND CNN-SLIDING WINDOW MODEL". IEEE 2017.