# Survey on Stock Market Prediction using Machine Learning Algorithms

**Vipul Bhagwat[1], Naman Kumar[2], Rajat Kulkarni[3], Prajwal AM[4], Sagar Salunke[5]**

*[1-4]Dept. of Computer Engineering, Pimpri Chinchwad College of Engineering Pune, Maharashtra, India*
*[5]Professor, Dept. of Computer Engineering, Pimpri Chinchwad College of Engineering Pune, Maharashtra, India*

**Abstract -** Stock Market Prediction is a challenging and trending topic for researchers in recent years. Many companies invest in stock market so that predicting the future movement of stock will minimise their risk and will increase their profit. A lot of study has been carried out in recent years using traditional methods, machine learning algorithms and deep learning methods to predict the market. Prediction can be made on previous historical data available. In this paper we have done a survey on different algorithms which are used in stock prediction like different regression techniques, sentiment analysis, SVM, Random forest, Neural networks, LSTM.

***Key Words***: **Stock prediction, Regression, SVM, Random forest, Neural networks, LSTM, Sentiment Analysis**

## A. INTRODUCTION

Stock market prediction has been very tiresome and troublesome since markets's existence. Stock market is an area where prediction does not follow any rules as the nature of the market is very volatile. Due to its volatile nature and high risk high returns on investments, 95% of the traders make losses in the stock market because they try to gamble by randomly speculating the prices or movement and lack a proper trading setup.

The share market is based on the concept of demand and supply. If the demand for a particular company's stock is higher and the supply is low then that company's share price would tend to increase and if the demand for company's share is low then the company share value tends to decrease. The successful prediction of a stock's price by its analysis could lead to a significant profit. For this, extremely large historic data sets are used to depict varying conditions and thus reaffirming the belief that the time series patterns possess significant predictive power with a high probability to generate profitable trades and high returns for investment in business.

Stock market prediction is performed without considering the sentiment analysis. This is one of the reasons for the efficiency to be low. A system is needed where the share price can be calculated efficiently and accurately considering all the factors.

In this paper we have done a study on different machine learning algorithms and deep learning algorithms like linear regression, Neural networks, LSTM which are used in stock market prediction.

## B. RELATED WORK

Dou Wei[1] has analysed and compared a variety of neural networks prediction method and finally chosen LSTM((Long Short-Term Memory).The data used for prediction is taken from Oriental fortune website by using crawler and stored in database after performing cleaning operations.The paper uses attention-based LSTM(long short-term memory) and consists of four parts : input Layer, hidden layer, attention layer and output layer. The inputs to the models were trading date, opening price, closing price, lowest price, highest price and daily volume of the stock .For standardization of input data maximum and minimum method is used and formula is as follows:

Standardized input data = (original data - minimum)/(maximum - minimum)

Dinesh Bhuriya[2] have used different regression techniques to predict TCS stock price .They have mainly implemented linear, polynomial and RBF regression model out of which linear regression model provides best result based on confidence value. They have used Open price,High price, Low price and Number of trend as input variable and Close price as output.

Omveer Singh Deora[3] have used two different machine learning models for stock market prediction. LSTM is used for historical data and CNN for current data. For both these models their respective data sets are split in 80:20 ratio, where 80% is used for training and the remaining 20% is used for testing.

The purpose of applying both these models is that LSTM models are better skilled to find any relations in the data set and use them to predict future prices whereas CNN primarily emphasize on current data for its prediction not requiring any information on historical data.

Dr. Devpriya Soni[4] has proposed a model which would check itself at each stage for the correctness in analyzing share trends. This model is a simplified one and is only applied on the historical data. The accuracy provided by the authors is 96.60% which is a very high number. We can use this model and combine it with sentiment analysis for a more thorough and efficient prediction.

The proposed model:

1. Divide the dataset into two parts-testing and training.
2. In the training set:
    a. Calculate the change in share price over a period of time for every day and take mean
    b. Share price can increase or decrease
    c. Prepare a matrix z for storing zeroes and ones having no. of rows equal to equal to the size of training set and columns equal to no. of factors included for the share
    d. If share price increases assign 1 else 0
    e. Count no. of 0s and 1s-if no. of 1s>0s then the share is a good share else a bad share
    f. Create a prediction matrix y
    g. If the stock shows positive growth then the value of the share price is increased by the change we calculated earlier and the first row of the newly created prediction matrix is marked as 1 otherwise if the stock shows negative growth then the value is decreased and correspondingly the first row of the prediction matrix is marked as 0.
3. In the testing set:
    a. Prepare a similar matrix x by storing 1s and 0s on the basis of actual rise and fall of the share price
    b. Now check first day's prediction by comparing values in x and y matrix
    c. If the match is found, prediction is right and we further proceed for the next day accordingly by what we did for the previous one( i.e. if share price is increased in the previous prediction then for the next prediction as well ,we do the same and vice versa).
    d. If the values of x and y are not matching then we correct the error by a factor of twice the change we calculated earlier of the share price value for the next prediction.
    e. So for example if the share price falls and we predicted a rise so for the next prediction we now alter the predicted value of the previous day by a factor of double the change and reduce its(predicted price) value accordingly.
    f. Also the next field of y matrix is marked as opposite to the previous one. In this way, the predicted price is calculated and subsequently the accuracy of prediction by comparing it to actual values.

Aparna Anant Bhat[5] proposed a system which was composed of modules like Data Collection, Technical Analysis, Prediction with Neural Networks and Sentiment Analysis.

In data collection, they gathered data of historical prices and volumes of the stocks from NSE India and google finance. This data was used for technical analysis components and neural networks. News and blog articles were also collected with the date of publication for sentiment analysis. But the predictions for intra-day trading were not available everyday at times and the frequency of opinions, news articles and comments varied for different working days.

In technical analysis they implemented the technical indicators such as MACD,EMA and RSI using the historical data collected. They made predictions based on these technical indicators.

For prediction using neural networks, the authors tested the neural networks model with different sets of input and hidden layer neurons and different activation functions. They used activation functions like sigmoid and hyperbolic tangent functions for triggering the neurons. They used back propagation to further optimise the prediction values.

The training data consisted of closing prices and volumes of each day.

In sentiment analysis, the authors tried to check the positive and negative sentiments from the reviews /news/ blog articles to further improve the accuracy.

In this research, the authors concluded that though combining Technical Analysis with technological methods like Neural Network gave better performance, addition of Sentiment Analysis further increased the results as stock prices very much dependent on day to day activities and emotions in the market and the fundamental analysis.

Juan Ricardo Rivera Peruyero[6] worked on 2 simple speculation methods in terms of the mean losses and benefits they produce under different market conditions, after 200 days of activity. They developed an application to mine the prices of financial assets from free web information providers.

They developed the application using Matlab because this software offers a lot of facilities for testing new algorithms quickly and visualising the results. With Matlab it was also easy to implement functions for mining web information.

They extracted the free information from the finance section of Yahoo and Google. The application starts once the markets of interest are closed, the series of day prices are automatically updated and, after the signals are computed, the new orders to be applied the next day are presented, always at the opening market price. With the application it was possible to visualise the generation of the signals and the orders day by day. Therefore, it is possible to see in which cases the system fails and provides and gain an idea of how to improve strategies.

Liu,Yifan and Guangzhong[7] (2018 IEEE) have used the LSTM ( Long Short Term Memory) Recurrent Neural Networks to analyse the stock data and predict the future value. Their 3 Layer Long Short Term Memory model shows that good forecasting results and the accuracy is as high as 72% for short term data, eventually as the model improves i.e more stack layers of Long Short Term Memory model, the accuracy of prediction and forecasting of the stock price will be increased.

Meghna Misra[8] had discussed prediction of the stock market using different Machine Learning algorithms. They had compared the Linear Regression, Support Vector Machine and Random Forest algorithms for getting higher accuracy and least error. The analysis shows that Support Vector Machine has high accuracy on non linear classification data whereas Linear Regression is the preferred algorithm if the available model is that of regression and the Random Forest shows high accuracy on binary classification model and multilayer perceptron offers the least error in prediction of the stock trend.

## C. ALGORITHMS

### 1) Regression

In Regression we predict Stock price by implementing a model based on one or more attributes like closed price, open price, volume etc.The goal of regression is to model the linear relationship between the variables which are dependent and independent.

$$y = \beta0 + \beta1\ x1 + \beta2\ x2 \ldots + \beta k\ xk \ (1)$$

y – Outcome, β0 – Intercept, β1, β2, βk – Partial Regression Coefficient, x1, x2, xk – Input Parameters.

### 2) Support Vector Machine

Support vector machine comes under a supervised learning model which also includes associated learning algorithms to identify the data. Classification and regression methods are used in this process to analyse the prediction. The market price is predicted by implementing the margins adopted. The support vector y's value is calculated with the help of two attribute case

$$y = w0 + w1\ x1 + w2\ x2 \ (1)$$

y – Outcome, w0, w1, w2 – these 3 weights to be learned by SVM model, x1, x2 – input parameters.

### 3) LSTM

LSTM neural network is a special kind of recurrent network [9].Two problems which are generally faced by traditional recurrent neural network are Long term dependency problem in RNNs and Vanishing Gradient & Exploding Gradient which can be solved using LSTM. LSTM keep the error at a more constant level, so that a recursive network can be a lot of time to learn, so as to open the establishment of a long distance causal link.

LSTM is made up of three gates, Input gate, Forgot gate and Output gate and these gates use sigmoid activation function.

The equations for gates in LSTM are:

$$i_t = \sigma\ (w_i\ [h_{t-1}, x_t] + b_i)\ (1)$$

$$f_t = \sigma\ (w_f\ [h_{t-1}, x_t] + b_f)\ (2)$$

$$o_t = \sigma\ (w_o\ [h_{t-1}, x_t] + b_o)\ (3)$$

where $i_t$ = input gate, $f_t$ = forgot gate, $o_t$ = *Output gate,*

$\sigma$ = sigmoid function, $wx$ = weight for the respective gate(x) neurons, $h_{t-1}$ = output of the previous lstm block, $x_t$ = input at current timestamp, $bx$ = biases for respective gates

In the above equations the first equation is for Input Gate which tells us that what new information we're going to store in the cell state.

Second equation is for the forget gate which tells the information to throw away from the cell state(memory).

Third equation is for the output gate which provides the activation to the final output of the lstm block at timestamp 't'.

The heart of a LSTM network is it's cell or say cell state which provides a bit of memory to the LSTM so it can remember the past.The equations for the candidate cell state, cell state and the final output are :

$\tilde{C}_t = \tanh(w_c\,[h_{t-1}, x_t] + bc)$ (4)

$C_t = i_t * \tilde{C}_t + f_t * C_{t-1}$ (5)

$h_t = o_t * \tanh(C_t)$ (6)

where $C_t$ = cell state(memory) at timestamp(t)

$\tilde{C}_t$ = candidate for cell state at timestamp(t)

Now, from the above equation we can see that at any timestamp(t), our cell state knows that what it needs to consider from the current timestamp ( $i_t * \tilde{C}_t$) and what it needs to forget from the previous state( $f_t * C_{t-1}$ ).

Lastly, we filter the cell state and after that it is passed through the activation function which predicts what portion should appear as the output of the current lstm unit at timestamp t. A block diagram of LSTM is given below.

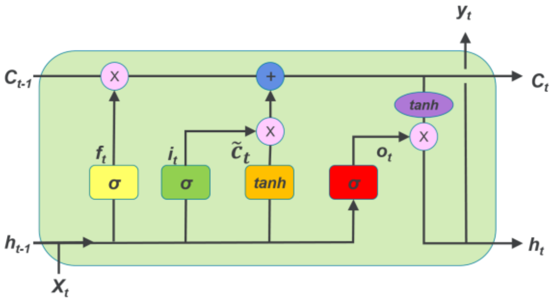

**Fig1**: LSTM Block Diagram

## CONCLUSION

The purpose of this survey is to study different traditional methods, machine learning and deep learning algorithms which are used in stock market prediction. These algorithms include different regression techniques, SVM, Random forest, Neural networks, LSTM. SVM shows high accuracy on non-linear classification data whereas LR is the preferred algorithm if the available model is that of regression, due to its high confidence value. LSTM are preferred over other neural networks because of its memory cell and gives better accuracy. It is also observed that the accuracy of a model increases if we add sentiment analysis to it as the stock market is sensitive towards people's sentiments.

## REFERENCES

[1] Dou Wei, "Prediction of Stock Price Based on LSTM Neural Network" in International Conference on Artificial Intelligence and Advanced Manufacturing(AIAM), 2019..

[2] Dinesh Bhuriya, Girish Kaushal, Ashish Sharma, Upendra Singh, "Stock Market Prediction Using A Linear Regression" in International Conference on Electronics, Communication and Aerospace Technology ICECA, 2017.

[3] Omveer Singh Deora, Pawan Jha, S.T. Sawant Patil, T.B. Patil, S. D. Joshi, "Monitoring and Training Stock Prediction System For Historical & Live Dataset using Lstm & Cnn," (IJITEE)(2019)

[4] Dr. Devpriya Soni, Sparsh Agarwal, Tushar Agarwal , Pooshan Arora, Kopal Gupta, "Optimised Prediction Model For Stock Market Trend Analysis,"(IC3)(2018)

[5] Aparna Anant Bhat, Sowmya Kamath S, "Automated Stock Prediction and Trading Framework for Nifty Intraday Trading" in 4th ICCCNT - 2013 July 4-6, 2013, Tiruchengode, India.

[6] Juan Ricardo Rivera Peruyero , Pere Marti-Puig, "Web-based system for evaluating day trading strategies," 2011 7th International Conference on Next Generation Web Services Practices 253

[7] Siyuan Liu, Guangzhong Liao, Yifan Ding, "Stock Transaction Prediction Modeling and Analysis Based on LSTM," - IEEE (2018)

[8] Meghna Misra, Ajay Prakash Yadav, Harkiran Kaur, "Stock Market Prediction using Machine Learning Algorithms: A Classification Study," (ICRIEECE) IEEE - (2018)

[9] Divyanshu Thakur, LSTM and its equations. https://medium.com/@divyanshu132