International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 12 | Dec 2020 www.irjet.net p-ISSN: 2395-0072

Stock Market Trend Prediction using Supervised Machine Learning Algorithms

Naeimuddin[1], B Vijayakumar[2]

[1]*Vidya Jyothi Institute of Technology, Aziznagar Gate, Chilkur Balaji Road, Hyderabad, 500075, INDIA*
[2]*Vidya Jyothi Institute of Technology, Aziznagar Gate, Chilkur Balaji Road, Hyderabad, 500075, INDIA*

Abstract - *Stock trend prediction is the difficult problems, due to stochastic nature which is disturbing both researchers and analysts over half a century. And also for those who are eager to invest in stock market but due to the risk of fluctuation in stock they are not much confidence to invest in stock market, so to overcome this problem our research work focuses to observe fluctuations in stock prices. In this research work four prediction algorithms are proposing using historical data to predict the stock market movements. The proposed supervised algorithms are K-Nearest Neighbor (KNN), Random Forest, Support Vector Machine (SVM) and Linear Regression. The historical data will be collected from Yahoo finance, Kaggle, Towards Data Science, NSE India. The results obtained from the different supervised algorithms for all four algorithms. We have obtained the result from different company dataset and analyzed, and overall performance and we got the random forest algorithm is the best for prediction we got highest 95.8% as a result.*

Key Words: Machine learning 1, prediction model 2, KNN 3, Random forest 4, SVM 5, Linear regression 6.

1. INTRODUCTION

Very arrogant role of a stock trend in the rapid money-spinning homeland like India. Such like India and other workup country growth decide the performance of the market if the stock market will in a loss then country money-spinning will also be in loss or stock will go higher then also country money-spinning will be high .in other words the country growth is dependent on the stock growth. At any place, a maximum of 10% of people emerges in the investment of the stock exchange because of misconceptions that buying or selling the stock is fake and fraudulent. If this doubt about the stock market can be changed then only it will bring more confidence in peoples. Strategy can aware and bring more people towards the stock trend at a place. The more likely desire output of predicting way alter the people's attitudes. Machine learning algorithms also help to predict future trends. So for this prediction, we are using four predicting models name linear regression, Support vector machine, K-nearest neighbor, and Random forest. Here, we are using these algorithms for stock market trend prediction to predict the future values which will help people to invest their money for more profit and for more exact value of a stock, and only doing this prediction on stock trend it will also help to grow country growth and economy. In this paper we are comparing all algorithms with each other and getting the best fit for large data set as well as for small data set, we have a different kind of company stock market data set such as Google, Amazon, MCD, and IBM. We have collected these stock market data set from various sites like Kaggle, NSE-India, etc.

2. SUPERVISED ALGORITHMS

Supervised learning is teacher-oriented, or we can say that where we get desired output, in which the system already knows about output, that what output will come after, supervised learning to which we train and teach them about the difference between two variable, where a machine has to classify according to the input given by the user. Where the algorithm learns from defining the shape, color, and functions, algorithms such as SVM, KNN, logistic regression are the supervised machine learning algorithms. Where multiple regression has the parameter which will be calculated through depreciate the sum of error.

$$\sum_{i=1}^{n}(yi - yj)_2 \quad (1)$$

Where, yi actual data of output variable, yj will be predicted data, n total numbers of record in the dataset.

The prediction will be achieved by weight and estimating by the feature with the known actual value called supervised learning algorithm.

2.1. Random Forest: The algorithm random forest is a conjecture, which combined the decision tree into leveled form for less quantity of data samples which scale to enhance the model accuracy and over- equilibrium control. The size below the sample is almost the coequal same as the earliest size where the case is underlined by the other input.

2.2. K-Nearest Neighbours: Definition: K nearest neighbour sometimes called slow learner because it does not make the intramural pattern. But on the other side it balances training data conditions, the division is calculated by effortless engagement of the majority of the close points.

2.3 SVM: The support vector machine which supports the depiction of data set like space marks divide into sections with as distinct space as long as achievable. The new model was drawn in that gap which was estimated that they would

be in a category depending on where space will carry off into.

2.4. Linear regression: Linear regression is a widely popular model of archetypical. Which is commonly used by people for predictive archetypical modeling. The aim of the continuous variable is to continue, and the non-continuous set of data will differ, the behavior of the backline is a line.

3. RELATED WORK

Sneh Kalra et al. in 2019, in this paper authors, did research on the fluctuation of stock market prices with respect to the relevant new articles of a company. They used classifier Naïve Bayes to separate negative or positive statements for prediction purposes based on daily news variance the social media data, blogs data may be considered for future work [1].

Aditya Menon et al. in 2019, this paper is focused on a review of neural model for forecast the stock tread after reviewing on a neural model they think that The long short term memory algorithm for predicting the economic information in confluence into the trendy era, this would be prioritized algorithm for forecasting [2].

Ashish Sharma et al. in 2017, they found that regression analysis is mostly used for stock market trend prediction they survey of regression technique for stock prediction using stock market data. In the future result could improve by using more numbers variables [3].

Mu yen chen et al. in 2019, authors did research to calculate the impact of news articles on the stock prices using deep learning approach LSTM (long short-term memory) and they think this study can predict the stock market trend [4].

Andrea Picasso et al. in 2019, in this research, authors worked which will alliance the economic and elemental analysis for market trend prediction through the various kind of application and automation methods neural network is machine learning technique the problem of trend stock and those are charts with forecasting data. As an input data sentiment of a news article is exploited. According to their research the problem in the most problematic accomplishment among the use of information about news astral one-off. To overcome this problem in the future the proper feature fusion technique will be suitable for the future [5].

Gangadhar Shobha et al. in 2018, this paper provided a full overview of machine learning techniques which will help to reader for use of equations and concept the author discussed about three type of all machine learning technique and also various kind of metrics like accuracy, confusion matrix, recall, RMSE, precision and quintile of errors. The author thinks that this review can help those people who are new to machine learning because most of the people confuse to use most of the machine learning techniques for prediction or others [6].

Suryoday Basak et al. in 2018, the author developed an experimental framework for predicting stock prices whether the price goes up or down in this experiment author uses the two algorithms name as a random forest classifier and Gradient boosted decision' n trees, and they got more accuracy in comparison to others research papers where others experiments got 50% to 67% results on the other hand according to the author of this paper they got 78% result accuracy for long term window. In the future, they could use the build boosted tree model for short term data window [7].

Arash Negahdari kia et al. in 2018, as the stock prediction so many experiments and models, have been developed for prediction purpose on historical data like as in this paper the author present HyS3 graph-based semi-supervised model and through a network views Kruskal based graph algorithm called ConKruG. In the future they think social media data, Twitter data could be used for the prediction of stock for better results using these algorithms [9].

Bruno Miranda et al. in 2019, in this paper the survey of bibliographic techniques that focus on text area for research the author works on the prediction of financial market values by using the machine learning models support vector machine (SVM) and neural network with data set from North American market new models may have opportunities for north American market data for prediction purpose in future [10].

K. Hiba Sadia et al. in 2019, author aim for this paper us to preprocess the raw data firstly then they are doing a comparison between random forest and SVM algorithm the main aim of the author is to find out the best algorithm for stock trend prediction in the last they have given the best-fit algorithm for future stock forecasting which is random forest algorithm for future work they think that for getting more accuracy in result the adding of more parameters can be good [11].

A. Akash et al. in 2019, the author introduced two more algorithm name as "LS SVM" which is least square support vector machine and another one is "PSO" (particle swarm optimization) the work "PSO" basically select the best-unbounded parameter with the "LS SVM" to reduce the overfitting and some technical indicators which will basically enhance the result accuracy. On the other hand at the same time, the proposed algorithm is being compared with artificial neural network model [12].

Aparna Nayak et al.in 2016, in this paper authors, woked to predict the stock market trend by using the supervised learning methods, here authors predicted the data based on daily live data which is directly calling by the program using yahoo financial website and also predicting the monthly

based prediction where in this paper they got a better result for daily live prediction instead of monthly prediction further future work they think if we consider more sentiments to the monthly [prediction that would also generate the best result [13].

Nuno Oliveira et al. in 2016, in this paper the author purposed a methodology by which they can access the value of stock prediction and microblogging data they used, for stock prices and return indices and some more like a portfolio. For this experiment, they have used huge data of Twitter, for all this experimental work they use Kalman filter to merge the microblogging data and some external sources and as a result, they found twitter data and blogging data were relevant for the purpose of forecasting these datasets were very useful. This result can be improved by using some more and different data such as social media datasets and others [14].

Han lock Siew et al. in 2017, the author in this paper used regression technique for finding the accuracy in the forecasting of a stock trend for all this experiment they used WEKA software which used for data mining and machine learning algorithms to execute them, the dataset they used which contains heterogeneous values and which is used for handling of currency values and functional ratios. The dataset for calculating the stock movement is collected from Bursa Malaysia for forecasting purposes. For the future extension, the authors thought that the forecasting using the regression method can be improved by using the more standardized ordinal format of data [15].

Smruti rekha das et al. in 2019, in this paper authors, used firefly method for forecasting the stock prices as an input dataset author collect from four different websites name as NSE-India, BSE, S&P 500 and FTSE, and all collected dataset is well transformed by using proper mathematical formulas by using the backpropagation, neural network and more two methods used for prediction, forecasting according to the time horizon of alternate days 1 day, 3 days, 5 days and so on. For future work, there may be some chance to get more accurate results by giving more parameters to the implemented algorithms [16].

Dattatray P.Gandhmal et al. in 2019, authors had written the paper on the review of stock prediction techniques. In this paper, the authors reviewed about 50 published research papers according to the publication years, and the authors suggested the best technique for prediction. KNN and fuzzy-based techniques as the authors suggested are the best techniques according to the review such as KNN, SVM, SVR, and much more but these two techniques can be more effective for the purpose of using historical data. In the future, they will review more papers to get the best-fit algorithm for prediction [17].

4. PROPOSED WORK

In this research paper, we implemented trend forecasting of stock market "the stock market trend prediction" using supervised methods like "linear regression", "random forest algorithm", "support vector machine algorithm", and K-nearest neighbor algorithm

4.1. SYSTEM ARCHITECTURE

Here is the proposed work system architecture, which we have depicted through the stepwise structure, In the first step we are starting by giving raw data to our trained algorithm which will pre-process the data by using python libraries which is also a feature extraction part, Where it will the cleanup data by using data pre-processing method after that we have divided our data into two parts where 70% of our data is trained and remaining 30% of data which is for testing by using the trained algorithm after all this process we will only get our predicted data, as shown in figure 1.

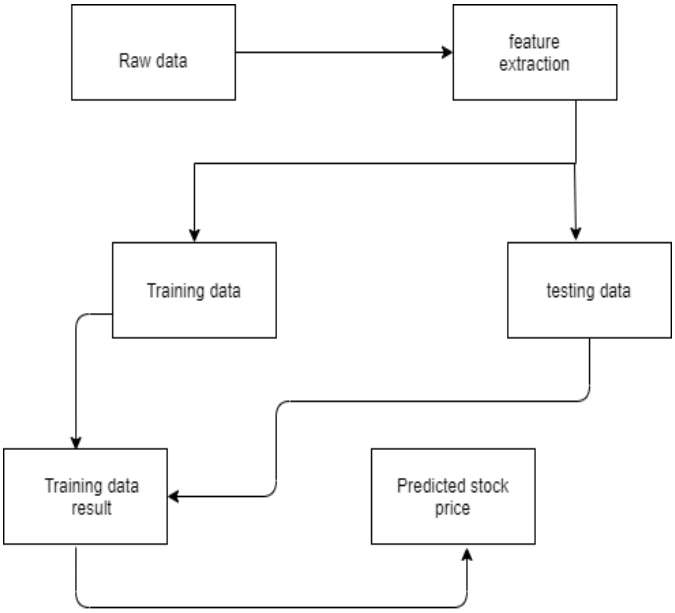

Fig -1 Proposed System Architecture

4.2 DATA PREPARATION

Raw data is at high risk for noise, lost values, and inconsistencies. Data quality affects the results of data mining. To help enhance the affection of information or, as a result, which is the output of extraction the unstructured information returns in advance to enhance the effectiveness or simplification of extraction operations. This is the only step that is very tough in the data extraction that negotiates and tries to rebuild and alteration of the original information.

4.2.1. Data integration

The chance for your information or data anatomizing function would include in data integration, that means the information will alliance from many cradles into a compatible stash, such information in the stash. Which can combine a huge amount of information, data files, and more in some files. So we can get many hurdles at the time of information compilation. The combination of tricks could deceptive. Whereas the other word companies which could be matched from many sources of data. Which is basically called business index the causes. Such as, a computer engineer or analyst cannot be sure about the person's information details in one database, as well as the customer number for the same business as a refers. Huge information and data repositories usually contain data at a very huge level which is called metadata. Which we say that data is about data, which normally can help to obtain or abate the mistakes in integration process retrenchment which is also trouble the adjective which should not any requirement, therefore, that is not based on the \\ " other table, such as annual income"

4.2.2. Data transformation

In data conversion, data is converted or compiled into appropriate mining forms. Data modification may include some points such as:

1. Formalize, the affection of measured details that can be depicted in the range of 0 to 1.
2. Self-composed, unwanted information or can say noise that casually contains the data. These processes having bent and turning of information.
3. Agregassion, where summarizing or collating data is used. The information which deals on the basis of daily could be compiled or calculate the total income of annually and monthly subtotal. Which is commonly maintained to build an information set of to the anatomizing whole collective grained.
4. The data which is generalized, where zero level and 'old' (green) information which cut out with the high place of information by the exploit of conceptual ordering. Such as, a separate feature which is arterial might be tailored to other ordering abstraction, such as a rural region same as, number values, or age, may be included in the map of high-level concepts, such as young, middle-aged, and older.

4.2.3. Removing extra values or cleaning of data

The dataset which is basically anatomized by the extraction method which is not completed by the process (deficit of adjective which is specific interest affection which contains aggregated into), bodacious/ noisy (means that which have error values and external data which is exactly different what we expect). Capricious not compatible with another fact (having inconsistency into code area which basically used to clarify dataset). The noisy changeable which are at the same area of huge existing in the real-world source of data or its repositories, the occurring of incomplete information have many reasons. Attributes in concern might not eternally operable. As the data of transaction which depends on customer information, some information cannot be encaged where it may be advised arrogant during login. Appropriate information cannot be recorded due to misunderstandings, or due to mechanical malfunction. Data that did not match any other recorded data may have been deleted. In addition, historical recording or modification of data may not be considered. Missing data, especially duplicates with missing values some of the symbols, might be entered. Given information might be arresting, with not corrective liability set of data, due to a given array of dataset tools given could improper. That could error of people and machine automation arising from the presence of information. Inaccuracy in information transfer might be raise. That is technical boundaries, likewise bounded buffer assize in linking and confide or syncing. The information which is not correct might have appeared from the changeable name or meetings and use of information canon. Coequal requires information is to be fine. The information purification methods beaver "clean up" information through filling with lost values, easily out sensitive information, recognizable and deleting sales, Disassociate capriciously. Noisy information might have some disarrangement about the extraction process. Albeit maximum extraction methods which are having some way of negotiation with imperfections and worthless information which is not every time solid. On the other hand, they can focus to abate adding enough information there is a task at hand. Accordingly, a practical preprocessing way to use own information in other ways to clean up data.

5. EXPERIMENT

In this experiment we implement the four supervised machine learning algorithms, name as linear regression, support vector machine, random forest, and k-nearest neighbor using these algorithms we predict the stock market trends.

5.1 Dataset

Information with which we are going to be work with downloaded or composed from Kaggle, towards data science, and NSE-India in the form of an excel file with the extension of .csv. The data is available in the numeric form contains a date, low, open, high adjacent close price with 1260*12(row, column). In historical data values, daily price movements are performed using the close stock price, and the difference between adjacent day close prices is calculated by subtracting the previous day price from today's approximate price. If the value obtained is a negative number then the trend is down otherwise high, as depicted in table 1.

Table -1: Amazon dataset.

S.No	Date	Open	High	Low	Close
0	04-01-2010	136.25	136.61	133.13	133.89
1	05-01-2010	133.42	135.47	131.80	134.69
2	06-01-2010	134.60	134.72	131.64	132.25
3	07-01-2010	132.00	132.32	128.80	130.00
4	08-01-2010	130.55	133.679	129.029	133.5200

5.2 Result

In this research "stock market trend prediction using supervised machine learning algorithm" results are provided in this section all of these processes implemented using a python programming language. These results obtained from the different supervised algorithms for all four algorithms we have obtained the result from different company dataset and analyzed, and overall performance will be considered by comparing all the results and then we will get our best algorithm for prediction.

Table 2 - Predicted value for Microsoft

Name of machine learning algorithms.	Microsoft dataset	
	RMSE %	r2_score %
Random forest	3.1	95.8
Support vector machine	3.4	93.2
K-Nearest neighbor	3.4	93.9
Linear regression	4.0	90.5

Table 3 - Predicted value for Google

Name of machine learning algorithms.	Google	
	RMSE %	r2_score %
Random forest	45.01	94.6
Support vector machine	90.03	74.8
K-Nearest neighbor	44.63	94.3
Linear regression	63.5	88.0

Table 4 - Predicted value for Amazon

Name of machine learning algorithms.	Amazon	
	RMSE %	r2_score %
Random forest	68.6	93.8
Support vector machine	73.1	76.8
K-Nearest neighbor	81.0	90.1
Linear regression	86.2	89.3

Table 5 - Predicted value for Alibaba

Name of machine learning algorithms.	Alibaba dataset	
	RMSE %	r2_score %
Random forest	13.5	92.1
Support vector machine	19.4	82.7
K-Nearest neighbor	14.1	90.8
Linear regression	21.8	82.8

In the above table 2, 3, 4, 5 we have provided all expected values for all four algorithms, values for RMSE (Root Mean Square Error) and the r2_score value which is our predicted value, from companies name like Microsoft, Google, Amazon, and Alibaba the dataset which we are using is already preprocessed which we have collected from some online websites, algorithms are random forest, support vector machine, K-nearest neighbor and linear regression are used.

Table 6 - Comparison of results with each algorithm

Companies	Machine learning algorithms.			
	Random forest	SVM	KNN	Linear regress.
Microsoft	95.8	93.2	93.9	90.5
Google	94.6	74.3	94.3	88
Amazon	93.8	76.8	90.1	89.3
Alibaba	92.1	82.7	90.8	82.8

5.3 Comparison of the predicted result

In this step we have compared all algorithms predicted values with each other to get the best prediction algorithm, here in table 6. We have given different company dataset for each algorithm and we are getting values in percentage. In the given table we are getting a higher value for the *Random forest* algorithm which performed well for all the given data set and also the lowest value for all the dataset we are getting form *SVM* and also second highest result we are getting from k-NN method where linear regression is also third highest. So as we have discussed before that the aim of this paper is to find out the best prediction algorithm from the given four algorithms, by the result which we have got the *Random forest* is the best-fit algorithm for the prediction purpose, as shown in table 6. The comparison of the four algorithms is depicted in figure 2.

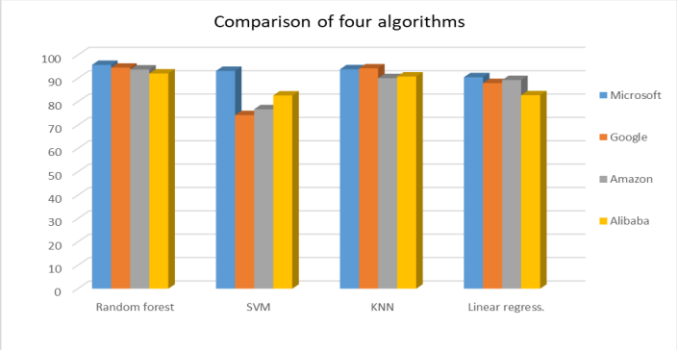

Fig. 2 – Comparison of four algorithms

6. CONCLUSION

In this paper, we implemented Random forest, SVM, Linear regression, and K-Nearest Neighbor algorithms which are supervised algorithms. We have implemented all four algorithms which have been functional for the stock prediction, as we have calculated the result for all different algorithms, we got to know that random forest algorithms are the best and more suitable for prediction purpose. As we have calculated this result through different kind of data points form lateral data, this algorithm random forest is the best for prediction and it will be very effective and profitable for those who invest their money in the stock market, hence this is the trained algorithm using large dataset collection of historical data with this algorithm we got accuracy in between 91% to 95%. Which is actually great output. On the other hand, as we have implemented three more algorithms where KNN performs the second best algorithm for the prediction for this algorithm KNN we got the accuracy in between 90 % to 93%. And for linear regression, we got the accuracy in between 80% to a maximum of 90 %. And we got accuracy for SVM which is very less in percentage as compared to these three algorithms which are 75% to 85%, this is the smallest number of accuracy we got in our experiment. And these results in the purposed method are better than the previously published papers. By comparing all of these algorithm values by percentage and depicted in the table and also represented in the graphical format we got the highest value for random forest algorithm which is the best algorithm for our purpose work and will be the best fit for prediction purpose.

REFERENCES

[1] Kalra S, Prasad JS. Efficacy of News Sentiment for Stock Market Prediction. Proc Int Conf Mach Learn Big Data, Cloud Parallel Comput Trends, Prespectives Prospect Com 2019. Published online 2019:491-496. doi:10.1109/COMITCon.2019.8862265

[2] Menon A, Singh S, Parekh H. A review of stock market prediction using neural networks. 2019 IEEE Int Conf Syst Comput Autom Networking, ICSCAN 2019. Published online 2019:1-6. doi:10.1109/ICSCAN.2019.8878682

[3] Sharma A, Bhuriya D, Singh U. Survey of stock market prediction using machine learning approach. Proc Int Conf Electron Commun Aerosp Technol ICECA 2017. 2017;2017-Janua:506-509. doi:10.1109/ICECA.2017.8212715

[4] Chen MY, Liao CH, Hsieh RP. Modeling public mood and emotion: Stock market trend prediction with anticipatory computing approach. Comput Human Behav. 2019;101(September 2018):402-408. doi:10.1016/j.chb.2019.03.021

[5] Picasso A, Merello S, Ma Y, Oneto L, Cambria E. Technical analysis and sentiment embeddings for market trend prediction. Expert Syst Appl. 2019;135:60-70. doi:10.1016/j.eswa.2019.06.014

[6] Shobha G, Rangaswamy S. Machine Learning. Vol 38. 1st ed. Elsevier B.V.; 2018. doi:10.1016/bs.host.2018.07.004

[7] Basak S, Kar S, Saha S, Khaidem L, Dey SR. Predicting the direction of stock market prices using tree-based classifiers. North Am J Econ Financ. 2019;47(December 2017):552-567. doi:10.1016/j.najef.2018.06.013

[8] Kia AN, Haratizadeh S, Shouraki SB. A hybrid supervised semi-supervised graph-based model to predict one-day ahead movement of global stock markets and commodity prices. Expert Syst Appl. 2018;105:159-173. doi:10.1016/j.eswa.2018.03.037

[9] Henrique BM, Sobreiro VA, Kimura H. Literature review: Machine learning techniques applied to financial market prediction. Expert Syst Appl. 2019;124:226-251. doi:10.1016/j.eswa.2019.01.012

[10] Zhou F, Zhou H min, Yang Z, Yang L. EMD2FNN: A strategy combining empirical mode decomposition and factorization machine based neural network for stock market trend prediction. Expert Syst Appl. 2019;115:136-151. doi:10.1016/j.eswa.2018.07.065

[11] Sirimevan N, Mamalgaha IGUH, Jayasekara C, Mayuran YS, Jayawardena C. Stock Market Prediction Using Machine Learning Techniques. 2019 Int Conf Adv Comput ICAC 2019. 2019;(4):192-197. doi:10.1109/ICAC49085.2019.9103381

[12] Jadhav AA, Biradar N, Bhaldar H, Mathpati MS, Wadekar R, Scholar R. International Journal of Innovative Research in Computer and Communication Engineering Design and Analysis of Triple Band Miniaturized Antenna for Wearable Application. 2019;(March). doi:10.15680/IJIRCCE.2019

[13] Nayak A, Pai MMM, Pai RM. Prediction Models for Indian Stock Market. Procedia Comput Sci. 2016;89:441-449. doi:10.1016/j.procs.2016.06.096

[14] Oliveira N, Cortez P, Areal N. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Syst Appl. 2017;73:125-144.

doi:10.1016/j.eswa.2016.12.036

[15] Siew HL, Nordin MJ. Regression techniques for the prediction of stock price trend. ICSSBE 2012 - Proceedings, 2012 Int Conf Stat Sci Bus Eng "Empowering Decis Mak with Stat Sci. 2012;(December):99-103. doi:10.1109/ICSSBE.2012.6396535

[16] Das SR, Mishra D, Rout M. Stock market prediction using Firefly algorithm with evolutionary framework optimized feature reduction for OSELM method. Expert Syst with Appl X. 2019;4:100016. doi:10.1016/j.eswax.2019.100016

[17] Gandhmal DP, Kumar K. Systematic analysis and review of stock market prediction techniques. Comput Sci Rev. 2019;34:100190. doi:10.1016/j.cosrev.2019.08.001

[18] Yadav A, Jha CK, Sharan A. Optimizing LSTM for time series prediction in Indian stock market. Procedia Comput Sci. 2020;167(2019):2091-2100. doi:10.1016/j.procs.2020.03.257

[19] Bustos O, Pomares-Quimbaya A. Stock market movement forecast: A Systematic review. Expert Syst Appl. 2020;156. doi:10.1016/j.eswa.2020.113464

[20] Jahan I, Sajal SZ, Nygard KE. Prediction model using recurrent neural networks. IEEE Int Conf Electro Inf Technol. 2019;2019-May:390-395. doi:10.1109/EIT.2019.8834336

[21] Iyer M, Mehra R. A survey on stock market prediction. PDGC 2018 - 2018 5th Int Conf Parallel, Distrib Grid Comput. Published online 2018:663-668. doi:10.1109/PDGC.2018.8745715

[22] Hajek P. Forecasting stock market trend using prototype generation classifiers. WSEAS Trans Syst. 2012;11(12):671-680.

[23] Golmohammadi K, Zaiane OR, Diaz D. Detecting stock market manipulation using supervised learning algorithms. DSAA 2014 - Proc 2014 IEEE Int Conf Data Sci Adv Anal. Published online 2014:435-441. doi:10.1109/DSAA.2014.7058109