

Exploring Machine Learning Approaches for Analysis and Prediction in the Indian Stock Market

Arti Buche¹ and Dr. M. B. Chandak²

^{1,2}Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Email: artibuche@gmail.com, chandakmb@gmail.com

Abstract—Predicting the Indian stock market with accuracy is a considerable challenge, mainly due to its nonlinear time series nature. Nevertheless, the advent of diverse Machine Learning methodologies has sparked an extensive investigation in this domain. Primarily, scholarly inquiries have concentrated on harnessing Supervised Machine Learning algorithms for predicting the prices of individual stocks and forecasting movements in stock indices. Furthermore, researchers have explored the utilization of Unsupervised Machine Learning algorithms and statistical approaches to anticipate stock prices. Another area of exploration involves constructing portfolios that maximize profits while minimizing risks. Moreover, the influence of social media, which generates an abundance of data, has become a crucial factor impacting experts' decisions to buy or sell stocks. This survey aims to identify the various approaches researchers employ to improve the accuracy of machine learning algorithms in analyzing the Indian Stock Market. It highlights the importance of taking into account all pertinent aspects of the stock market prior to making predictions.

Index Terms— Machine learning, Stock Market Indicators, Sentiment Analysis, Hybrid approach, and Indian Stock Market.

I. INTRODUCTION

As of December 2019, the Indian stock market is recognized as the ninth largest globally, with a market capitalization of US \$2.02 trillion. Future projections suggest that the market capitalization is anticipated to surpass US \$6 trillion by 2027, offering substantial growth potential for investors. The primary motivations for investing in the stock market are to outpace inflation rates and to achieve quick financial gains. Through diligent research and informed decision-making, investments in stocks can maximize profits while minimizing risks.

The Indian stock market is predominantly comprised of two major stock exchanges: the Bombay Stock Exchange (BSE) and the National Stock Exchange (NSE). The BSE, established in 1875, holds the distinction of being Asia's oldest stock exchange. Presently, the Indian stock market encompasses approximately 5,400 listed companies, with around 500 of them accounting for nearly 90% of the market capitalization. Within the Bombay Stock Exchange (BSE), there is the SENSEX, which is a stock index composed of 30 large and highly traded companies.

On the other hand, the NSE was founded in 1992 and commenced trading in 1994. The National Stock Exchange (NSE) lists nearly 1,900 companies and has introduced the NIFTY, an equity benchmark index that represents the weighted average of 50 Indian company stocks spanning diverse sectors. The NIFTY serves as a significant indicator of market performance on the National Stock Exchange.

II. TOOLBOX OF MACHINE LEARNING METHODS

The field of Machine Learning has gained immense popularity due to its extensive applications in prediction, image recognition, speech recognition, medical diagnosis, and the financial industry, particularly in trading. The growth of Machine Learning can be attributed to factors such as enhanced computational power and the exponential generation of data by various digital devices. The primary categories of Machine Learning algorithms include Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

A. Supervised Learning

During this process, the system goes through a training phase where it builds a model based on a provided set of input data and its corresponding outputs. Following that, the model is evaluated using a separate set of inputs to validate the outputs. Eventually, the system gains the ability to generate outputs for any real-time input data.

B. Unsupervised Learning

In this scenario, the system is trained to construct a model solely based on the input data, without access to the corresponding output. The model then autonomously operates to uncover the correct output based on the given inputs.

C. Reinforcement Learning

In this approach, the system is trained using a reward-based framework to construct a model. The system learns by receiving rewards for accurate performances and penalties for incorrect performances, aiming to maximize rewards and minimize penalties. Through this iterative process, the system continuously improves its performance.

The following section provides a brief overview of the most extensively researched and commonly used Machine Learning algorithms.

K-Nearest Neighbours (KNN): KNN, known as K-Nearest Neighbors, is a simple supervised classification algorithm that can also be used for regression tasks. It works on the principle that similar data points are closer to each other. When provided with an input, KNN identifies the nearest neighbors using Euclidean distance and utilizes this information to classify the data. KNN finds applications in diverse domains, such as stock price prediction and portfolio construction.

Naïve Bayes Classifiers: Naïve Bayes classifiers are a set of probabilistic classification algorithms based on Bayes' Theorem. The Naïve Bayes classifier is recognized for its scalability and efficiency, as it requires minimal computational time and processor cycles. Its speed and effectiveness in classifying multi-class problems make it particularly valuable for real-time stock market prediction. Moreover, Naïve Bayes classifiers are widely used in text classification and sentiment analysis tasks.

Random Forest: Random Forest is a popular supervised Machine Learning algorithm primarily used for classification tasks, although it can also be applied to regression. The final decision is made based on the collective agreement among the decision trees. It has shown remarkable accuracy in handling large datasets, making it well-suited for analyzing stock market data.

Multiple Linear Regression (MLR): Multiple Linear Regression (MLR), also known as Multiple Regression (MR), is a statistical method used to determine the linear correlation between multiple independent variables and a dependent variable. MLR allows for the prediction of the dependent variable's value based on the provided values of the independent variables. As a result, MLR can be valuable in predicting stock market trends and stock prices.

Support Vector Machine (SVM): Support Vector Machine (SVM) is a supervised Machine Learning algorithm used to classify data points in an N-Dimensional space by utilizing a hyperplane. In addition to classification, SVM is also applicable in regression analysis, making it a versatile technique. With its capability to capture intricate non-linear relationships among data points, SVM is well-suited for predicting stock index movements and stock prices.

Neural Network (NN): These algorithms draw inspiration from the structure of neurons in the human brain. Artificial Neural Networks (ANN), or Neural Networks, consist of interconnected neurons organized in multiple layers. The first layer is the input layer, and the last layer is the output layer. Hidden layers can exist between the input and output layers. Each layer contains neurons that are connected to neurons in adjacent layers. In FFNN with backpropagation, an error signal is propagated backward from the output layer to the input layer if the

output pattern does not match the expected output. Neural Networks are highly effective for analyzing time series data, making them widely applied in stock market predictions.

K-Means Clustering: K-Means Clustering is a straightforward unsupervised Machine Learning method used to group or cluster data based on similar characteristics. By grouping unlabelled data with similar attributes, K-means clustering creates K clusters. This technique holds particular value in the creation and optimization of stock portfolios, as well as in performing sentiment analysis.

Linear Regression: Linear regression is a supervised learning algorithm used for predicting a continuous outcome based on one or more input variables. It establishes a linear relationship between the input variables and the target variable. The algorithm estimates the best-fitting line by minimizing the sum of squared differences between the predicted and actual values.

Logistic Regression: Logistic regression is another supervised learning algorithm used for binary classification tasks. It models the relationship between the input variables and the probability of a binary outcome. Logistic regression applies the logistic function to transform the linear regression output into a probability value, enabling classification.

III. RELATED WORK

This survey presents an overview of the research carried out on the Indian Stock Exchanges, specifically BSE and NSE. Numerous studies have utilized historical data sourced directly from BSE or NSE, or obtained from platforms like Yahoo Finance. Web scraping techniques are commonly used to extract stock data from websites such as Yahoo Finance. The survey further discusses common approaches employed for stock price prediction and predicting stock index trends.

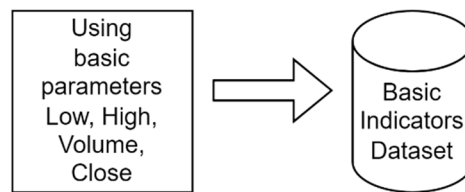


Figure 1. Basic Indicators Dataset: A Single-Layered Approach

In the initial stages, early studies on the stock market adopted a straightforward approach, as depicted in Figure 1. Supervised machine learning algorithms predominantly relied on fundamental parameters such as high price, low price, close price, and trading volume to indicate stock price trends or make predictions.

In [13], the conducted research adopts the aforementioned approach by employing the Neural Network (NN) algorithm and focusing exclusively on the Close price of the trading day. The study explores two variations of NN: Feed Forward NN with backpropagation and Radial Basis NN. Its objective is to predict both the stock price and its trend. The dataset utilized comprises the closing prices of nine stocks over four consecutive days from the NSE. The Feed Forward NN demonstrates better performance in predicting the stock price trend, while the Radial Basis NN yields superior results in predicting the actual stock price.

In [9], the approach illustrated in Figure 1 is once again utilized, employing the Neural Network algorithm with basic parameters. The research introduces a novel element by incorporating the Binarized Genetic Algorithm to optimize the weights during training, leading to the development of the BGANN (Binarized Genetic Artificial Neural Network) model. To enhance accuracy, Min-Max Normalization is performed during data pre-processing. The dataset utilized in this study consists of bank data collected from the NSEINDIA website, covering a three-year period from 2014. The results demonstrate that the BGANN model outperforms SVM, NN, and ARIMA in terms of error rate. The research endeavors of the study strive to integrate fundamental analysis in future investigations.

Technical indicators such as Simple Moving Average (SMA), Exponential Moving Average (EMA), Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), Commodity Channel Index (CCI), and more are derived from the basic indicators, as illustrated in Figure 2. In the case of SVM and NN, these indicators are utilized to enhance prediction accuracy.

In [19], the research focuses on Cuckoo Optimized SVM for Stock Market Prediction. The utilization of Cuckoo search as a parameter optimization technique for SVM aims to improve the accuracy of the SVM classifier. The CS-SVM (Cuckoo Search-SVM) model showcases enhanced prediction accuracy, surpassing both SVM and

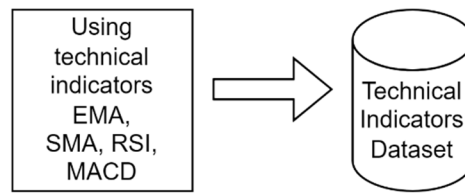


Figure 2. Technical Indicators Dataset: A Single-Layered Approach

ANN. The system utilizes basic indicators such as open, high, low, and volume, as depicted in Figure 1, along with five technical indicators including EMA, RSI, MACD, MFI, and SO shown in Figure 2. Daily closing prices are collected using Yahoo Finance as the data source. The research findings propose that integrating social media sentiments and taking into account the influence of crude oil and dollar price could significantly improve the accuracy of predictions.

In [1], the research integrates both basic stock parameters and a selection of technical parameters, following the approach depicted in Figure 1 and Figure 2. The CLFANN (Computationally Efficient Functional Link Artificial Neural Network) model leverages ADE (Adaptive Differential Evolution) and LMS (Least Mean Square) algorithms to train the network weights. MLP (Multi-Layer Perceptron) encounters several limitations, including slow convergence, local minimum, overfitting, and generalization issues. The weights optimized using ADE exhibit significantly improved performance compared to the enhanced LMS, effectively addressing the limitations of MLP. The dataset comprises 3400 records from Reliance (BSE India) and IBM, Oracle (DWSG-US). Error measures such as MAPE, RMSE, and MAE are employed for evaluation.

It has been observed that global political news plays a crucial role in stock market movements. For example, political tensions between America and Iran resulted in a decline of approximately 1.5% to 2% in stock markets, while indications of peace caused a rebound of around 1-1.5%. Currently, public sentiments exert a considerable influence on the stock market. Although Figure 3 attempts to illustrate this impact, considering other factors is necessary to enhance accuracy.

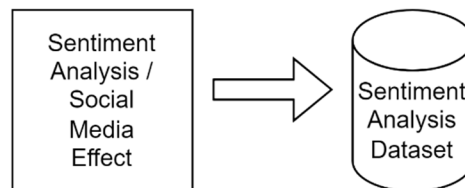


Figure 3. Sentiment Analysis Dataset: A Single-Layered Approach

The study [11] specifically targets pharmaceutical stocks and constructs a dictionary to build a sentiment analysis model. By solely considering news sentiments and excluding other parameters, the model achieves a directional accuracy of 70.59% in predicting stock price trends. To gather news articles, the researchers utilize the reliable moneycontrol.com, while the 'pattern' Python library is employed to convert text into numerical vectors. The study proposes that a word-weighting approach could be advantageous in sentiment analysis. Furthermore, the incorporation of approaches like ARIMA or LSTM alongside sentiment analysis may further enhance the directional accuracy.

The hybrid approach illustrated in Figure 4 integrates both machine learning (ML) algorithms and sentiment analysis, leading to enhanced results by taking into account multiple factors that influence the stock market.

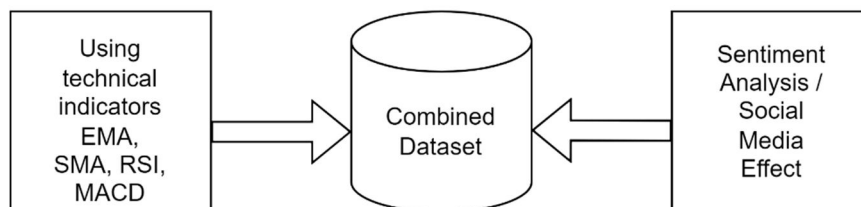


Figure 4. ML-Based Sentiment Analysis in Hybrid Approach

The study documented in [16] follows the approach depicted in Figure 4. Initially, the performance of three machine learning (ML) algorithms - artificial neural network (ANN), support vector machine (SVM), and

multiple linear regression (MLR) - is evaluated, with ANN outperforming SVM and MLR. Subsequently, ANN is combined with sentiment analysis to develop a prediction model for generating stock prices. The prediction model focuses on both monthly and daily predictions, specifically forecasting the opening price for the following day.

The model presented in [6] combines neural networks (NN) with sentiment analysis, following a similar approach as illustrated in Figure 4. In this model, the input layer consists of 5 neurons, with 4 neurons representing the closing prices of the previous 4 days and the fifth neuron representing the sentiment score of the previous day. The study suggests the use of deep multilayer neural networks with more than two hidden layers to determine text sentiments effectively.

In [4], separate models are developed: one for daily trend prediction and another for monthly prediction. The daily prediction model combines historical prices with sentiment analysis, following a similar approach as depicted in Figure 4. The historical data is trained using supervised machine learning algorithms, with Decision Boosted Tree outperforming SVM and Logistic Regression. The daily prediction model achieves an accuracy rate of approximately 70%. On the other hand, the monthly prediction model indicates a weak correlation between the trends of the previous month and the current month. For sentiment analysis, data is collected from Twitter and two distinct news websites, while stock data is sourced from Yahoo Finance.

The study [20] specifically focuses on the SVM classification method, following the approach outlined in Figure 4, where technical indicators are employed to predict the CNX NIFTY index value. This research investigates four different kernels of the SVM classification model. The dataset consists of only 1046 samples from CNX NIFTY, with basic parameters (open, high, low, and close) obtained from Yahoo Finance. From these basic indicators, 20 additional technical features are calculated and considered.

A different research study [15] utilizes the method illustrated in figure 4 to predict stock prices. The study incorporates a hybrid approach that combines Neural Network and Genetic Algorithm. A soft computing technique is implemented to analyze news related to the stock market and social media data. The method is tested exclusively on the CUB stock acquired from the NSE website.

A new methodology named the Multi-Layer approach, which combines multiple ML algorithms, is depicted in Figure 5. The research study [8] concentrates on the S&P BSE and CNX Nifty indices in India. It introduces a two-stage fusion approach, similar to the one illustrated in Figure 5, involving SVR (Support Vector Regression) in the initial stage. In the second stage, ANN (Artificial Neural Network), RF (Random Forest), and SVR are employed, leading to the creation of the SVR-ANN-RF and SVR-SVR fusion models for predicting stock market indices. Ten technical indicators are chosen as inputs for these models.

The methodology portrayed in Figure 5 is also implemented in the research documented in [10]. The study concentrates on developing and evaluating three hybrid models: ARIMA-SVM, ARIMA-ANN, and ARIMA-RF. These models combine both linear and non-linear techniques to capture the relationship in time series data. The outcomes indicate that the ARIMA-SVM hybrid model surpasses other models in terms of profit, volatility, and risk-adjusted performance. The study concludes that the proposed hybrid model performs well and outperforms individual models.

The research documented in [3] proposes a hybrid framework that resembles the multi-layer approach depicted in Figure 5. The suggested study incorporates SVM as the first layer and LSTM as the second layer. Comparative analysis of the hybrid model with SVM, LSTM, and Back Propagation indicates that the proposed model outperforms other algorithms in terms of both accuracy and time efficiency. The dataset utilized in the study comprises TATA Global stock data from NSE covering the period from 2012 to 2018. The model's predictions were made for both one day ahead and one week ahead.

The research documented in [5] not only focuses on predicting stock prices one day in advance but also provides recommendations on whether to trade or not for a given stock. The study adopts a multi-layered approach, as illustrated in Figure 5. Initially, the technical indicators are optimized using a GA (Genetic Algorithm) technique and then employed in a decision tree. Furthermore, two additional systems, one utilizing Decision Tree - SVM and the other employing Decision Tree - SVM with feature selection on the optimized technical parameters, were examined.

In the paper [14], a hybrid framework is proposed that merges Support Vector Machine (SVM) with K-Nearest Neighbor (KNN). The first step entails utilizing SVM with different kernel functions to forecast profit or loss. The output of SVM is subsequently utilized as input for K-NN, which identifies the nearest neighbor from the training set to predict the stock index for various time horizons, including 1 month, 1 week, and 1 day. The outcomes obtained from the framework also offer insights for analyzing the closing price, volatility, and momentum of the stock market.

The study [17] explores the development of a stock market portfolio and the formulation of an appropriate stock selection strategy. The primary objective of any portfolio is to maximize profit while minimizing risk. In this research, the initial step involves applying k-means clustering to a diverse sample of 53 stocks listed on the NSE. The monthly closing return serves as the basis for classifying the stocks into three categories: low stock price, medium stock price, and high stock price. This study bridges the effects of unsupervised learning with supervised learning, making it distinct in its approach.

In [18], data mining techniques are utilized to categorize stocks into distinct clusters, optimizing the stock portfolio. The findings indicate that K-Means clustering produces the most condensed clusters in comparison to SOM and Fuzzy C-means. The study employs timely stock returns and valuation ratios to evaluate the performance of the stocks.

Another hybrid model proposed in [2] aims to forecast stock returns from the NSE. The model combines a non-linear model, RNN, with two linear models, Exponential Smoothing and Autoregressive Moving Average, to create a unique Hybrid Prediction Method (HPM). The study integrates predictions from RNN, ES, and ARMA to establish the HPM. The dataset utilized in the research includes 6 stocks from diverse sectors during a 16-week period, with an equal split between training and test datasets, each accounting for 50%. It is important to note that RNN is exclusively used for training, while the HPM is employed for testing. The performance of the predictive models is evaluated using MSE and MAE, revealing a significantly positive correlation coefficient between the actual and predicted returns.

A distinct and unique study is presented in [12], focusing on predicting the increase or decrease in stock market performance. While conventional stock prediction models rely on basic or technical using machine learning, this study takes a different approach. It utilizes common macroeconomic factors like commodity prices, market history, and foreign exchange rates to forecast BSE. In the analysis, four machine learning algorithms, namely SVM, RF, Gradient Boosting, and Adaptive Boosting, are considered. The study incorporates market history data spanning 2 years from June 2015 to December 2016, while commodity and foreign exchange rates data from investing.com for 9 months are used. Among the evaluated models, the AdaBoost algorithm demonstrates the highest accuracy.

The research presented in [7] distinguishes itself by considering the relationship between data dimension and model complexity. It recognizes that higher-dimensional data can lead to increased model complexity. To address this, the study focuses on implementing the minimal variability OWA (Order Weighted Averaging) operator. By combining subtractive clustering with a hybrid fuzzy-based network inference system based on the proposed model, the research achieves superior results compared to other models. The experiment conducted centers around the Bombay Stock Exchange Index (BSE30).

IV. DISCUSSION

The survey conducted in this study highlights the wide array of research conducted on the Indian stock market. It identifies multiple prominent approaches utilized for predicting stock price trends and stock index trends. The Indian stock market data serves as the foundation for implementing nearly all machine learning algorithms.

Certain research studies concentrate on a limited number of stocks, which raises concerns about their performance when applied to a different stock dataset. Given the highly unpredictable nature of the stock market, the performance of a specific research approach can vary over time. However, it is commonly observed that finely-tuned Support Vector Machine (SVM) and Artificial Neural Network (ANN) models often produce better results compared to other algorithms. In order to improve accuracy, many recent research efforts employ a hybrid or multi-layered approach, resembling the ones illustrated in Figure 4 or Figure 5.

V. CONCLUSION

Numerous research studies concentrate on a single dimension, utilizing only one key parameter in the prediction process. However, despite the existence of various hybrid or multilayered approaches, they often fall short in encompassing all the factors that influence stock prices or stock price indices. The majority of the studies neglect fundamental parameters and macroeconomic variables, which have a significant impact on stocks. Moreover, the current impact of social media data on the stock market cannot be ignored. It is evident that there is a gap in research that aims to encompass all pertinent parameters. Furthermore, when considering all parameters, it is necessary to determine whether a multilayered or hybrid approach is more effective. Assigning weights to the parameters based on historical evidence and regularly evaluating their performance may be beneficial.

REFERENCES

- [1] Rout A K, Biswal B, and Dash P K, "A hybrid FLANN and adaptive differential evolution model for forecasting of stock market indices", *International Journal of Knowledge-Based and Intelligent Engineering Systems*, 18(1) 23–41, 2014.
- [2] Rather A M, Agarwal A, and Sastry V N, "Recurrent neural network and a hybrid model for prediction of stock returns", *Expert Systems with Applications*, 42(6) 3234–3241, 2015.
- [3] Mundra A, Mundra S, Verma V K, and Srivastava J S, "A deep learning based hybrid framework for stock price prediction", *Journal of Intelligent and Fuzzy Systems*, 38(5) 5949–5956, 2020.
- [4] Nayak A, Pai M M M, and Pai R M, "Prediction Models for Indian Stock Market", *Procedia Computer Science*, 89 441–449, 2016.
- [5] Nair B B, and Mohandas V P, "An intelligent recommender system for stock trading", *Intelligent Decision Technologies*, 9(3) 243–269, 2015.
- [6] Vanipriya C H, and Thammi Reddy K, "Indian Stock Market Predictor System", *Advances in Intelligent Systems and Computing*, 249 17–26, 2014.
- [7] Kaur G, Dhar J, and Guha R K, "Stock market forecasting using ANFIS with OWA operator", *International Journal Artificial. Intelligence*, 12(2) 102–114, 2014.
- [8] Patel J, Shah S, Thakkar P, and Kotecha K, "Predicting stock market index using fusion of machine learning techniques", *Expert Systems with Applications*, 42(4) 2162–2172, 2015.
- [9] Srinivas K, Srinivasa Rao V, and Sreemalli M, "Binarized genetic algorithm with neural network for stock market prediction", *ACM International Conference Proceeding Series*, 8–11, 2018.
- [10] Kumar M, and Thenmozhi M, "Forecasting stock index returns using ARIMA-SVM, ARIMA-ANN and ARIMA-random forest hybrid models", *International Journal of Banking, Accounting and Finance*, 5(3) 284–308, 2014.
- [11] Arti Buche and Dr. M. B. Chandak, "Stock Market Forecasting Techniques: A survey", *Journal of Engineering and Applied Sciences*, Volume 14, Issue 5, Pages 1649–1, 2019.
- [12] Kohli P P S, Zargar S, Arora S, and Gupta P, "Stock prediction using machine learning algorithms", *Advances in Intelligent Systems and Computing*, 698 405–414, 2019.
- [13] Charkha P R, "Stock price prediction and trend prediction using neural networks", *Proceedings - 1st International Conference on Emerging Trends in Engineering and Technology ICETET*, 592–594, 2008.
- [14] Nayak R K, Mishra D and Rath A K, "A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices", *Applied Soft Computing Journal*, 35 670–680, 2015.
- [15] Suganya R, and Sasi Kumar A, "Stock price prediction using tech news based soft computing approach", *International Journal of Advanced Trends in Computer Science and Engineering*, 9(2) 2049–2054, 2020.
- [16] Nivetha R Y, and Dhaya C, "Developing a Prediction Model for Stock Analysis", *Proceedings - 2017 International Conference on Technical Advancements in Computers and Communication ICTACC*, 2017.
- [17] Gupta S, Bandyopadhyay G, Biswas S and Upadhyay A, "A Hybrid Machine Learning and Dynamic Nonlinear Framework for Determination of Optimum Portfolio Structure", *In Lecture Notes in Networks and Systems*, 74 437–448, 2019.
- [18] Nanda S R, Mahanty B, and Tiwari M K, "Clustering Indian stock market data for portfolio management", *Expert Systems with Applications*, 37(12) 8793–8798, 2010.
- [19] K. N. Devi, V. M. Bhaskaran and G. P. Kumar, "Cuckoo optimized SVM for stock market prediction", *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, India, pp. 1–5, 2015.
- [20] Upadhyay V P, Panwar S, Panchariya R and Merugu R, "Forecasting stock market movements using various kernel functions in support vector machine", *ACM International Conference Proceeding Series*, 12–13, 2016.