Retraction Notice

The Editor-in-Chief and the publisher have retracted this article. An investigation uncovered evidence of compromised peer review and determined the paper to be out of the scope of the journal. The Editor and publisher no longer have confidence in the results and conclusions of the article.

SK, AJ, AN, GSY, and RK disagree with the retraction. PG and SR either did not respond or could not be reached.

Machine learning-based probabilistic profitable model in algorithmic trading

Shubham Khandelwal,[a] **Piyush Gupta**,[a] **Aman Jain**,[a] **Ajay Nehra**,[a,*] **Gyan Singh Yadav**,[a] **Riti Kushwaha**,[b] **and Selvanambi Ramani**[c]

[a]Indian Institute of Information Technology, Kota, India
[b]Bennett University, Greater Noida, India
[c]Vellore Institute of Technology, Vellore, India

Abstract. Machine learning models are nowadays becoming ubiquitous in algorithmic trading and investment management. These models are mostly used in the pre-trade analysis phase to determine the buy or sell decisions using various machine learning techniques. We aim to implement a machine learning-driven approach using various technical indicators to predict stock market prices and then accordingly make a decision about buying or selling. First, an effective trading strategy is discussed that selects the potentially profitable stocks, and then the technical indicators such as simple moving average (SMA), exponential moving average (EMA), relative strength index (RSI), and moving average convergence divergence (MACD) are calculated for those potentially profitable stocks. Then supervised machine learning algorithms such as multiple linear regression, support vector machine regression, and decision tree regression are applied, where the close price of the stock is predicted using technical indicators for the next day, and based on that buy or sell signals are generated. The model is then tested on 12 different SNP500 stocks, one for every month in 2018, with the mean squared error (MSE) varying between 30.33 and 48.16 and the root MSE varying between 5.51 and 6.93, where the error is calculated on the difference in the number of days when the stock price actually increases and the predicted number of days for various models. © *2023 SPIE and IS&T* [DOI: 10.1117/1.JEI.32.1.013039]

Keywords: algorithmic trading; machine learning; technical analysis; trading strategy.

Paper 221294G received Nov. 13, 2022; accepted for publication Jan. 25, 2023; published online Feb. 21, 2023.

1 Introduction

In this era of automation and technology, financial markets are also being automated. Algorithmic trading[1] refers to the use of computer programs or algorithms to automate one or more stages of the trading process: data analysis, trading signal generation (i.e., buy and sell decisions), and trade execution. The speed and accuracy of algorithmic trading are very high and it is impossible for human traders to work at such high speed or accuracy, that is why nowadays algorithmic trading is preferred over manual trading. Since the prediction of the share market is very vague as there are no verified rules or methods to guarantee, estimate, or precisely predict the price of a share in the share market, many methods such as technical analysis,[2] time series analysis, statistical analysis, and fundamental analysis[3] are used to attempt to predict the price in the share market and make buy or sell decisions. The involvement of computers in any of these steps will improve the speed and accuracy, and remove human emotions while decision making.

This paper is a blend of the knowledge of trading and machine learning to describe how machine learning can be used to predict whether the algorithmic trading strategy is profitable or not. Earlier, trading in stock markets used to be manual, wherein human traders used to do the analysis and then they placed the trades. There were many problems in manual trading; for instance, the speed of humans is not comparable with computers, so manual trading is slow. Moreover, human sentiments are involved in manual trading, which can sometimes lead to

*Address all correspondence to Ajay Nehra, ajay.cse@iiitkota.ac.in

wrong decisions. Hence, the concept of algorithmic trading emerged. Algorithmic trading uses computers and algorithms to do the analysis and place the trades.[4]

In this paper, a trading strategy is discussed in Sec. 4, which decides whether a stock is potentially profitable or not. It has three main conditions, i.e., inside days, uptrend check, and volatility contraction. The uptrend condition identifies the stocks with upward price momentum. Inside days and volatility contraction work on the condition that if the demand currently is low, then in the future the demand will rise, and so shall the price.

Then the technical analysis is performed on potentially profitable stocks found from the strategy. Various technical indicators,[5] which include simple moving average (SMA), exponential moving average (EMA), relative strength index (RSI), average directional index (ADX), moving average convergence divergence (MACD), etc.,[6] are then calculated. SMA and EMA are trend indicators, which will provide information about the uptrend or downtrend. RSI and ADX are momentum indicators, which capture information about the momentum of price-increase or price-decrease. MACD is a trend-following momentum indicator, which captures both pieces of information. These indicators are preferred over others because these are the most widely used; in addition, they provide most of the information needed related to trend and momentum information.

We have used technical indicators in conjunction with machine learning algorithms because there are some disadvantages to using technical indicators alone. Sometimes, they may give signals that can be opposite to the market. Also, a stock price could have made a substantial move already by the time the trend is identified using technical analysis.

So these technical indicators are used as input parameters along with volume data in supervised machine learning algorithms such as multiple linear regression,[7] decision tree regression,[8] and support vector machine (SVM) regression.[9] Since the data from the stock market is readily available,[10] supervised machine learning algorithms are used here; also, since there is no rule in the stock market for the growth of stocks, complex models will not generalize better. Hence these algorithms are selected. As discussed, these three machine learning algorithms predict whether the price of the potential stocks will go up or down; if up, then the "buy" signal is generated, otherwise a "do not buy" signal is generated.

After an extensive survey of the related documents, we identified a gap in that there is a lack of the fusion of technical analysis and machine learning-based approaches to predict the profitability of algorithmic trading strategies. Our approach is targeting this gap with the development of a trading strategy to find potentially profitable stocks and then use technical analysis and machine learning to predict their profitability.

The research contributions that we have made through this paper are:

1. Development of an algorithmic trading strategy, which gives potentially profitable stocks based on three major conditions. [Sec. 4]
2. Designing a system that utilizes the information of potentially profitable stocks to apply technical analysis and then combines it with machine learning algorithms to predict profitability. [Sec. 5]

The remainder of this article is arranged as follows: beginning with this introduction and in the next section, the background knowledge is shared. In the next section, the literature survey is carried out on the previous studies. Section 4 describes an effective trading strategy to find potential stocks to buy on a particular day. Section 5 describes the system design. Then the results are discussed. The conclusions are then discussed in the last section.

2 Background Knowledge

There are two major components of the machine learning-based system, the first is the technical analysis related part and the second is the part that makes the prediction. In the system first, the input features are to be calculated. For this system, these input features are the technical indicators. Then, these are used to predict the profitability using supervised machine learning algorithms. In this section, the background information, which is required to understand the paper is discussed. First, there is a discussion of various technical indicators, which include SMA, EMA,

RSI, ADX, MACD, etc. Then there is a discussion about supervised machine learning algorithms used for the prediction.

2.1 *Technical Indicators to Be Used as Input Features for Machine Learning Algorithms*

Technical indicators are heuristic or pattern-based signals produced by price and volume movements. These indicators capture the trend, volume, and momentum information. These indicators can be used to predict future price movements using them as input features in machine learning algorithms.

Candlestick charts. Figure 1 shows a popular way to represent stock prices, which is easy to interpret and very useful. The green candle represents the prices for a profit-making day, while the red candle indicates that it is a loss-making day. The two tails or shadows on the upper and lower sides represent the high and low prices on that particular day. On the green candle, the closing price is higher than the opening price, and the opposite is true for the red candle.

SMA. A simple moving average or SMA is a mathematical calculation that takes the arithmetic mean of a given set of prices over a specific time period in the past. An SMA for a smaller number of days is more sensitive to price changes as compared to an SMA for a greater number of days.[11]

EMA. An EMA is a type of moving average that weights, i.e, places a higher importance and significance on, the most recent day's data. The other name for this is the exponentially weighted MA. It is more sensitive to price changes than SMA.[11]

MACD. Moving average convergence divergence, or MACD, is a momentum indicator that follows a trend and it shows the relationship between two moving averages of a stock's price. An EMA of nine days of the MACD is called the signal line, which is plotted on top of the MACD line, using this we can generate buy and sell signals.[12]

RSI. The relative strength index (RSI) is also a momentum indicator. It is used in technical analysis to analyze the magnitude of price changes that are recent. It tells whether the stock is overbought or oversold.[13]

ADX. The average directional index or ADX is a technical indicator that is used to decide the strength of a trend, i.e., how strong is the upward or downward price momentum.[14]

Of these chosen indicators, SMA and EMA are trend indicators, which will provide information about the uptrend or downtrend. RSI and ADX are momentum indicators, which capture information about the momentum of price-increase or price-decrease. MACD is a trend-following momentum indicator, which captures both pieces of information. Some other indicators are parabolic stop and reverse, which is a trend indicator, and stochastic oscillator, which is a momentum indicator. Since the information about trend and momentum is already captured by other indicators, these indicators are not used.

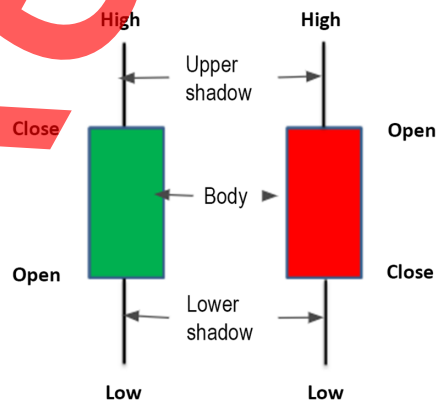

Fig. 1 Candlestick chart.

2.2 *Supervised Machine Learning Algorithms to Predict the Profitability of a Strategy*

In supervised machine learning, the machine is trained using perfectly labeled data, and based on that the machine learns and predicts the output. Some of the supervised algorithms are discussed here; it is also discussed how they can be used for finding the profitability of the stocks.

Since the data from the stock market is readily available,[10] hence supervised machine learning algorithms are used here. Some of the most popular supervised machine learning algorithms are linear regression, logistic regression, naïve Bayes network, SVM, K-means clustering, decision trees, etc.[15]

Since there is no rule in the stock market for the growth of stocks, so complex models will not generalize better, hence multiple linear regression is selected. SVM is robust to high dimensional data and has good generalization ability, hence SVM regression is used, and the decision tree is non-parametric, hence outliers do not affect the model much, so it is also selected.

2.2.1 *Multiple linear regression*

Multiple linear regression, commonly known as multiple regression, is a statistical technique that uses several independent variables to predict the outcome of another variable, which is generally known as the dependent variable. The formula for multiple regression is

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon, \tag{1}$$

where for $i = n$ observations: $y =$ dependent variables, $x_j =$ independent variables, $\beta_0 = y$ -intercept (constant terms), $\beta_p =$ coefficients of the slope for each independent variable, and $\epsilon =$ the error term of the model (which is also known as the residuals)

In multiple regression, we try to fit a plane to the data points using the independent variables, which are more than one variable. Then we use that plane to predict the value of the dependent variable. In this scenario, we use the closing price of the stocks as a dependent variable. For independent variables, we use opening price, volume, SMA10, EMA10, RSI14, and ADX.

2.2.2 *Decision tree regression*

Decision tree builds regression models or classification models in the form of a tree-like structure. It breaks down a dataset into smaller subsets while at the same time incrementally developing an associated decision tree. Finally, we get a decision tree with decision nodes and leaf nodes as a result as depicted in Fig. 2.

To use such a tree for regression to predict the value of the dependent variable, the data point is classified as to the correct node using its independent variables. The value of the dependent variable will be the mean of the dependent variable values of all the other data points present in that node.

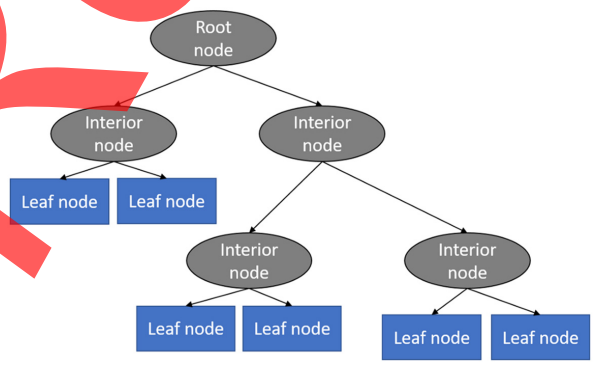

Fig. 2 Decision tree.

In this scenario, we use the closing price of the stocks (Close) as a dependent variable. For independent variables, we use the opening price (Open Price), volume, SMA10, EMA10, RSI14, and ADX.

2.2.3 *Support vector machine regression*

SVMs are supervised, learning models. They have associated learning algorithms that help in analyzing the data used for classification and regression. In SVM regression, the straight line which has to be fit is known as the hyperplane.

The main goal of an SVM algorithm is to find a hyperplane, which is in an n-dimensional space that evidently classifies the given data points. The data points on every side of the hyperplane that are the closest to the hyperplane are known as the support vectors. These support vectors tend to influence the position and direction of the hyperplane and hence they help in building the SVM.

The principle of support vector regression (SVR) is the same principle as the SVMs. The basic idea behind SVR is to find the line of best fit. The hyperplane that has the maximum number of points is the best fit line in SVR. Here also we use the same set of independent variables and dependent variables.

3 Literature Survey

The use of machine learning and technical analysis for stock market prediction has been the subject of much research in recent years. Many different approaches have been proposed, with varying levels of success. In Sec. 3.1, the literature survey of different papers using approaches based on machine learning algorithms is discussed such as SVM, random forest, and logistic regression. In Sec. 3.2, the literature survey of different papers using deep-learning algorithms such as artificial neural network (ANN), convolutional neural network (CNN), and long short-term memory (LSTM) is discussed. In Sec. 3.3, the literature survey of different papers using hybrid approaches based on technical analysis is discussed.

3.1 *Machine Learning Based*

Cheng et al.[16] tried to investigate the fact that whether the ensemble methods can be employed to improve the performance of the base learner in financial time series prediction. They found that ensemble algorithms are powerful in improving the performances of base learners in financial time series prediction. When compared with random subspace and stacking, bagging provides a more stable and better improvement.

Yuan and Luo[17] used high-frequency data of stock prices to construct different input features, and then used a decision tree model to forecast the price movement. They reached an accuracy of up to 70%. The variables that they used in their model were not refined and could have been constructed in more depth.

Patel et al.[18] compared and contrasted several different machine learning algorithms, including SVMs, decision trees, and ANNs. They find that a combination of these techniques (fusion) outperforms any single technique alone. The paper includes a detailed analysis of how each machine learning algorithm performed on the data. They did not take into account the news related to the stock market.

Cakra and Distiawan Trisedya[19] used financial news to enhance the feature representation which was based on a linear regression model for stock market prediction. They are using naïve Bayes and random forest algorithms are used to calculate sentiment regarding a company. The results show that created sentiment analysis model using the random forest algorithm can classify tweet data with 60.39% accuracy, and the one with the naïve Bayes algorithm can classify tweet data with 56.50% accuracy. Their model could be improved using other natural language processing methods such as part of speech tagging and word weighing.

Khan et al.[20] used *K*-nearest neighbor (KNN), naïve Bayes, and SVMs before and after applying principal component analysis (PCA). They found that the KNN has the highest

accuracy. The model could further be improvised using social media data in addition to the price data to predict.

Milosevic[21] used decision trees, SVMs with sequential minimal optimization, JRip, random trees, random forest, logistic regression, naïve Bayes, and Bayesian networks and compared the results to predict long-term stock price movement. They were able to correctly predict whether some company's value will be 10% higher or not over the period of one year in 76.5% of cases. The limitation of this study is that the models are not created out of data that were not limited in time.

Rossi[22] used a semi-parametric method known as boosted regression trees (BRT) to forecast stock returns and volatility at the monthly frequency. The analysis shows that the relation between the predictor variables constituting the conditioning information set and the investors' optimal portfolio allocation to risky assets is, in most cases, nonlinear and nonmonotonic.

Vanukuru,[23] Huang and Tsai,[24] Zhang et al.,[25] and Pai et al.[26] used different variants of SVR for prediction. The SVR parameters that they used could be optimized using more advanced algorithms such as genetic algorithms.

Gururaj and Shriyav[7] used linear regression and SVMs to predict the stock market and try to find out the pros and cons of using both these techniques to predict values and compare both algorithms. They found that the SVM is better than linear regression in terms of root mean square error, mean absolute error, mean squared error (MSE) and R-squared error.

Illa et al.[27] used random forest and SVM. They used these procedures to determine whether the cost of a stock will be higher than its cost on a given day to make profitable trading strategies. They found that the random forest model outperforms the SVM. In Table 1, all discussed approaches are summarized.

3.2 *Deep-Learning Based*

Yoo et al.[28] tried to investigate various global events and their issues in predicting stock markets. For this, they compared various techniques such as SVMs, case-based reasoning, and ANN to predict the stock market and then compared the results. They found that neural networks offer the ability to predict market directions more accurately than other existing techniques. Their model did not take into account the historical events in the stock market.

Zhong and Enke[29] used kernel-based principal component analysis and ANNs to predict the daily stock market return. For this, they used the dimensionality reduction algorithms such as PCA and fast robust PCA. They found that combining the ANNs with the PCA gives slightly higher classification accuracy as compared to the other models they used. The selection of the kernel function for the KPCA algorithm could be improved. Wu et al.[30] used a graph-based CNN-LSTM model in conjunction with leading technical indicators used a variety of news collections, including options, historical data, and futures, and involved the stock sequence array to predict the prices of stocks. They found that the neural network framework combined with convolution and long–short-term memory units achieved better performance for statistical methods and traditional CNN and LSTM in prediction tasks. In Table 2, summary of various deep-learning-based approaches is presented.

3.3 *Hybrid Approaches*

Zhu and Zhou[2] analyzed the usefulness of technical indicators, more specifically the moving average rules from an asset allocation perspective. For this, they used an optimal generalized moving average (GMA). They found that the optimal GMA is robust to model specification and outperforms the optimal dynamic strategies substantially derived from the wrong models. Their model was not able to find how past prices and trading volumes reveal the strategies of the major market players.

Peachavanish[5] proposed a method using cluster analysis to identify a group of stocks that has the best trend and momentum characteristics at a given time. They found that the best combination of these indicators is determined by way of cluster analysis and results show that the proposed method to select stocks from the Thai stock market and trade them using equal-weight monthly portfolio rebalancing can outperform the market in the long run.

Table 1 Literature survey of machine learning based approaches.

Research paper title	Technology used	Results
A comparison of ensemble methods in financial market prediction[16]	Random subspace, stacking, and bagging	Bagging provided more stable and better results
Test on the validity of futures market's high-frequency volume and price on forecast[17]	Decision tree	Reached an accuracy of up to 70%
Predicting stock market index using a fusion of machine learning techniques[18]	SVM, decision tree, and ANN	The MAE for the prediction performance of *n*-day ahead of time for CNX Nifty where *n* varies from 1 to 30 lies between 52.48% and 278.37%
Stock price prediction using linear regression based on sentiment analysis[19]	Naïve Bayes and random forest	With random forest 60.39% accuracy, and with the naïve Bayes 56.50% accuracy
Predicting trend in stock market exchange using machine learning classifiers[20]	KNN, naïve Bayes, SVM, and PCA	The KNN has the highest accuracy
Equity forecast: Predicting long-term stock price movement using machine learning[21]	Decision tree, SVM, random forest, and logistic regression	Correctly predict whether some company's value will be 10% higher or not over the period of one year in 76.5% of cases.
Predicting stock market returns with machine learning[22]	BRT	Relation between the predictor variables and the optimal portfolio allocation to risky assets is nonlinear and nonmonotonic.
Stock market prediction using machine learning[23]	SVR	The SVM algorithm was effective for predicting stock index movements as it eliminated the problem of over-fitting.
A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting[24]	SVR	The average MAPE error lies between 5.5% and 6.2%.
Support vector regression with modified firefly algorithm for stock price forecasting[25]	SVR	The MAE for different types of SVR model lies between 3.2% and 6.6%.
Stock market prediction using linear regression and support vector machines[7]	Linear regression and SVM	The SVM is better than linear regression
Stock price prediction methodology using random forest algorithm and support vector machine[27]	Random forest and SVM	The random forest model outperforms the SVM.

Table 2 Literature survey of deep-learning-based approaches.

Research paper	Technology used	Results
Machine learning techniques and use of event information for stock market prediction: a survey and evaluation[28]	ANN	NNs offer the ability to predict market directions more accurately than other existing techniques.
Forecasting daily stock market return using dimensionality reduction[29]	Kernel-based PCA and ANN	Combining the ANNs with the PCA gives slightly higher classification accuracy as compared to the other models they used.
A graph-based CNN-LSTM stock price prediction algorithm with leading indicators[30]	Graph-based CNN-LSTM model	The neural network combined with convolution and LSTM units achieved better performance for statistical methods than traditional CNN and LSTM.

Table 3 Literature survey of hybrid approaches.

Research paper	Technology used	Results
Technical analysis: an asset allocation perspective on the use of moving averages[2]	Optimal generalized moving average	The optimal GMA is robust to model specification.
Stock selection and trading based on cluster analysis of trend and momentum indicators[5]	Cluster analysis with trend and momentum indicators	Using cluster analysis they were able to outrun the market in the long run
A novel hybrid model for stock price forecasting based on metaheuristics and support vector machine[32]	Metaheuristics with SVM	Able to outperform the other methods in terms of quality and accuracy in the US stock market.

Sedighi et al.[32] used a novel hybrid model for stock price forecasting based on metaheuristics and a SVM. The outcomes, were obtained by running on datasets taken from 50 largest companies of the U.S. Stock Exchange from 2008 to 2018 and their approach outperforms the other methods in terms of quality and accuracy. Table 3 summarizes all discussed hybrid approaches.

Overall, the literature suggests that the use of machine learning and technical analysis for stock market prediction can be an effective approach, with many studies showing that these methods are able to outperform traditional methods of stock market prediction.

4 Effective Strategy to Find Potential Stocks to Buy on a Particular Day

There are three main conditions to mark a stock as a potential stock, which can be bought as its price will likely go up. These are uptrend, inside days, and volatility contraction. An uptrend is a sustained increase in the price of a security or market index over a period of time. It is generally considered to be a bullish sign, as it indicates that demand for that stock is increasing and that investors are willing to pay higher prices. Inside days refer to days in which the range of the trading day's high and low prices is contained within the previous day's high and low prices. An inside day can be a sign of indecision or a pause in the current trend, and it can sometimes precede a change in the direction of the trend. Volatility contraction refers to a decrease in the level of price volatility in a security or market index. It is often seen as a sign of a potential trend reversal, as it can indicate that the market is consolidating before making a move in a particular direction. These three conditions can help to find potentially profitable stocks.

4.1 *Check for Uptrend*

We keep a track of two different MAs of a particular stock. One is a slow-MA which we find using an effective MA finder algorithm. The other MA is a fast-MA which is half the duration of the slow-MA. Then we check if the current fast-MA is higher than the current slow-MA and if the price of the stock is higher than the fast-MA, if so then we say that the stock is in the uptrend. The graph in Fig. 3 shows the uptrend condition for Merck & Co. Inc. (MRK) stock.

4.2 *Check for Inside Days*

For a stock that is already in an uptrend, we check whether this stock is in an inside day situation for the particular day on which we are checking it. An inside day means that the high and low prices of the stock on that particular day are completely inside the high and low prices of the previous day. In other words, the candlestick in the chart for that day should be completely inside the previous day's candle. The graph in Fig. 4 shows the inside days condition for MRK stock.

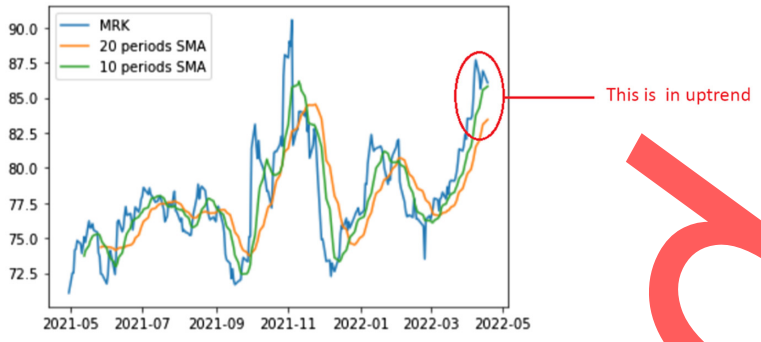

Fig. 3 Diagram for uptrend.

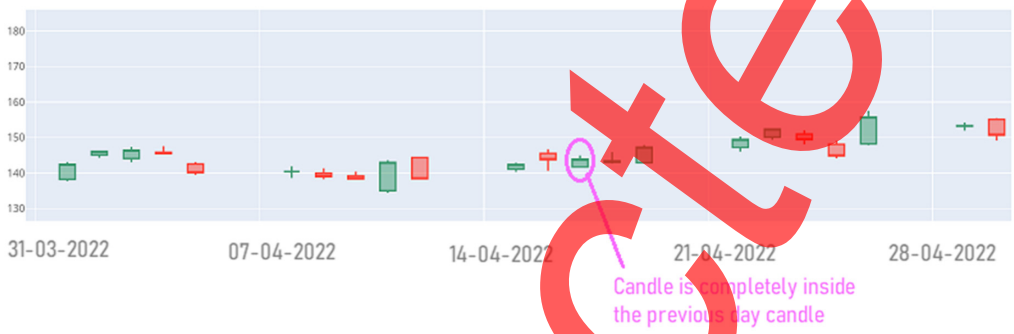

Fig. 4 Diagram for inside days.

4.3 *Check for Volatility Contraction*

For a stock that is already in an uptrend and is on an inside day, we check for volatility contraction. Stock will be in volatility contraction if its traded volume is among the least several traded volumes in the last 20 to 25 days. The graph in Fig. 5 shows the volatility contraction condition for MRK stock.

If a stock fulfills these three criteria, then it is likely that in the upcoming days the price of the stock will rise. So we consider it as a potentially profitable stock.

5 System Design

In this section, the complete methodology is discussed. To begin with, the abstract flow diagram is discussed, which is a high-level view of the system. Then the data-flow diagram is discussed, which is a complete view of all the processes involved.

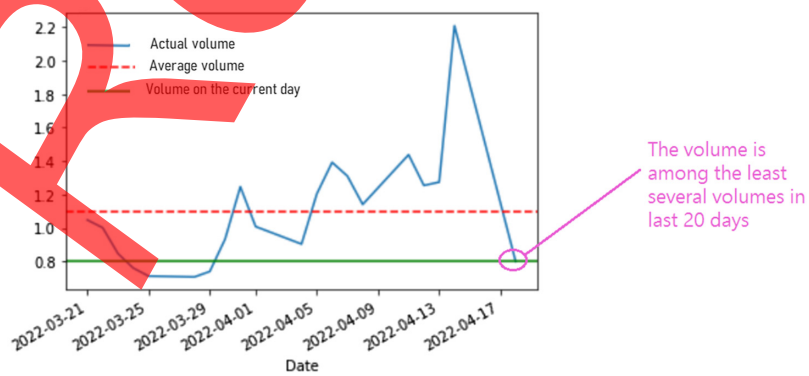

Fig. 5 Diagram for volatility contraction.

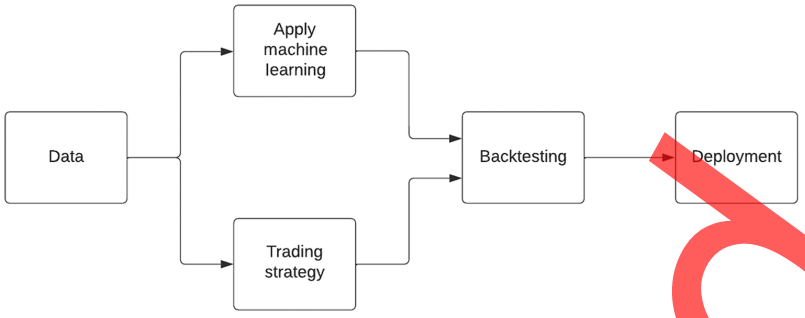

Fig. 6 Abstract flow diagram.

5.1 *High-level View of the System*

1. Sourcing and managing the data. This is the most important step of the process. Data collection is very important because it can give insights into the trends in the past. This data is necessary to train machine learning models also, which is an important step. Many popular APIs are available to collect the data. The most popular is the Yahoo Finance API, which is used in this paper.
2. Creating the trading strategy. It includes design, evaluation, and a combination of factors that can generate high alphas. An effective trading strategy is discussed in Sec. 4, which will tell about the potentially profitable stocks.
3. Apply machine learning. Since the previous data on stocks is easily available, so it is implied to apply supervised machine learning algorithms to find profitable stocks. In Sec. 2.2 various supervised learning algorithms are discussed.
4. Backtesting of the strategy developed. Before trading in the real stock market it is necessary to test the strategy and the model developed. For this purpose, backtesting is used. In backtesting, the model is tested on the historical data and if the model does well on the Historical data, then only it is used in the real market. The abstract flow diagram is presented in Fig. 6.

5.2 *Detailed Data Flow System Diagram*

Here, first we have Entity Yahoo Finance,[33] which we can use to get the stock price data, as well as serve the purpose of getting the historical data for back-testing. Then the other entity is for data retrieval and transformation, which preprocesses the data. It takes the raw data obtained from Yahoo Finance as input and removes rows will null values. It gives the processed dataset as output.

On the formatted data, the trading strategy is applied, which can also be seen from Fig. 7. Here, those three conditions discussed in Sec. 4 are checked. This gives us potentially profitable stocks.

Then, we do the predicted trend analysis, i.e, calculating the technical indicators, which can be seen from the diagram. The input is the processed dataset for potentially profitable stocks and the output is the calculated technical indicators, which are added to the corresponding dataset.

These datasets with indicators are then provided as inputs to the machine learning algorithms, which are here decision tree regression, multiple linear regression, and SVM regression, this can be seen from the middle part of the diagram. The models are trained on the dataset.

Then these algorithms will predict results for the upcoming days. These results can be verified by back-testing, where first the price is predicted for future days, then actual price data is fetched from Yahoo Finance for the predicted days and then it can be decided whether the price of the stock went up or down on a particular day. If the price actually went up and the model also predicted the same, or the price actually went down and the model also predicted the same then the prediction is considered correct otherwise wrong. The results from all 3 models are combined by voting, where if any 2 or all 3 models predict that the price will go up then up is taken as the final result otherwise down will be the final result. Based on this up or down signal the trades can be executed.

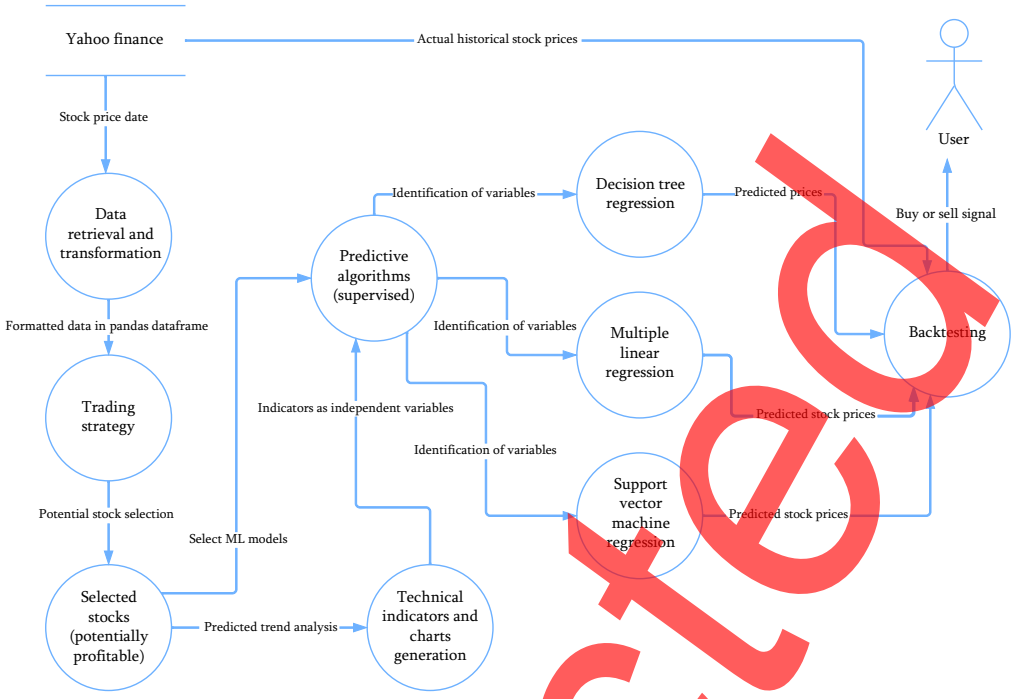

Fig. 7 Data flow diagram.

6 Results and Inferences

We applied the trading strategy on S&P500 stocks between January 1, 2018 and December 31, 2018 to find the potentially profitable stock each month. Then the technical indicators were calculated for all the stocks for the respective months. Then these indicators were used as input parameters for all three supervised machine learning models. Then the stock was tested for the next 50 days starting from the next month on-wards. The list of the potentially profitable stocks found along with the number of days when the stock price actually increased and the number of days when each of the machine learning models predicted that the prices will increase is given in Table 4.

Table 4 Potentially profitable stocks along with the number of days.

Month	Stock	Actual days	Multiple linear regression	SVM regression	Decision tree regression
January	AAP	25	19	19	15
February	NFLX	19	24	25	20
March	BSX	22	24	21	21
April	TRIP	21	17	29	31
May	UA	19	27	25	29
June	AMZN	20	26	27	14
July	HCA	17	22	19	12
August	AMD	18	25	31	11
September	ILMN	24	22	21	21
October	ABMD	24	23	28	22
November	FTNT	28	18	23	16
December	DIS	24	22	21	21

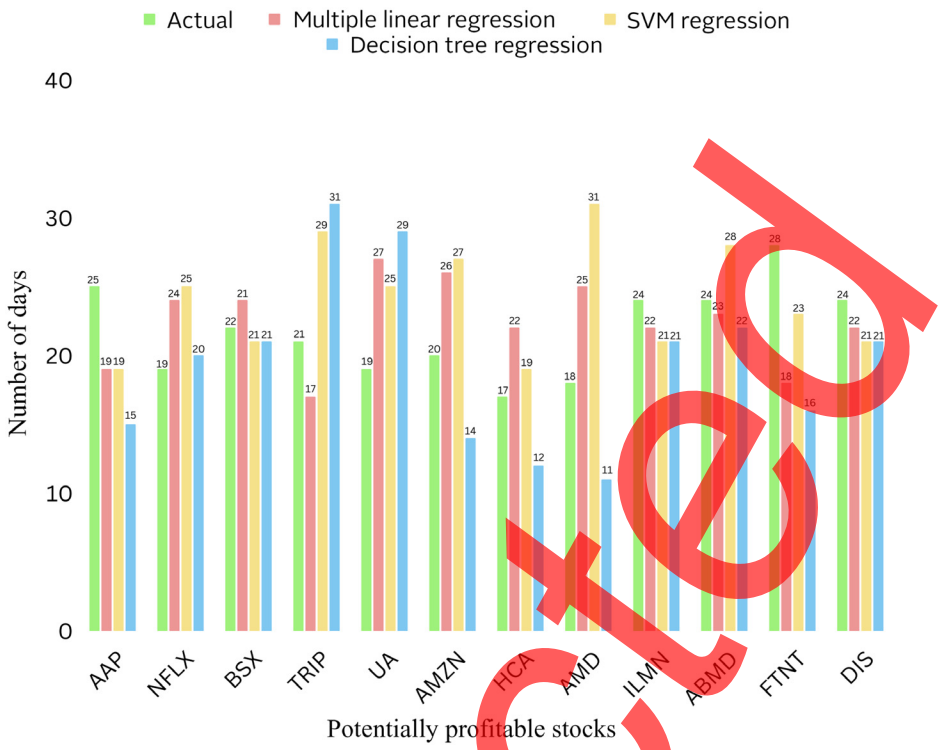

Fig. 8 Actual and predicted number of days by different models for potentially profitable stocks.

The histogram for the actual number of days when the price increased for each stock and the predicted number of days for all three models is shown in Fig. 8.

The formula for the percentage error used is:

$$\%\text{error} = \frac{|\text{actual days} - \text{predicted days}|}{\text{actual days}} \times 100. \tag{2}$$

Based on the formula, the error calculated for the stocks for the three supervised machine learning models are shown in the graph as shown in Fig. 9.

Other error metrics used are MSE and root mean squared error (RMSE). MSE and RMSE are commonly used metrics to evaluate the performance of a machine learning model. They are used to measure the difference between the predicted values and the actual values in a dataset.

The formula for the MSE is

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(D_i - \hat{D}_i)^2. \tag{3}$$

The formula for the RMSE is

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(D_i - \hat{D}_i)^2}, \tag{4}$$

where n is the number of stocks tested, here $n = 12$, D_i is the actual number of days when the stock price increases, and $\hat{D}_i$ is the predicted number of days by each machine learning algorithm.

MSE and RMSE are significant because they provide a quantitative measure of the difference between the predicted and actual values. They allow us to evaluate the accuracy of the model's predictions and identify areas where the model is performing well or poorly.

Table 5 shows the MSE and RMSE for different machine learning models on the 12 resultant stocks.

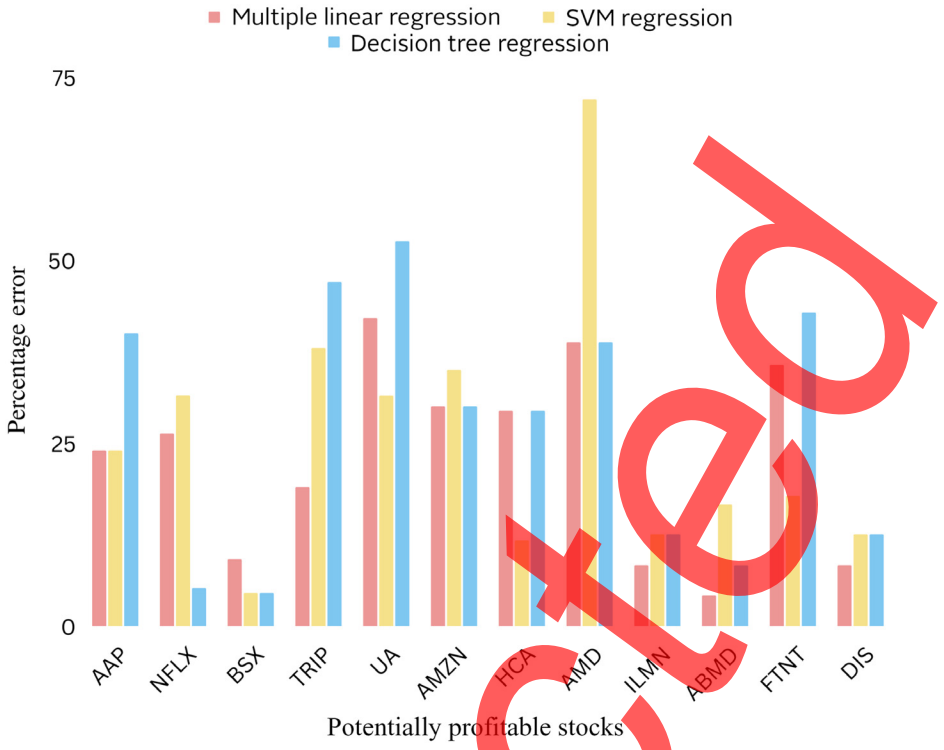

Fig. 9 Error in prediction.

Table 5 MSE and RMSE for different machine learning algorithms.

Type of error	Multiple linear regression	SVM regression	Decision tree regression
MSE	30.33	37.83	48.16
RMSE	5.51	6.15	6.93

The RMSE achieved is between 5 and 7, which is the maximum number of days the model is predicting wrong. It implies that our model is a good tool for profitability as out of 30 days, it is predicting correctly ~23 to 25 days.

7 Conclusion and Future Work

In this paper, we have proposed a supervised machine learning and technical analysis based model to predict Stock profitability in algorithmic trading. We tested our predictive models on twelve different potentially profitable stocks. The raw data for the stock price was collected using Yahoo Finance for the period January 2018 to December 2018. Results were presented on the performance of the three different machine learning models based on the correct number of days in prediction. It was observed that multiple linear regression produced the least MSE and RMSE for prediction and the decision tree produced the highest MSE and RMSE.

The prediction of the stock market can be made robust by analyzing the sentiment of the people, which is driven by news related to various stocks and general market conditions. So technical analysis can be combined with sentiment analysis to make better predictions of the market, in the future.

References

1. J. Chen, "Algorithmic trading," (2022).

2. Y. Zhu and G. Zhou, "Technical analysis: an asset allocation perspective on the use of moving averages," *J. Financial Econ.* **92**(3), 519–544 (2009).
3. N. Rouf et al., "Stock market prediction using machine learning techniques: a decade survey on methodologies, recent developments, and future directions," *Electronics* **10**(21), 2717 (2021).
4. S. Seth, "Basics of algorithmic trading: concepts and examples," (2022).
5. R. Peachavanish, "Stock selection and trading based on cluster analysis of trend and momentum indicators," (2016).
6. M. Wu and X. Diao, "Technical analysis of three stock oscillators testing MACD, RSI and KDJ rules in SH & SZ stock markets," in *4th Int. Conf. Comput. Sci. and Netw. Technol. (ICCSNT)*, Harbin, China, pp. 320–323 (2015).
7. V. Gururaj and R. Shriyav, "Stock market prediction using linear regression and support vector machines," (2019).
8. M. Miró-Julià, G. Fiol-Roig, and A. P. Isern-Deyà, "Decision trees in stock market analysis: construction and validation," *Lect. Notes Comput. Sci.* **6096**, 185–194 (2010).
9. Z. Hu, J. Zhu, and K. Tse, "Stocks market prediction using support vector machine," in *6th Int. Conf. Inf. Manage., Innov. Manage. and Ind. Eng.*, Vol. 2, pp. 115–118 (2013).
10. G. Bland, "Yahoo finance API - a complete guide," (2021).
11. J. Fernando, "Moving average (MA): purpose, uses, formula, and examples," (2022).
12. B. Dolan, "MACD indicator explained, with formula, examples, and limitations," (2022).
13. J. Fernando, "Relative strength index (RSI) indicator explained with formula," (2022).
14. C. Schaap, "ADX: the trend strength indicator," (2022).
15. F. Y. Osisanwo et al., "Supervised machine learning algorithms: classification and comparison," *Int. J. Comput. Trends Technol.* **48**, 128–138 (2017).
16. C. Cheng, W. Xu, and J. Wang, "A comparison of ensemble methods in financial market prediction," in *Fifth Int. Joint Conf. Comput. Sci. and Optim.*, pp. 755–759 (2012).
17. J. Yuan and Y. Luo, "Test on the validity of futures market's high frequency volume and price on forecast," in *Int. Conf. Manage. of e-Commerce and e-Government*, Shanghai, China, pp. 28–32 (2014).
18. J. Patel et al., "Predicting stock market index using fusion of machine learning techniques," *Expert Syst. Appl.* **42**(4), 2162–2172 (2015).
19. Y. E. Cakra and B. Distiawan Trisedya, "Stock price prediction using linear regression based on sentiment analysis," in *Int. Conf. Adv. Comput. Sci. and Inf. Syst. (ICACSIS)*, pp. 147–154 (2015).
20. W. Khan et al., "Predicting trend in stock market exchange using machine learning classifiers," (2016).
21. N. Milosevic, "Equity forecast: predicting long term stock price movement using machine learning," ArXiv abs/1603.00751 (2016).
22. A. G. Rossi, "Predicting stock market returns with machine learning," (2018).
23. K. Vanukuru, "Stock market prediction using machine learning," (2018).
24. C.-L. Huang and C.-Y. Tsai, "A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting," *Expert Syst. Appl.* **36**, 1529–1539 (2009).
25. J. Zhang, Y.-F. Teng, and W. Chen, "Support vector regression with modified firefly algorithm for stock price forecasting," *Appl. Intell.* **49**, 1–17 (2019).
26. P.-F. Pai et al., "Time series forecasting by a seasonal support vector regression model," *Expert Syst. Appl.* **37**, 4261–4265 (2010).
27. P. K. Illa, B. Parvathala, and A. K. Sharma, "Stock price prediction methodology using random forest algorithm and support vector machine," *Mater. Today Proc.* **56**, 1776–1782 (2022).
28. P. Yoo, M. Kim, and T. Jan, "Machine learning techniques and use of event information for stock market prediction: a survey and evaluation," in *Int. Conf. Comput. Intell. for Model., Control and Autom. and Int. Conf. Intell. Agents, Web Technol. and Internet Commerce (CIMCA-IAWTIC'06)*, Vol. 2, pp. 835–841 (2005).
29. X. Zhong and D. Enke, "Forecasting daily stock market return using dimensionality reduction," *Expert Syst. Appl.* **67**, 126–139 (2017).

30. J. M.-T. Wu et al., "A graph-based CNN-LSTM stock price prediction algorithm with leading indicators," *Multimedia Syst.* (2021).
32. M. Sedighi et al., "A novel hybrid model for stock price forecasting based on metaheuristics and support vector machine," *Data* **4**(2), 75 (2019).
33. R. Aroussi, "Yahoo! Finance's API i.e. yfinance 0.2.9," yfinance PyPI, https://pypi.org/project/yfinance/ (accessed 7 February 2023)

Shubham Khandelwal is a final year BTech student in the Computer Science and Engineering Department at the Indian Institute of Information Technology, Kota. His research interests are in the fields of machine learning, algorithmic trading, and natural language processing.

Piyush Gupta is a final year BTech student in the Computer Science and Engineering Department at the Indian Institute of Information Technology, Kota. His research interests are in the fields of machine learning and cloud computing.

Aman Jain is pursuing a BTech degree in the Computer Science and Engineering Department at the Indian Institute of Information Technology, Kota. His research interests are in machine learning and natural language processing.

Ajay Nehra completed his doctoral research from the Department of Computer Science and Engineering at Malaviya National Institute of Technology, Jaipur, India, in July 2019. He received his MTech degree in computer science and engineering from Central University of Rajasthan, India, in 2012. He is currently working as an assistant professor at the Indian Institute of Information Technology, Kota. His current research areas include software-defined networking, information security, algo trading, and AI in pedagogy.

Gyan Singh Yadav received his MTech and PhD degrees from the Department of Computer Science and Engineering, IIITDM Jabalpur. He is an assistant professor in the Department of Computer Science and Engineering, IIIT Kota. He has published several research papers in journals and conferences of international repute. His research interests include cryptography, steganography, quantum computing, and cyber security issues.

Riti Kushwaha received her PhD from Malaviya National Institute of Technology. She is a researcher in the field of artificial intelligence, machine learning, and deep learning. With over 10 years of research experience, she has published several articles in highly regarded journals, showcasing her in-depth knowledge and innovative thinking in the field of AI, ML, and DL. Currently, she is working at Bennett University.

Ramani Selvanambi received his BE degree in computer science and engineering from Madras University, his MTech degree in computer science and engineering from Bharathidasan University, and his PhD in computer science and engineering from Vellore Institute of Technology (VIT), Vellore. He is an associate professor at the School of Computing Science and Engineering, VIT, Vellore, India.