Machine Learning for Predicting the Stock Price Direction with Trading Indicators

Md. Siam ANSARY

Abstract—There are a number of possible advantages in utilising trading indicators and machine learning to predict the direction of stock prices. It is crucial to remember that stock price prediction is difficult by nature and that there is no way to ensure accuracy. The financial markets are extremely information-rich, dynamic, and complex. Large datasets may be processed and analysed by machine learning algorithms far more quickly than by people, which makes it possible to spot patterns or trends that might not be immediately obvious. By using historical price and volume data, the algorithms can be trained to identify patterns that could predict future moves. Because they offer insights into possible market moves, predictive models can help with risk management. Because ML models are always learning from fresh data, they can adjust to changing market conditions. This flexibility is essential in markets where a variety of factors impact the market. Several machine learning models have been used in this experimental effort to monitor the direction of stock prices, and the outcomes are extremely encouraging.

***Keywords*—stock, dhaka stock exchange, yahoo finance api, machine learning, classification.**

Nomenclature

ENP	Ecole Nationale Polytechnique.
ML	Machine Learning.
AI	Artificial Intelligence.
SVM	Support Vector Machine.
KNN	K Nearest Neighbor.
RF	Random Forest.
DT	Decision Tree.
MLP	Multilayer Perceptron.
AdaBoost	Adaptive Boosting.
API	Application Programming Interface.
MFI	Money Flow Index.
RSI	Relative Strength Index.

I. Introduction

A number of factors, such as complexity, volatility, and the abundance of variables that might affect market movements, make stock market research difficult. Financial markets are intricate systems that are impacted by a wide range of variables, such as emotions among investors, firm performance, economic indicators, and world events. A thorough and multidimensional approach is necessary to comprehend the interactions among these elements. Every day, the stock market produces enormous amounts of data. It can be very difficult to go through all of this data and find trustworthy and pertinent sources. To make wise selections, traders and investors must be able to discern between irrelevant signals and meaningful ones. Due to their dynamic nature, financial markets are susceptible to sudden changes in news, events, or macroeconomic conditions. It's never easy to stay current with information and adjust to shifting market conditions. Market movements are significantly impacted by investor behaviour, which is driven by herd mentality, emotions, and market psychology. Research on stock markets becomes even more challenging when one attempts to comprehend and forecast human behaviour. It is difficult to forecast future price movements in financial markets due to the inherent risk and uncertainty involved. Economic downturns, market corrections, and unforeseen events can all have a big influence on the results of investments. Financial markets are becoming more interconnected and susceptible to global events and developments as a result of globalisation. Events in one region of the world that are political, economic, or geopolitical may reverberate throughout markets worldwide. The regulatory landscape is subject to change, which has an effect on how businesses function and how investments are managed. Maintaining up-to-date knowledge of regulatory modifications is essential for precise market analysis. A balance between quantitative analysis (financial statements, ratios, etc.) and qualitative analysis (industry trends, management quality, etc.) is necessary for stock market research. The research method becomes more complex when both elements are integrated. Examining price charts, patterns, and indications is part of technical analysis. Although it can offer insightful information, chart interpretation is a discipline that takes expertise and skill. Meaningful analysis requires data that are accurate and reliable. It is imperative for traders and investors to guarantee that the data they depend on is current, precise, and accurately depicts the actual market conditions. Traders and investors are susceptible to biases or overconfidence that impair their judgement. Effective market research requires retaining objectivity and being conscious of cognitive biases. It takes a combination of experience, knowledge, analytical abilities, and a methodical approach to decision-making to successfully navigate these hurdles. A lot of market players also use technology, such as data analytics and machine learning, or consult experts to improve their research skills. Compared to people, machine learning algorithms can process and analyse massive datasets much more quickly, which makes it easier to spot patterns or

Manuscript received December 28, 2022; revised December 14, 2023.

Md. S. ANSARY is with the Computer Science and Engineering Department, Ahsanullah University of Science and Technology, Dhaka, BANGLADESH (e-mail address: siamansary.aust@gmail.com).

Digital Object Identifier (DOI): 10.53907/enpesj.v3i2.178

trends that might not be immediately obvious. Automated market data analysis and fast prediction generation are possible with machine learning models. This can be especially useful in quick-paced marketplaces where making decisions quickly is essential. Patterns in past price and volume data can be used to train machine learning algorithms to identify possible future movements. This involves seeing patterns on charts or technical indicators that are hard for people to interpret. With their ability to forecast future market moves, predictive models can help with risk management. In order to reduce the risks brought on by market volatility, traders and investors can modify their methods in light of the anticipated direction. A variety of indications and factors that can affect stock values can be quantitatively analysed thanks to machine learning. More unbiased and data-driven decision-making may result from this. To evaluate the performance of machine learning models, past data can be used for backtesting. This enables investors and traders to assess the performance of their methods in various market scenarios. Because machine learning models are always learning from fresh data, they may adjust to shifting market conditions. In markets where a variety of factors impact the market, this adaptability is essential. In this experiment, we used a variety of trading indicators to attempt to forecast the direction of the stock price in order to make investment decisions. Prediction techniques should be flexible enough to accommodate the rapid fluctuations that the stock exchange can encounter. For these kinds of trials, methods other than machine learning models are less appropriate because ML classifiers are more adept at learning from datasets and adapting to potential changes.

We review the previous research on our experiment, its methodology, and its results in the parts that follow.

II. Literature Review

Several studies have focused on the use of machine learning algorithms to forecast stock values. Various strategies have been investigated in order to detect patterns in historical data and predict future price changes. Effective feature engineering is critical for enhancing machine learning model performance. As potential elements for predicting stock prices, researchers studied several financial indicators, technical analysis metrics, and sentiment analysis from news or social media. Random forests and gradient boosting are two common ensemble learning approaches. These methods combine numerous models' predictions to improve overall accuracy and resilience. Chen and Hao [1] concentrated on stock market index prediction, which is important for investing and applications that aim to maximise earnings at minimal risk. In order to effectively predict stock market indexes, their study highlights the use of machine learning (ML). Specifically, they propose a hybrid framework that combines feature weighted support vector machines and feature weighted K-nearest neighbours. To evaluate the effectiveness of the suggested approach, the researchers ran tests on the indices of the Shanghai and Shenzhen stock exchanges. The findings show that for short-, medium-, and long-term predictions of the Shanghai Stock Exchange Composite Index and Shenzhen Stock Exchange Component Index, the created model has the ability to attain higher prediction skills. A Fuzzy Metagraph (FM) based method for stock market classification, prediction, and decision making for short-term investors in the Indian stock market was put out by Anbalagan and Maheswari [2]. The research employed the FM technique to train the system while using Technical Indicators. Stocks listed on the Indian Bombay Stock Exchange (BSE) are used to assess the success of the suggested model. According to the findings, the FM-based model produces performance that is satisfactory and has very little risk error. He et al. [3] conducted research on pharmaceutical companies' stock price movement prediction during the COVID-19 pandemic. The researchers gathered extensive financial data unique to the company, including historical and present stock prices, key financial indicators, worldwide industry indices, and information about COVID-19. They had gathered information from a variety of sources, including Our World in Data [4] and the Yahoo! Finance API [5]. These disparate datasets were combined into a single, day-oriented dataset that could be used to train several predictive models. The researchers used worldwide industry indexes and created a set of characteristics especially for COVID-related data. The models' capacity for prediction was much improved by the addition of these features. The findings showed that adding features from industry indexes increased AUROC by an average of 1.7%. Additionally, the addition of COVID-related features led to a 6.5% average AUROC rise. The research investigated how the exclusion of early COVID era data affected the performance of the model. With an average increase in AUROC of 3.9% and an average increase in accuracy of 1.9%, the results showed significant gains. The decision tree model using the gradient boosting approach was able to attain a 70% AUROC and 66% accuracy as a result of this modification. Ballings et al. [6] attempted to forecast the direction of stock prices using information obtained from European companies that were publicly traded. Their primary goal was to determine which kind of algorithm worked better for single learners or ensemble learners. Logistic Regression, Neural Networks, K Nearest Neighbour, Support Vector Machine, Random Forest, AdaBoost, and Kernel Factory were the techniques employed in this study. Random Forest fared the best among the methods. For cross validation, they had employed five times the twofold. As features of this experiment, various financial indicators such as liquidity indicators, solvency indicators, and profitability indicators were used. Kara et al. [7] used ANN and SVM to forecast the direction of a stock price index. The Istanbul Stock Exchange data was used in the study. Artificial Neural Networks obtained 75.74% accuracy, while Support Vector Machines achieved 71.52% accuracy. In the experiment, various indicators were used. Basak et al. [8] used Random Forest and Gradient Boosted Decision Trees to forecast the direction of stock market prices. As features for the ML models, the work used the Relative Strength Index, Stochastic Oscillator, Williams Percentage Range, Moving Average Convergence Divergence, Price Rate of Change, and On Balance Volume. Parmar et al. [9] used the Yahoo Finance dataset to predict stock prices using Open, High, Low, Close, and Volume information. 20% of the dataset was used for testing, while the remaining 80% was used to train models. For stock price movement classification, Naik and Mohan [10] used data from the National Stock Exchange of India. As features, indicators such as the Simple Moving Average, the Exponential Moving Average, the Momentum Indicator, the Stochastic Oscillator, the Moving Average Convergence Divergence, the Relative Strength Index, the William R, the Accumulation Distribution Index, and the Commodity Channel Index were employed.

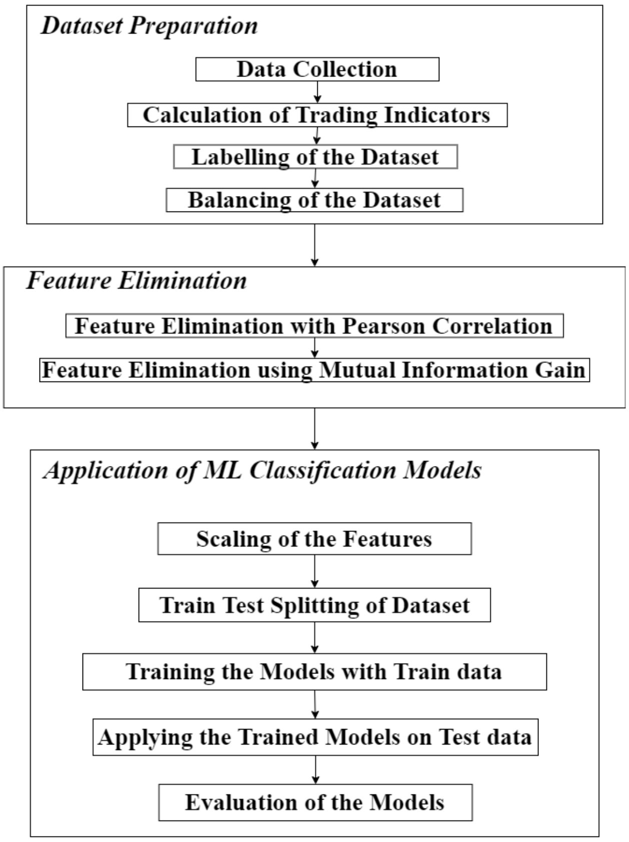
Dataset Preparation
Data Collection
Calculation of Trading Indicators
Labelling of the Dataset
Balancing of the Dataset
Feature Elimination
Feature Elimination with Pearson Correlation
Feature Elimination using Mutual Information Gain
Application of ML Classification Models
Scaling of the Features
Train Test Splitting of Dataset
Training the Models with Train data
Applying the Trained Models on Test data
Evaluation of the Models

Fig. 1: Steps of Proposed Methodology

III. Proposed Methodology

A. Dataset Preparation

Firstly, stock data has been collected. This collected data has the Open, High, Low, Close, Volume values of each of the stocks. We have extracted important features from the dataset. Using Technical Analysis Library [11], based on the Open, High, Low, Close and Volume values of each stock, different trading indicators have been calculated so that they can be used as features. The calculated indicators are as below.

- Moving Average (MA)
- Balance of Power (BOP)
- Money Flow Index (MFI)
- Momentum (MOM)
- Rate of Change (ROC)
- Relative Strength Index (RSI)
- Weighted Close Price (WCP)
- Average Price (AP)
- Median Price (MP)
- Typical Price (TP)
- On Balance Volume (OBV)

With the MFI and RSI values of each stock, the stocks have been labeled. If the value of MFI is greater than 80 or the value of RSI is greater than 70, a stock is overbought. Overbought condition signifies the market is upwards. If the value of the MFI is less than 20 or the value of RSI is less than 30, a stock is oversold. Oversold signifies that at the current moment, the stock is going downwards. The stocks that are neither overbought nor oversold, are neutral. In such a manner, all the stocks have been labeled. After labelling the stocks as upwards, downwards or neutral, we have observed most stocks to be neutral, some stocks to be of upwards direction and very few records to be of downwards direction. Hence, for classification task, the dataset is not balanced. Hence, at first, we have used Random Undersampling on Neutral Stocks so that number of neutral records come down to the number of upwards stocks and then, applied Random Oversampling so that number of downwards stocks match the numbers of upwards stocks. We have used two datasets for our work. For the first dataset, we have at first collected data using the Yahoo! Finance's API [5]. Initially, we have collected 5630 samples. The samples are of the period from 01 January 2000 to 17 May 2022. After labelling the stocks, we have observed 5033 stocks to be neutral, 426 stocks to be of upwards direction and 171 records to be of downwards direction. The total number of records of the balanced dataset becomes 1278 where each class has same amount of records. For yhe second dataset, we used data of Dhaka Stock Exchange (DSE). Initially, there aere 90041 records in the collection. Following the classification of stocks as moving upwards, downwards, or neutral, we discovered that 66004 stocks were in a neutral state, 14648 stocks were traveling upward, and 9389 records were heading downwards. The balanced dataset contains 43944 records overall, with an equal number of each kind.

B. Feature Elimination

After ensuring that the dataset is balanced, we attempted to determine whether all of the characteristics in our dataset are important, as unnecessary features might raise the complexity of the problem and cause overfitting. On the dataset, we used the Pearson Correlation method. Figure 2 shows the heatmaps of the values for the first dataset and Figure 3 shows the heatmaps of the values for the second dataset. The threshold values for the two datasets have been set to 0.8 and 0.9, respectively, to remove superfluous features. We used the Mutual Information Gain technique on the existing columns. The threshold value has been set to 0.5 in order to further exclude inconsequential features, suggesting that features with information gains less than 0.5 have been deleted.

C. Application of ML Models

Following feature elimination, we used Standard Scaling on the remaining features. Standard scaling was used to bring all features into a similar value range because various feature ranges can produce undesired bias. Then, we divided our dataset into two parts. 80% of the dataset was utilised for training, while the remaining 20% was used for testing. Five classification models were used. They are listed below.

- K Nearest Neighbor Classifier:The KNN classifier is a simple and effective supervised machine learning technique. KNN is a sluggish, non-parametric learning method. Non-parametric means that no assumptions are made about the

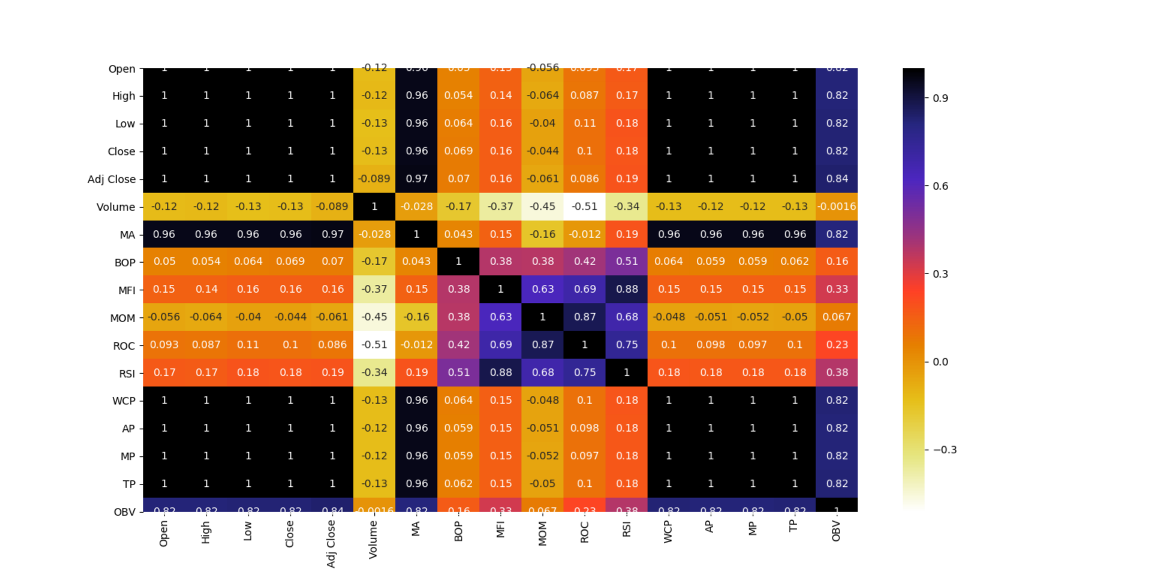

Fig. 2: Pearson Correlation Heatmap of Features for first Dataset of Yahoo! Finance's API

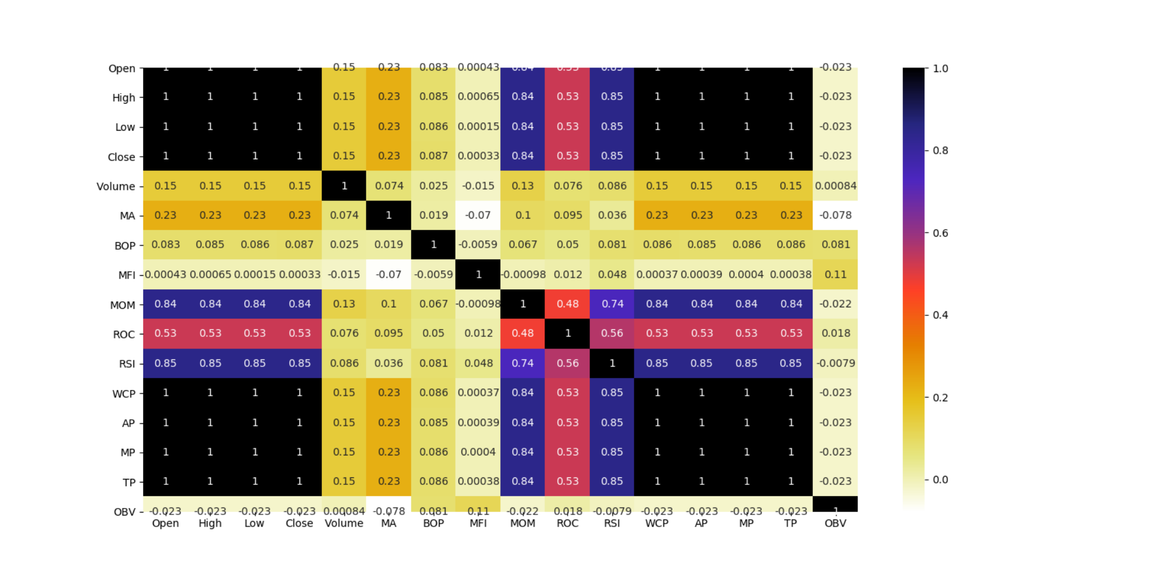

Fig. 3: Pearson Correlation Heatmap of Features for second dataset of Dhaka Stock Exchange

underlying data distribution, and lazy learning indicates that the method does not construct an explicit model during the training phase. It instead remembers the training data. Given a new, unknown data point, the method uses a distance measure to identify the K nearest data points in the training set. The parameter K (number of neighbours) must be chosen carefully because it has a substantial impact on the model's performance. A little K can cause overfitting, whereas a large K can cause underfitting. The ideal value of K is determined by the dataset's unique properties. The distance metric used to compare the similarity of data points is critical in KNN. While Euclidean distance is the most commonly used, alternative metrics such as Manhattan distance or Minkowski distance might be used depending on the nature of the data. Because the KNN method is sensitive to feature scale, feature scaling is frequently required. Normalisation or standardisation of the features guarantees that no single feature dominates the distance calculation. The main computational expense of KNN occurs during the prediction phase, when the distances between the new point and all training points must be calculated. Large datasets can be computationally expensive. The curse of dimensionality can impair KNN, which means that as the number of features increases, the distance between points tends to become more uniform, potentially diminishing the algorithm's effectiveness. Techniques like as feature selection and dimensionality reduction can help to alleviate this problem. KNN is a versatile and intuitive method, but its effectiveness is dependent on careful tweaking of hyperparameters and consideration of data properties. It is frequently used as a baseline model for comparison with more complex algorithms, and its simplicity makes it excellent for short exploratory data analysis.

- Decision Tree Classifier: The Decision Tree Classifier is a common supervised machine learning method. It is a tree-like model in which each core node represents a characteristic or attribute, the branches represent decision rules, and each leaf node corresponds to the anticipated outcome. A decision tree is a hierarchical structure in which nodes indicate decisions based on attributes. The root is the top node, and the leaves are the bottom nodes, each representing a class label or a numeric value. In the tree, decision nodes reflect questions regarding the input data. These queries lead to succeeding nodes, and the structure continues until a leaf node is reached, offering the ultimate judgement or forecast. To select the appropriate split for each node, the method employs metrics such as Gini impurity or information gain. The tree-building procedure entails recursively partitioning the dataset based on the features chosen. This process is repeated until a stopping requirement, such as a preset tree depth, a minimum amount of samples in a leaf, or no further improvement is possible, is met. It explores the decision tree from root to leaf, using the decision rules, to create a forecast for a new data point. The prediction is subsequently assigned to the class label or value of the leaf node.

- Random Forest Classifier: The Random Forest Classifier is a classification technique that creates numerous decision trees during training and reports the mode of the classes. To generate numerous subsets, the training dataset is randomly sampled with replacement. The forest's decision trees are then trained on one of these subsets. A random subset of features is considered for each split in a decision tree rather than all features. This adds to the randomness and diversity of the trees. The class indicated by the majority of trees is chosen as the final prediction in classification problems. When compared to individual decision trees, Random Forests are less prone to overfitting. The variety generated by training on diverse subsets of data and features aids in generalisation to previously unseen material. Because of its durability, ease of application, and strong generalisation performance, Random Forests are commonly utilised in practice.

- AdaBoost Classifier: AdaBoost is an ensemble learning method that combines the predictions of numerous weak learners to generate a strong and accurate predictive model. It lends additional weight to occurrences that were incorrectly classified by earlier models, allowing subsequent models to focus on the faults. Each data point in the training set is given a weight, which is changed at each iteration based on the performance of previously trained models. Misclassified points are weighted more heavily, emphasising them in subsequent iterations. Weak learners are trained in a sequential manner, with each new learner correcting the mistakes of the preceding ones. The final forecast is a weighted average of the individual weak learners' predictions. AdaBoost is especially successful in dealing with complex cases that are difficult for a single weak learner to correctly classify. AdaBoost prioritises troublesome cases by applying higher weights to misclassified instances. The algorithm is adaptive in the sense that it adapts its strategy based on the mistakes made by previous models. AdaBoost's versatility enables it to continuously enhance its performance. Each weak learner's contribution is weighted based on its accuracy during the final prediction. More precise models have a greater impact on the final decision. AdaBoost is a sophisticated algorithm that has gained popularity due to its ability to improve the performance of poor learners and generate robust models. It is frequently employed as a building block in ensemble methods and has the potential to dramatically increase forecast accuracy.

- Gradient Boosting Classifier: Gradient Boosting is an ensemble learning strategy that combines the predictions of numerous weak learners, typically decision trees, to create a powerful predictive model. The Gradient Boosting Classifier, in particular, is utilised for classification problems. Gradient Boosting gradually constructs an ensemble of weak learners. Each new learner corrects the mistakes committed by the existing learners as a group. Using gradient descent optimisation, the approach minimises a loss function. It computes the loss gradient with respect to the ensemble prediction and modifies the model parameters to minimise the loss. Gradient Boosting is concerned with minimising the residuals, or mistakes, of the existing ensemble's predictions. To predict the residuals, new trees are trained, eventually reducing the overall error. A shrinkage parameter, often known as the learning rate, governs each tree's contribution to the ensemble. A slower learning rate necessitates more trees for the same degree of complexity, but it can improve model robustness. Gradient Boosting involves overfitting prevention approaches such as limiting tree depth, introducing a penalty term for tree complexity, and subsampling data during training. Gradient Boosting is a powerful technique that has shown effectiveness in a variety of disciplines. When high predicted accuracy is required, it is a common alternative, and careful adjustment of hyperparameters is often required for best performance.

D. Performance Evaluation

The performances of the ML models are evaluated with Accuracy, Precision, Recall and F-1 Score evaluation metrics. Figure 4 illustrates the performances of the models for the first dataset which has been created with the data of Yahoo! Finance API. The following Figure 5 shows how the models perform for the second dataset that has been made with the data of Dhaka Stock Exchange. From observations of the performances of the ML models, it can be noted that their attainments have been satisfactory.

IV. Conclusion and Future Works

Stock price direction is difficult to predict due to a number of factors, and financial markets are influenced by a complex interaction of various elements. Stock prices are vulnerable to erratic and chaotic movements that are influenced by a plethora of variables. Unexpected news, geopolitical events, economic statistics, and investor attitude, among other things, have an impact on the market. Because of these variables, properly forecasting the direction of stock values is difficult. According to the Efficient Market Hypothesis, markets assimilate all relevant information into stock prices rapidly and efficiently. If this hypothesis is correct, it indicates that predicting price fluctuations based on previous data or publicly available data is difficult. Human behaviour, including emotions such as fear

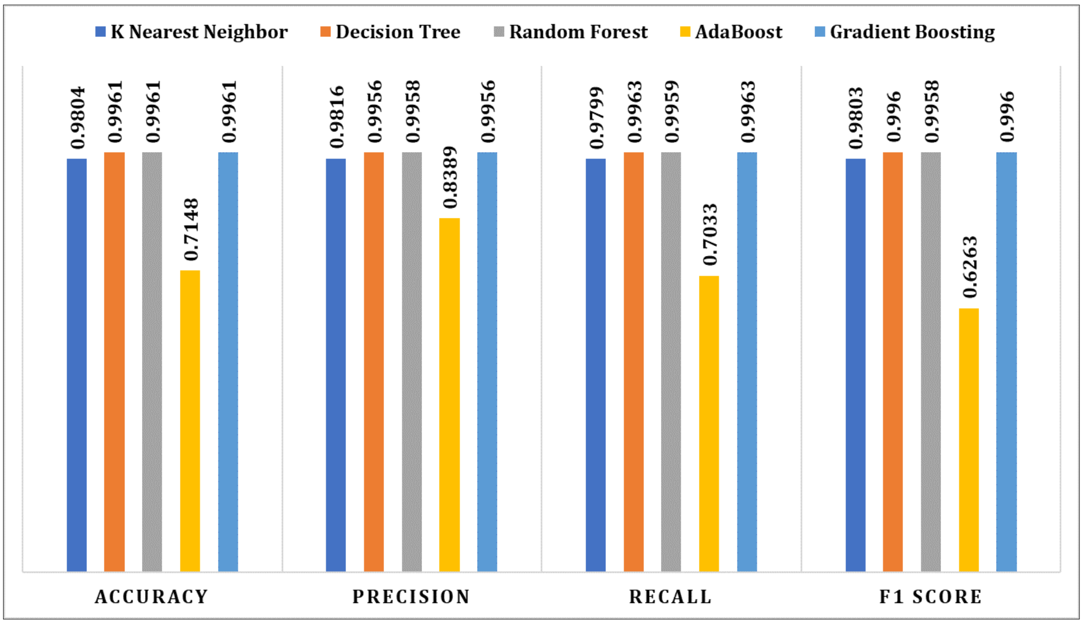

Fig. 4: Performance of ML Models for the first dataset

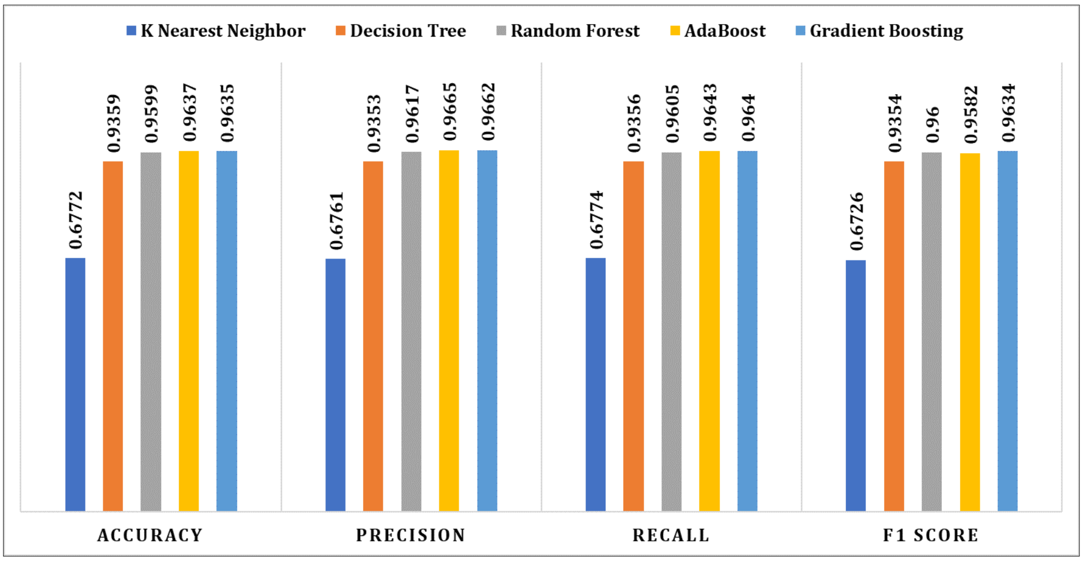

Fig. 5: Performance of ML Models for the second dataset

and greed, has a significant impact on stock values. According to behavioural finance, investors do not always make rational decisions, and market movements can be influenced by psychological factors that are difficult to quantify and anticipate. Not all market participants have simultaneous access to the same information. Insiders or large institutional investors may have access to privileged information, providing them an advantage over individual investors. Because of this information asymmetry, ordinary investors find it difficult to precisely predict stock price changes. In some circumstances, market manipulation occurs when specific businesses or individuals purposefully create misleading views in order to impact stock prices. This can add another element of uncertainty to the market. Stock prices can be greatly influenced by economic conditions, interest rates, inflation, and world events. It is difficult to correctly predict these macroeconomic factors, and their consequences on the market can be complex and varied. Financial markets create massive amounts of data, making it difficult to spot meaningful trends. Traditional models struggle to collect all essential information due to the vast volume of data and the speed with which markets operate. Market dynamics are frequently nonlinear, which means that little changes can have disproportionately huge consequences. This makes using linear models for accurate predictions difficult. Price movements in the short term can be influenced by factors such as news and sentiment, whereas long-term trends can be influenced by fundamental factors such as earnings, economic growth, and industry trends. The combination of these elements makes projecting both short-term and long-term trends difficult. It is critical to recognise the inherent uncertainty and risk in projecting stock values. Despite the various challenges and difficulties, in our experiment it has been seen that Machine Learning models can predict the direction of movement of prices of stocks with great efficiency. Of course it is not always possible to predict correctly but the more accurate the prediction is, the less risks there are. In the future, we intend to continue training our models with the new data of Yahoo! Finance API and DSE. Also, we will employ more pre processing techniques, ML models on more Stock Datasets so that we can scrutinize the capabilities of the ML approaches for predicting the direction of stock prices.

References

[1] Chen, Yingjun, and Yongtao Hao. “A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction." Expert Systems with Applications 80 (2017): 340-355.

[2] Anbalagan, Thirunavukarasu, and S. Uma Maheswari. “Classification and prediction of stock market index based on fuzzy metagraph." Procedia Computer Science 47 (2015): 214-221. https://doi.org/10.1016/j.procs.2015.03.200

[3] He, Beilei, Weiyi Han, and Suet Ying Isabelle Hon. “A Machine Learning Approach: Enhancing the Predictive Performance of Pharmaceutical Stock Price Movement during COVID." Journal of Data Analysis and Information Processing 10, no. 1 (2021): 1-21.

[4] “Our World in Data”, [Online] Available: https://ourworldindata.org/, Last Accessed on 27 March 2023.

[5] “Yahoo! Finance’s API”, [Online] Available: https://pypi.org/project/yfinance/, Last Accessed on 27 March 2023.

[6] Ballings, Michel, Dirk Van den Poel, Nathalie Hespeels, and Ruben Gryp. “Evaluating multiple classifiers for stock price direction prediction." *Expert systems with Applications* 42, no. 20 (2015): 7046-7056. https://doi.org/10.1016/j.eswa.2015.05.013

[7] Kara, Yakup, Melek Acar Boyacioglu, and Ömer Kaan Baykan. “Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange." Expert systems with Applications 38, no. 5 (2011): 5311-5319.

[8] Basak, Suryoday, Saibal Kar, Snehanshu Saha, Luckyson Khaidem, and Sudeepa Roy Dey. “Predicting the direction of stock market prices using tree-based classifiers." *The North American Journal of Economics and Finance* 47 (2019): 552-567.

[9] Parmar, Ishita, Navanshu Agarwal, Sheirsh Saxena, Ridam Arora, Shikhin Gupta, Himanshu Dhiman, and Lokesh Chouhan. “Stock market prediction using machine learning." In 2018 first international conference on secure cyber computing and communication (ICSCCC), pp. 574-576. IEEE, 2018.

[10] Naik, Nagaraj, and Biju R. Mohan. “Stock price movements classification using machine and deep learning techniques-the case study of indian stock market." In International Conference on Engineering Applications of Neural Networks, pp. 445-452. Springer, Cham, 2019.

[11] “Technical Analysis Library”, [Online] Available: https://ta-lib.org/, Last Accessed on 27 March 2023.

Md. Siam Ansary completed Bachelor of Science in Computer Science and Engineering from Ahsanullah University of Science and Technology of Dhaka, Bangladesh. Later, he completed Master in Information Technology degree from Institute of Information Technology of University of Dhaka. Currently, he is employed as a Lecturer at the Department of Computer Science and Engineering of Ahsanullah University of Science and Technology, Dhaka, Bangladesh. He is very much interested in research works related to Artificial Intelligence, Machine Learning, Natural Language Processing etc.