

Research Article

Stock Data Analysis by Machine Learning Techniques to Improve Predictive Modeling and Decision-Making in Financial Markets

Rasool F. Jader ^{1,*} , Mudhafar Haji M. Abd ¹ 

¹ Computer Science Department, Faculty of Science, Soran University, Soran, 44008, Iraq

*Corresponding Author: Rasool F. Jader, E-mail: rasool.jader@gmail.com

Article Info	Abstract
Article History	Stock price predictions remain a significant area of interest, particularly among experts in computer science and artificial intelligence, due to the dynamic nature of financial markets. Numerous studies have employed machine learning techniques to forecast stock market trends—an essential task for investors aiming to maximize returns. This study investigates the predictive capabilities of several machine learning models, including Support Vector Machine (SVM), Gradient Descent, Random Forest, and Adaptive Boosting. The models are applied to predict the behavior of the Bombay Stock Exchange (BSE), using key input features such as crude oil prices, gold and silver values, historical market performance, and foreign exchange rates. The predictive performance of each model is evaluated and compared. Among the factors considered, gold prices exhibit the strongest positive correlation with market performance. Notably, the Adaptive Boosting method outperforms the other models, demonstrating its potential for accurate stock market forecasting.
Received Mar 13, 2025	
Revised Jun 12, 2025	
Accepted Jun 15, 2025	
Keywords	
Machine Learning	
Stock Prediction	
Support Vector Machine	
Random Forest	
Gradient Boosting	
RNN	
Adaptive Boosting	
 Copyright: © 2025 Rasool F. Jader, and Mudhafar Haji M. Abd. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) license.	

1. Introduction

Recent advances in parallel computing, big data analytics, and machine learning, particularly deep learning, have transformed predictive modeling across industries [1, 2]. In the financial sector, these technologies enable the processing of vast and complex datasets to uncover hidden patterns and improve the accuracy of market forecasts [3]. Models such as deep neural networks, supported by high-performance computing, are now widely used to predict stock market trends by analyzing historical prices, macroeconomic indicators, and investor sentiment extracted from news and social media. These developments have reinforced the role of artificial intelligence in financial decision making.

Stock markets are crucial to economic development, particularly in emerging economies, where they serve as both indicators of economic health and mechanisms for raising capital [4]. However, despite their importance, stock market participation remains low in many regions, driven by factors such as perceived risk, limited financial literacy, and a lack of reliable forecasting tools. Machine learning (ML)-based predictive

models offer promising solutions to these barriers by learning from complex, high-dimensional data and adapting to rapidly changing market conditions [2]. Unlike traditional statistical models, machine learning (ML) algorithms can evolve with new information, thereby improving their predictive capacity over time. This adaptability makes them particularly suitable for dynamic financial environments.

Accurate stock price prediction, however, depends not only on the choice of algorithm but also on the relevance and quality of the input features. Prior research has often focused on developed markets such as the NYSE or NASDAQ, and primarily used technical indicators or price history as predictive inputs [5]. This approach overlooks the broader economic variables that significantly affect market behavior in emerging markets. Moreover, few studies have examined the impact of macroeconomic indicators, such as commodity prices and foreign exchange (FX) rates, on stock performance within localized contexts.

To address this gap, the present study focuses on the Bombay Stock Exchange (BSE), one of Asia's largest and most sensitive financial markets, and evaluates how macroeconomic indicators, including gold, silver, crude oil prices, and the INR/USD exchange rate, affect market efficiency. These features were selected for their historical economic significance in the Indian context. For example, India's status as a major oil importer directly links global crude oil prices to domestic corporate costs and inflation, while fluctuations in the foreign exchange rate impact international trade and investment. Our empirical correlation analysis confirms the predictive relevance of these features, with gold prices in particular showing a strong positive correlation with BSE performance.

In this work, we apply and compare multiple machine learning models: Adaptive Boosting (AdaBoost), Support Vector Machine (SVM), Gradient Boosting, and Random Forest on BSE data to evaluate their predictive accuracy. Unlike much of the existing literature focused on global indices, this study provides a localized, data-driven forecasting framework tailored to an emerging market. Our results not only validate the value of macroeconomic features in financial prediction but also demonstrate the effectiveness of ensemble methods, particularly AdaBoost, in delivering robust and interpretable stock forecasts. This contribution aims to inform both academic research and practical investment decision-making in volatile and under-studied financial environments.

2. Literature Review

The following section provides a concise discussion of several studies focused on stock market prediction using machine learning algorithms. In Desokey, et al. [6], clustering algorithms were developed by combining a genetic algorithm with the k-means algorithm. The primary objectives were to enhance stock market prediction through improved clustering and to examine the impact of this hybrid approach on prediction outcomes. To evaluate the accuracy of the clustering method, Chi-Square similarity measures and the sum of squared distances (SSD) were used.

Using sentiment characteristics and Latent Dirichlet allocation (LDA) features, Wang, et al. [7] carefully select official accounts from Sina Weibo, which is one of the major online social networks in China. In the research conducted, the authors evaluated the news content collected from these accounts. The RNN boost model, used to forecast stock volatility in the Chinese stock market, was developed by combining specific characteristics with technical indicators. Based on the experimental findings presented in the cited research, it appears that the proposed model is superior to other methods often employed and can make accurate predictions.

He, et al. [8] offer a data mining technique that can be used for researching and predicting trends in stock prices. The three processes that make up the approach are partitioning, analysis, and prediction. The k-means clustering approach is used to segregate the time series data related to stock prices. A linear regression analysis is conducted after the data has been partitioned to investigate the patterns present within each cluster. After that, the results of the linear regression are applied to provide predictions regarding trends for windowed time series data. In addition, it is designed for stock trading, but it can also be utilized to analyze the trend of any time series, including economic data.

Nikou, et al. [9] determine the degree to which machine learning algorithms can anticipate the stock market accurately. All daily closing price information from the iShares MSCI UK listed funds was used, beginning in January 2015 and continuing through June 2018. In the process of predicting, there are four different kinds of machine learning algorithms taken into consideration. When it comes to making accurate predictions, deep learning is superior to other methods, with support vector regression coming in second, producing fewer errors than neural networks and random forests.

Gurav and Sidnal [10] investigate several approaches to stock market analysis, mathematical models, the algorithms that are utilized the most frequently in the data science business, a review of several different types of machine learning algorithms, and a description of potential solutions. The purpose of this article is to study a range of aspects of dynamic stock market prediction. The essay is based on the concept that the reduction of risk associated with stock market investment is strongly connected to the reduction of inaccuracy in forecasting.

Machine learning techniques have been utilized in the stock trading industry to forecast stock price increases and decreases before the actual occurrence of a rise or fall in stock price. Shah [11] discusses Support Vector Machines, Linear Regression, Decision Stumps Prediction, Expert Weighting, and Online Learning, along with the advantages and disadvantages of each method. Additionally, the authors discuss how to implement each methodology. Additionally discussed are the aspects and variables that can be utilized to identify patterns in stock prices that may prove beneficial in the future for stock forecasting, as well as how Boosting can be combined with other learning algorithms to enhance the precision of such systems.

Sismanoglu, et al. [12] attempted to predict the future value of a specific stock by analyzing primary

financial data related to the market's exchange processes. Furthermore, a methodology was proposed for training and developing a knowledge base on a certain store, such as "IBM," by utilizing machine learning technologies. Stock time series values are obtained by the authors from the New York Stock Exchange. These numbers cover the period from 1968 to 2018. According to the results of their investigation, this approach yields exceptionally accurate stock forecasts.

Huang, et al. [13] took advantage of the SVM technique to forecast the direction of the weekly movement to investigate the degree to which the direction of financial movement may be predicted. A comparison is made between the performance of the SVM algorithm and that of Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Elman Backpropagation Neural Networks, as utilized by researchers, to evaluate the accuracy of the SVM method's prediction capabilities. According to the findings, SVM outperforms the other categorization methods. In addition, you should propose a combining model that makes use of SVM in conjunction with the other classification algorithms. This combined model performs the best out of all the other methodologies for prediction.

A study by Jader and Aminifar [14] discuss technical indicators for the stock market, mathematical models, the algorithms most commonly used in the data science industry, various types of machine learning algorithms, and a summary of potential solutions. Additionally, they aim to investigate several issues associated with dynamic stock market prediction. This concept is based on the idea that reducing the risk of investing in the stock market is closely tied to minimizing the number of forecasting errors.

In recent years, deep learning architecture such as Long Short-Term Memory (LSTM) networks and attention mechanisms have gained prominence in financial forecasting due to their ability to model temporal dependencies and focus on relevant inputs within sequential data. LSTM models, an extension of Recurrent Neural Networks (RNN), are particularly effective for time series analysis because they can retain long-term dependencies and mitigate the vanishing gradient problem. For instance, Chen, et al. [15] applied an LSTM-based model to stock price prediction and found it to outperform traditional ML models like SVM and Random Forest in terms of accuracy and temporal pattern recognition.

Moreover, the integration of attention mechanisms with LSTM networks has shown further improvements in performance by dynamically weighting the importance of different time steps or features. This approach helps the model "focus" on the most relevant parts of the input sequence when making predictions. Studies such as Saberikia, et al. [16] have demonstrated that attention-enhanced LSTM models can more effectively capture nonlinear relationships and market signals, leading to higher predictive accuracy and improved robustness across volatile financial datasets.

These advanced models, such as LSTM networks and attention-based architectures, offer a promising foundation for improving stock market forecasting. Their strength lies in modeling complex temporal relationships and dynamically focusing on the most relevant input features, which is particularly valuable in

financial time series prediction. LSTM networks are effective at capturing long-term dependencies, while attention mechanisms enhance the model's ability to weigh important patterns across time.

Although our current study emphasizes traditional machine learning models, such as AdaBoost and SVM, due to their simplicity, interpretability, and strong performance on structured data, these models may not fully capture the temporal dynamics, and sequential patterns often present in stock price movements. They are effective in static classification tasks but cannot learn from the order and timing of events—an important factor in financial time series forecasting.

As a result, future work could enhance this framework by incorporating models such as LSTM networks and attention-based architecture, which are specifically designed to handle sequential data and long-term dependencies. These models can better adapt to shifting market conditions and uncover deeper patterns across time, improving prediction accuracy. Such approaches are particularly promising for more complex and volatile datasets, and they may provide richer insights into market behavior, ultimately enabling more adaptive and responsive stock prediction systems.

Table 1. Summary table for literature review

Study	ML Method	Dataset	Key Features	Outcome
Nikou et al. (2019)	LSTM, SVM, RF	iShares MSCI UK ETF (2015–2018)	Historical prices	LSTM outperformed all with lowest RMSE
Gurav & Sidnal (2018)	SVM, Decision Trees	Indian stock market	Technical indicators	SVM performed best among tested methods
Chen et al. (2018)	RNN-Boost	Sina Weibo + Chinese market data	News sentiment, tech indicators	RNN-Boost showed superior accuracy
He et al. (2007)	K-means + Regression	US stock market	Time-windowed stock clusters	Improved trend prediction accuracy
Our Study	AdaBoost, SVM, RF, GB	BSE (2015–2017)	Gold, Oil, FX rates, Market history	AdaBoost highest accuracy (78.9%)

3. Background

In this segment of the research, four distinct machine-learning algorithms are examined and assessed based on the accuracy and reliability of their outcomes, as presented in Jader and Aminifar [14] and Jader, et al. [17]. The fundamental aim of machine learning is to enable systems to extract patterns and insights from previously collected data without requiring explicit programming or manual intervention. A key aspect of this learning process is the iterative application of models to data sets, through which the algorithms gradually improve their performance by identifying underlying patterns. This iterative process allows the models to generalize and adapt to new, unseen data independently rather than relying solely on prior examples.

Central to effective machine learning is the principle of autonomous learning, where the system self-directs its learning process by recognizing and internalizing patterns from data. This typically involves making informed predictions or decisions about future data points based on historical trends and relationships observed in the training data [18].

When determining the most suitable machine learning strategy for a given task, one must often choose between supervised and unsupervised learning approaches. This choice largely depends on the nature of the data, specifically whether labeled data with known outcomes (ground truth) is available. Supervised learning is generally used when such labeled data exists, whereas unsupervised learning is applied in scenarios where the data lacks predefined categories or labels.

More concretely, machine learning problems commonly fall into three broad categories: classification, regression, and clustering [19]. Classification involves predicting categorical labels, regression deals with forecasting continuous values, and clustering focuses on grouping similar data points without prior knowledge of the categories. These foundational problem types guide the selection and application of appropriate machine-learning techniques throughout the research [18].

The training of machine learning algorithms often involves the use of the three most frequent methods. There are three types of learning: supervised learning, unsupervised learning, and reinforcement learning. The most fundamental form of machine learning is the first one, which is called supervised learning. Here, the machine learning algorithm is trained on labeled data. Even though the data must be adequately labeled for this method to function, supervised learning is quite efficient when it is applied under the appropriate conditions. Within the context of supervised learning, the machine learning algorithm is provided with a miniature training dataset. A smaller portion of the larger dataset comprises this training dataset. It contributes to the algorithm's ability to acquire a fundamental comprehension of the problem, the solution, and the data points that need to be administered. Moreover, the second advantage of unsupervised machine learning is that it can handle data that has not been labeled. This is a particularly useful feature. Instead of relying on a predetermined and fixed problem statement, it can adapt to the data by continuously updating hidden structures. This enables more post-deployment development than supervised learning techniques do, particularly when compared to the former.

Additionally, the third one is known as reinforcement learning, which is directly influenced by the way individuals learn from data in their everyday lives. It is equipped with an algorithm that is capable of self-improvement and can learn from new situations through a process of trial and error. To implement reinforcement learning, the algorithm is placed in a working environment that includes an interpreter and a reward system. This approach, which is based on the psychological concept of conditioning, is effective. The interpreter receives the output result of each iteration of the algorithm and then evaluates the outcome to determine whether or not it is helpful. Four different machine learning algorithms are considered and associated with one another based on the precision of their results [9, 20].

3.1. Support Vector Machines

The Support Vector Machine, also known as SVM, is a method of supervised machine learning that can be applied to problems such as regression and classification. The SVM algorithm is founded on the

statistical learning theory [14]. Additionally, if you can apply SVM for recognizing undetected samples. There is a range of dimensions that are used to represent the data points. The SVM accomplishes the categorization in Figure 1; this hyperplane is responsible for dividing the various labels and generating them in a space featuring several dimensions, making it the most accurate and effective method for categorization.

Given the gravity of the circumstance, it is necessary to reduce the vector as much as possible. It does this by establishing a distinct decision boundary, which is analogous to the process of maximizing or optimizing the margin spaces that exist between two classes or objects. This is done to obtain the response that is optimal for the objective function. As presented in Figure 1.

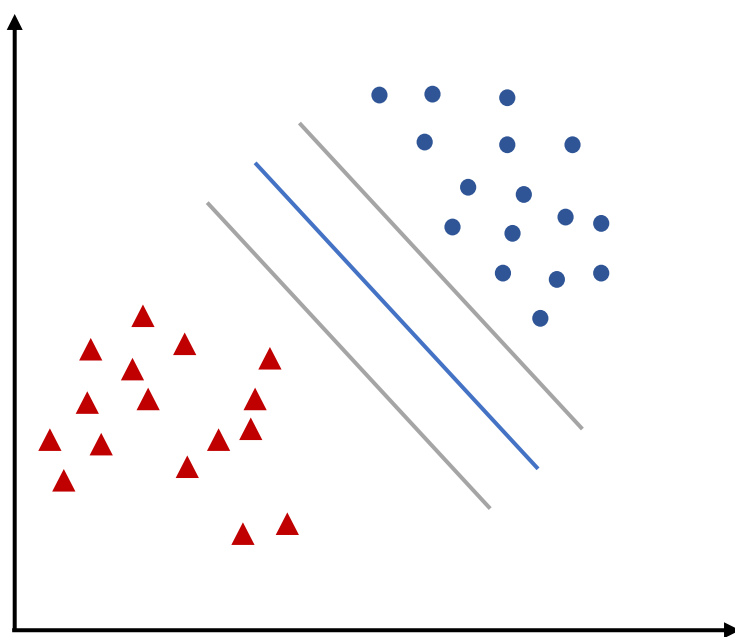


Figure 1. Support vector machine algorithm

SVM aims to find the optimal separating hyperplane that maximizes the margin between two classes, given a set of training data points.

$(x_i, y_i), i = 1, \dots, n, y_i \in \{-1, +1\}, \text{ and } x_i \in R^n$ (1)

The goal is to find a hyperplane defined as:

$$w^T x + b = 0 \quad (2)$$

where w is the weight vector, and b is the bias.

3.2. Random Forest

The random forest method has proven to be highly successful as a standard target regression and classification method. In settings where the number of variables is much higher than the number of observations [21], the random forest approach, which merges several randomized decision trees and accumulates their forecasts by averaging, as illustrated in the overall structure of random forest in Figure 2, has demonstrated remarkable success. Furthermore, it is capable of handling large computing issues, is adaptable to a

variety of single-task learning challenges, and provides customizable significance measurements [22].

Additionally, we can say that Random Forest is a machine learning technique that addresses classification and regression issues using a combination of strategies. Decision Tree and Bagging or Bootstrap Combination are used in this technique. As a result, it solves the problem of decision tree overfitting. The diversity of high-inconsistency algorithms, such as decision trees, is reduced by bagging. A Random Forest is a collection of unprocessed categorization and recession trees derived from divisions of a useful data set [23].

The processes in this algorithm are as follows: The CART model is employed in each of the sub-samples to produce the expected output based on the model, and samples of the training dataset have been created [24]. To create a mixture of each model's projected outputs, the bagging or categorizing procedure is utilized. As a result, the anticipated output using these processes has less prediction error than any individual model [25].

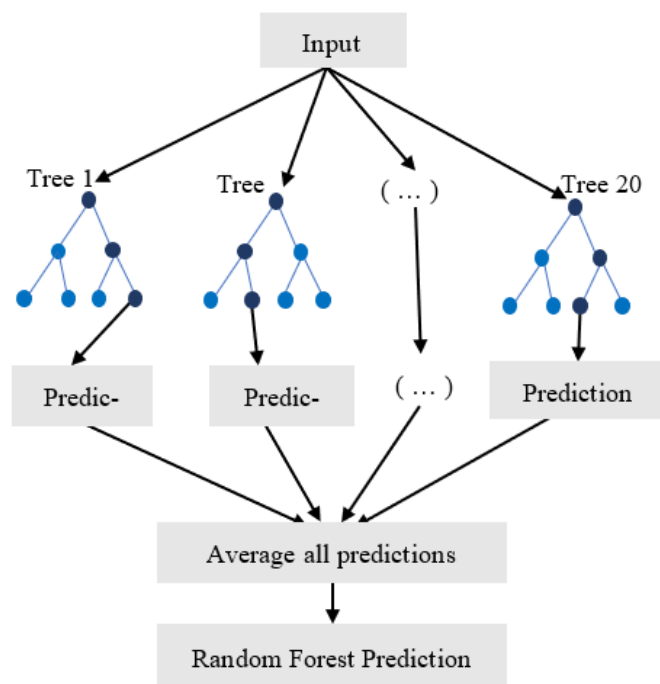


Figure 2. The general structure of a random forest method

3.3. Gradient Boosting

Gradient boosting methods are powerful machine learning techniques that have demonstrated considerable success in a wide range of practical applications. They are extremely configurable to meet the application's specific requirements, such as being trained on various loss functions.

Gradient Boosting is a type of gradient or increasing method by derivation that teaches patterns sequentially. The cost purpose is regularly reduced or optimized with each new model, and the model error will approach zero, as shown in Figures 3 and 4. On the boosting gradient, imagine a uniform distribution, for example, of $m1$, which is $\frac{1}{n}$ for all n observations. Next, the process operation according to these stages [25]:

Step 1: Select the original opinion.

Step 2: Compute the pseudo residuals.

Step 3: Predict the pseudo residuals.

Step 4: Create a prediction and calculate the residuals.

Step 5: Get a second prediction.

Step 6: Next Iteration.

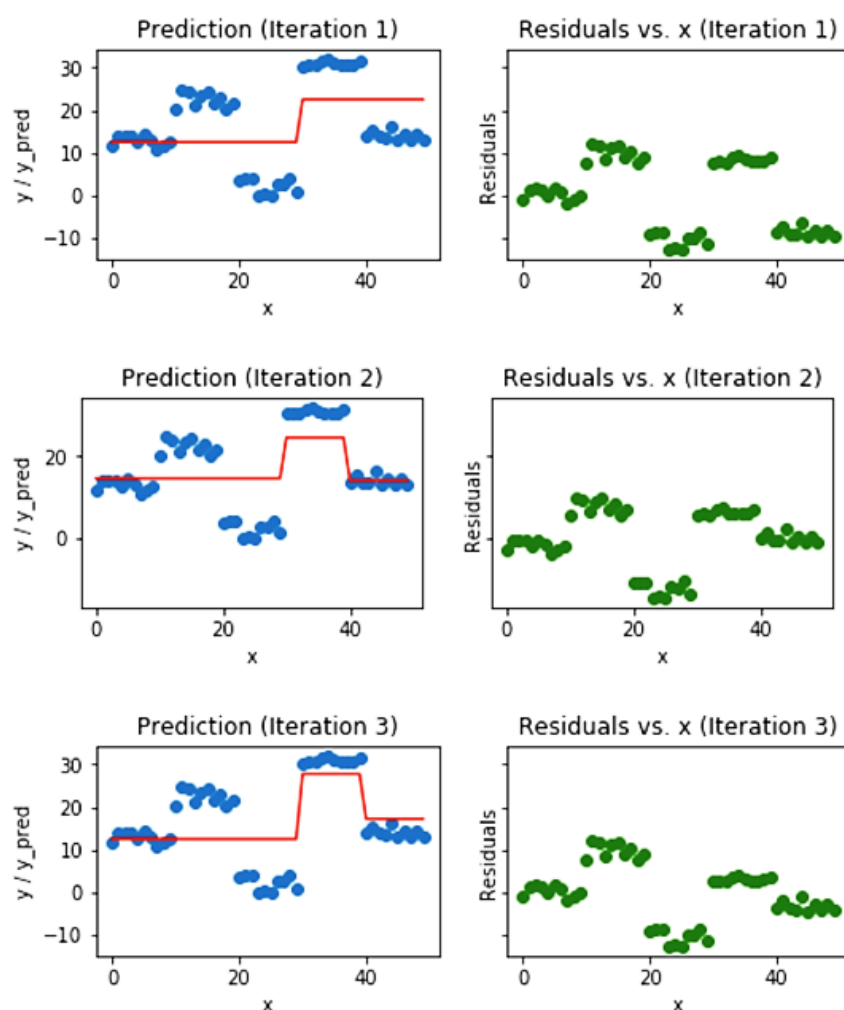


Figure 3. Visualization of gradient boosting prediction (First 4 Iteration)

3.4. AdaBoost and RNN Techniques

Recurrent neural networks (RNN) are the character of the design algorithm for succeeding data and are used by two big companies, Google for Google's voice search and Apple for Apple's Siri. It is one of the first algorithms that retains all its input data, which enables it to perform well for machine learning problems that require sequential information.

RNN is often used for the time series prediction tasks. It is regarded as a recurrent method because it performs the same task for every part in a series, and the current result or output depends on the past calculations [15].

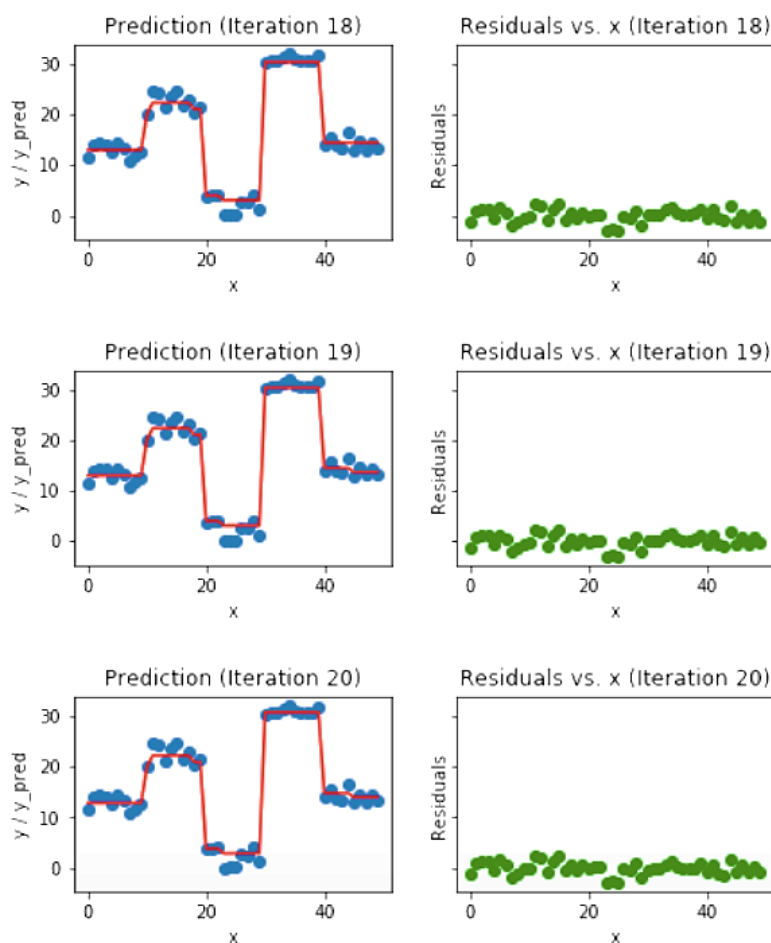


Figure 4. Visualization of gradient boosting prediction (18th to 20th Iteration)

Adaboost, also known as adaptive boosting, is one of the most effective strategies. To begin, regression and classification are two areas that can both benefit from the application of robust boosting techniques. The result is a powerful classifier that is created by combining some comparatively weak classifiers. The goal of adaptive boosting is to make accurate predictions based on the current data obtained by assigning an equal weight to each feature. A greater amount of weight is given to inaccurate observation if the data forecast is incorrect. Until there are no more mistakes, the iteration will continue continuously. The principal forecast is generated through the application of a fundamental algorithm. It is possible to anticipate the result by determining the weighted mean of any classifier. This is because the weights are generated based on incorrect earlier predictions and then revised later. One example of a model that projected a price based on the impact of social media news and previous stock data is presented in Figure 6.

4. Methodology

Before model training, the dataset underwent several preprocessing steps to enhance consistency and optimize learning performance. A key part of this process was normalizing numerical features to eliminate the impact of differing value scales across variables, such as commodity prices, foreign exchange rates, and historical BSE data.

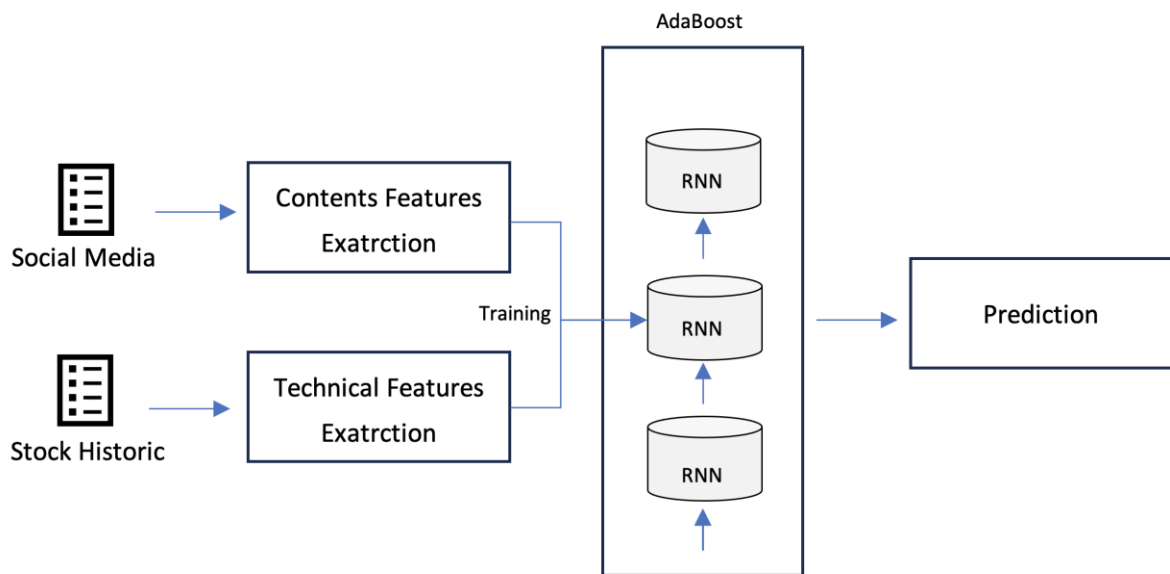


Figure 6. Prediction model created by AdaBoost vs. RNN technique

Two normalization techniques were applied:

a) Min-Max Normalization

For models sensitive to input scales, we applied min-max normalization to transform each feature into a bounded range of $[-3, 3]$. This was calculated as:

$$x' = \left(\frac{x - x_{min}}{x_{max} - x_{min}} \right) \times 6 - 3 \quad (3)$$

This transformation preserved the distribution of the original data while rescaling it to a fixed interval, improving convergence during training.

b) Z-Score Normalization (Standardization)

In a separate preprocessing configuration, especially for models like SVM and Gradient Boosting, we also tested z-score standardization, which rescales data based on the mean and standard deviation:

$$z = \left(\frac{x - \mu}{\sigma} \right) \quad (4)$$

This method centers the data around a mean of 0 with a standard deviation of 1, which can be beneficial when features exhibit different variances.

The choice between normalization methods was guided by empirical performance during model tuning. Both techniques helped reduce the dominance of high-scale variables and ensured numerical stability in training. The dataset was then split chronologically, maintaining the temporal structure of the data to prevent data leakage and preserve predictive integrity. Two train-test split configurations were used: one with 90% of the data for training and 10% for testing and another with 70% for training and 30% for testing, enabling robustness checks across varying data proportions.

The machine learning models, Random Forest, Support Vector Machine (SVM), Gradient Boosting,

and Adaptive Boosting were implemented using Python and relevant open-source libraries (e.g., Scikit-learn). Each model was trained using the normalized feature set, and hyperparameters were tuned empirically. Classification output was binary: a predicted label of +1 for an upward trend and -1 for a downward trend in stock prices. Model performance was evaluated using accuracy metrics across both training and test sets. Among the tested methods, Adaptive Boosting consistently demonstrated superior accuracy, achieving 78.9% with 90% training data and 76.7% with 70% training data.

4.1. Feature Selection and Justification

The selection of input features was guided by both domain knowledge and empirical correlation analysis, to improve prediction accuracy while maintaining model interpretability. The features incorporated into the machine learning models include:

- **Commodity Prices:** Gold, silver, and crude oil prices were selected due to their well-established impact on market sentiment, inflation, and industrial costs, particularly in energy-dependent economies such as India.
- **Foreign Exchange Rate (INR/USD):** The exchange rate was included as it directly affects multinational companies, import-export dynamics, and foreign investment inflows, which in turn influence stock prices.
- **Historical Stock Data:** Key indicators such as daily opening, high, low, and closing prices were used to capture short-term market momentum and volatility trends.

Additionally, we conducted Pearson correlation and covariance analyses to empirically assess the relationship between each feature and market movement (BSE trend). Results are presented in Table 3, where gold prices showed the highest positive correlation (+0.75) with BSE trends, followed by crude oil and exchange rates, which exhibited negative correlations consistent with expectations for emerging markets. By combining expert-guided selection with quantitative correlation checks, we ensured that only the most relevant and non-redundant features were included, thus improving model efficiency and reducing the risk of overfitting.

To predict the Bombay Stock Exchange, the primary objective of this article is to make use of several characteristics as data features. These characteristics include the values of commodities, the market's history, and the foreign exchange rate (BSE). Because these input attributes were continuous values spanning a variety of ranges, they were normalized within the range of [-3] and [3] to facilitate the classification process. Each of the input characteristics can have both positive and negative values, this is the situation. The purpose of this study is to examine and contrast the results obtained by each of the machine learning algorithms that were presented in the article. To reflect the negative or positive impact on the stock market, the outcome derived from each model is either a value of 1 or a negative value of 1. The representation of the suggested model is shown in Figure 7.

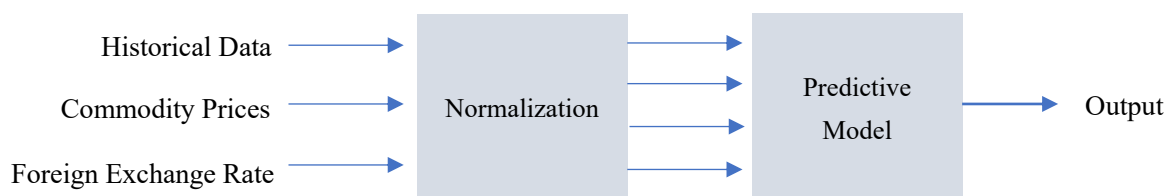


Figure 7. The proposed model

4.2. Effectors Element on Market

4.2.1. History of Market

Historical information for two years from 2015 - Jan to 2016 - Dec was collected. This historical information was not directly used in the prediction model. After using the model, it retains historical information that was used for training data, such as the effectors' elements, including changes in the opening market price, small amounts, changes in stock sizes over the period, and high prices.

4.2.2. Commodity Price

The price difference of several features like silver, gold, and crude oil directly affects the total change of the stock prices of the market.

4.2.3. Foreign Exchange

Foreign exchange plays a significant role in determining market efficiency, especially in emerging economies where currency fluctuations can directly impact trade, investment, and corporate earnings. In our study, foreign exchange, along with the other selected influencing factors, is used as a key input feature in the predictive model. All these variables are collectively processed to generate a binary output representing market movement: $[+1]$ for a positive price trend and $[-1]$ for a negative price trend. Specifically, the INR/USD exchange rate is included as a continuous input across all observations, reflecting its influence on market volatility and investor sentiment. The relationship between the input features and the resulting market prediction is visualized through the variation patterns illustrated in Figure 8.

4.3. Implement of Each Algorithm

Data of various qualities were collected and modified to be used as input data for the model. The output of this model is $+1$. This production depicts a positive market, indicating an increase in BSE prices, as well as a negative market, indicating a decline in BSE prices. Four alternative algorithms were used in the simulation, and the precision of each didactic and experimental separation approach was determined. In this method, the data is split into training and test data. The training set contains the output material, and the model learns from it so that it may be extended to new data in the future.

The applied machine learning methods were successfully implemented on a dataset downloaded from Data of 9 months from Jan 2017 to Sept 2017 in efficiency or accuracy for each model computed by various

data for training and testing dataset. It was obtained using the random forest method, with an accuracy of 68.4% on the training set and 73.0% on the test set. For the support vector machine, the obtained accuracy was 78.9% for 90% training and 73% for 70% training. The accuracy obtained by the gradient boosting method was %78.9 for % 90 training and %73.2 for 70% training. Finally, for the adaptive boosting method, the obtained accuracy was 78.9% on the training data and 77% on the test data. Also, as shown in Table 2.

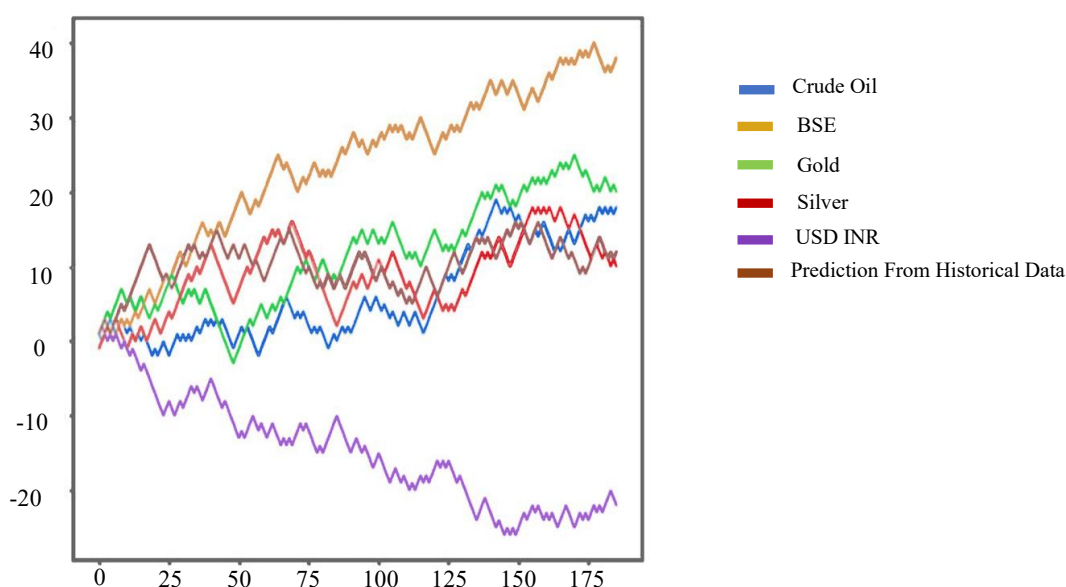


Figure 8. Variation between input and output features

Table 2. Accuracy of each machine learning algorithms based on different rates of training and test data

Machine learning algorithms	Dataset used for holdout data		
	(%) Training set	(%) Test set	(%) Accuracy
Adaptive Boosting	90	10	78.9
	70	30	76.7
Gradient Boosting	90	10	78.9
	70	30	73.2
Support Vector Machine	90	10	78.9
	70	30	73
Random Forest	90	10	68.4
	70	30	67.8

The study calculated the association and covariance between the various characteristics and market efficiency. Correlation and covariance are used to characterize the feature's resultant dependence, which is shown in Table 3.

We employed Grid Search with 5-fold cross-validation for each model. The following parameter ranges were explored:

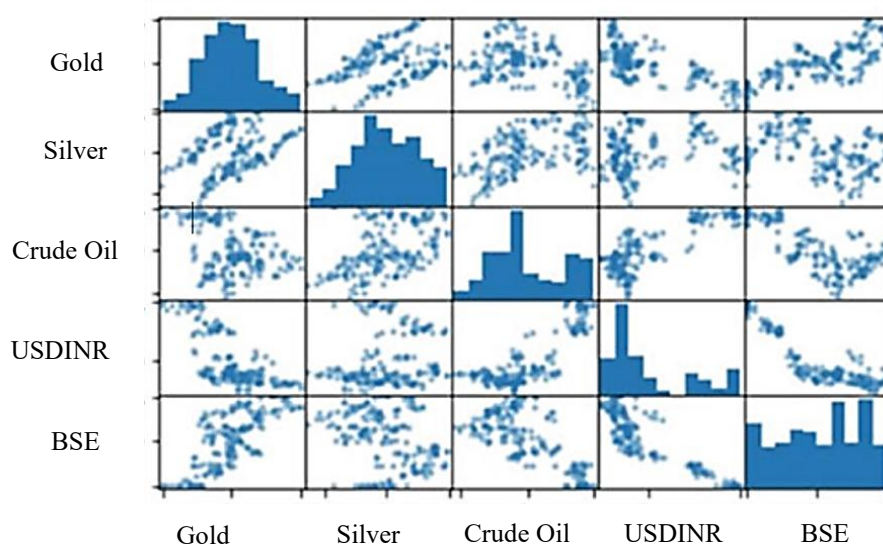
- **AdaBoost:** $n_estimators = [10, 50, 100]$, $learning_rate = [0.01, 0.1, 1.0]$
- **Gradient Boosting:** $n_estimators = [50, 100]$, $max_depth = [3, 5]$, $learning_rate = [0.05, 0.1]$
- **SVM:** $C = [0.1, 1, 10]$, $kernel = ['rbf', 'linear']$, $gamma = ['scale', 'auto']$

- **Random Forest:** $n_estimators = [50, 100]$, $max_depth = [None, 10, 20]$

Table 3. Correlation and covariance of different features and market efficiency

S. No	Features	Correlation	Covariance
1	Oil Rate	-0.7 451	-102579.41
2	Gold Rate	0.75 293	16781.41
3	Silver Rate	-0.2 866	-114.30
4	Foreign Exchange Rate	-0.8 961	-698.39

For a deeper understanding of the correlation between features used in models, as shown in Figure 9, a scatter plot illustrates the correlation between various features to identify the structured relationship between multiple features and market efficiency.

**Figure 8.** Variation between input and output features

4.4. Algorithm Implementation Details

To optimize model performance and avoid underfitting or overfitting, we conducted a comprehensive grid search using 5-fold cross-validation for each machine learning algorithm. The grid explored combinations of parameters based on standard guidelines and prior work in financial time series prediction. The following outlines the reasoning for each model's parameter choices:

Random Forest Classifier:

- $n_estimators=100$
- $max_depth=None$
- $criterion='gini'$
- $random_state=42$

Support Vector Machine (SVM):

- Kernel: 'rbf'

- $C=1.0$
- $\text{gamma}=\text{'scale'}$

Gradient Boosting Classifier:

- $n_estimators=100$
- $learning_rate=0.1$
- $max_depth=3$
- $subsample=1.0$

AdaBoost Classifier:

- $n_estimators=50$
- Base estimator: `DecisionTreeClassifier(max_depth=1)`
- $learning_rate=1.0$

All hyperparameters were selected through a combination of empirical tuning and a systematic grid search strategy, conducted within a 5-fold cross-validation framework on the training data. This approach ensured that the models were not only optimized for performance but also generalizable to unseen data. Grid search allowed us to evaluate multiple combinations of parameters for each algorithm, while cross-validation provided a robust estimate of model stability and helped prevent overfitting. The final parameter configurations were chosen based on the best-performing results across folds, prioritizing metrics such as accuracy, precision, recall, and F1-score. These evaluation metrics offer a more comprehensive understanding of model performance, especially in binary classification tasks involving imbalanced or noisy financial data. The complete set of performance results, including the selected parameters and their associated metrics, is presented in Table 4 for comparative analysis:

Table 4. Accuracy, precision, recall and F1-Score metrics for our models

Model	Accuracy	Precision	Recall	F1-Score
AdaBoost (90/10)	78.9%	0.80	0.78	0.79
SVM (90/10)	78.9%	0.77	0.76	0.76
Gradient Boosting	78.9%	0.78	0.77	0.77
Random Forest	68.4%	0.69	0.67	0.68

5. Discussion

The results of our experiments, as summarized in Table 2 and Table 4, demonstrate that Adaptive Boosting (AdaBoost) consistently outperformed other machine learning models in terms of accuracy and F1-score. This finding aligns with previous research by Gurav and Sidnal (2018), who found AdaBoost and SVM to be the most effective classifiers for financial prediction tasks involving technical indicators. Similarly, our results reinforce the claims made by Chen et al. (2018) that ensemble methods can outperform traditional

classifiers due to their ability to capture non-linear patterns in volatile datasets.

AdaBoost's superior performance in our study can be attributed to its iterative weighting mechanism, which emphasizes misclassified instances during training. This enables the model to progressively focus on more challenging cases, thereby improving the classification boundaries. In contrast to models like Random Forest or SVM, which treat all instances with equal importance, AdaBoost adapts its learning process to the underlying data distribution, making it particularly effective in noisy or imbalanced datasets such as financial time series.

Another reason for AdaBoost's effectiveness lies in its combination of multiple weak learners, typically shallow decision trees, which individually may have limited predictive power but collectively produce a robust classifier. This ensemble effect helps reduce both bias and variance, enabling the model to generalize better to unseen data.

From a real-world perspective, the robustness and high accuracy of AdaBoost suggest its applicability in automated trading systems, risk assessment frameworks, and portfolio optimization tools, where quick and reliable directional market predictions are crucial. Its ability to incorporate a variety of financial indicators (such as commodity prices and FX rates) and adapt to changing market conditions makes it suitable for deployment in emerging markets, where data may be more volatile and less structured compared to mature financial environments.

Furthermore, our correlation analysis revealed that gold prices had the strongest positive association with BSE performance. This observation is consistent with a study by Nikou et al. (2019), which also highlighted the predictive importance of commodity indicators in emerging economies. By integrating economically relevant variables into our AdaBoost framework, we not only improved model performance but also enhanced interpretability, offering insights that can inform both algorithmic strategies and human decision-making.

In future work, integrating adaptive boosting with deep learning models, such as LSTM networks or hybrid attention-based systems, may further enhance the model's temporal sensitivity and predictive capabilities. This could bridge the gap between explainability and complexity in financial forecasting models.

6. Conclusion

This study demonstrated that among the evaluated machine learning models, Adaptive Boosting (AdaBoost) consistently delivered the most accurate stock market predictions, as reflected in its superior accuracy, precision, and F1-score metrics. The strong performance of AdaBoost underscores the effectiveness of ensemble methods in capturing complex, non-linear relationships within financial time series data. By incorporating a diverse set of macroeconomic and market-related indicators, the analysis revealed that gold prices exhibited the strongest positive correlation with stock market efficiency, highlighting their value

as a predictive input, particularly in emerging markets like India.

These findings emphasize that both the selection of an appropriate machine learning algorithm and the inclusion of economically relevant features are critical to achieving high forecasting accuracy. The integration of robust ensemble models, such as AdaBoost, with carefully engineered feature sets significantly enhances the performance and reliability of predictive systems. This research offers practical insights and guidance for financial analysts, institutional traders, and investors aiming to leverage machine learning techniques for more informed, adaptive decision-making in dynamic and volatile market environments.

Funding: The authors declare that no funding was received for this research.

Declaration of Competing Interest: The authors declare they have no known competing interests.

Research Involving Human/Animal Participants: This article does not involve any studies conducted by authors on animals or human participants.

References

- [1] B. Jan *et al.*, "Deep learning in big data analytics: a comparative study," *Computers & Electrical Engineering*, vol. 75, pp. 275-287, 2019, doi: <http://dx.doi.org/10.1016/j.compeleceng.2017.12.009>.
- [2] N. L. Rane, M. Paramesha, S. P. Choudhary, and J. Rane, "Machine learning and deep learning for big data analytics: A review of methods and applications," *Partners Universal International Innovation Journal*, vol. 2, no. 3, pp. 172-197, 2024, doi: <http://dx.doi.org/10.2139/ssrn.4835655>.
- [3] T. Portovaras, V. Kyrychenko, T. Tsepikalo, O. Sokulskyi, and I. Krysovaty, "The role of modern information technologies in financial analysis and market forecasting," *Economic Affairs*, vol. 69, no. 1, pp. 735-742, 2024, doi: <http://dx.doi.org/10.46852/0424-2513.2.2024.33>.
- [4] A. Hailemariam and C. Guotai, "Stock market development and economic growth: Empirical evidence for emerging market economies," *International Journal of Economics, Finance and Management Sciences*, vol. 2, no. 2, pp. 171-181, 2014, doi: <http://dx.doi.org/10.11648/j.ijefm.20140202.19>.
- [5] N. Rouf *et al.*, "Stock market prediction using machine learning techniques: a decade survey on methodologies, recent developments, and future directions," *Electronics*, vol. 10, no. 21, p. 2717, 2021, doi: <http://dx.doi.org/10.3390/electronics10212717>.
- [6] E. N. Desokey, A. Badr, and A. F. Hegazy, "Enhancing stock prediction clustering using K-means with genetic algorithm," in *2017 13th international computer engineering conference (ICENCO)*, 2017: IEEE, pp. 256-261, doi: <http://dx.doi.org/10.1109/ICENCO.2017.8289797>.
- [7] J. Wang, Y. Zhou, W. Zhang, R. Evans, and C. Zhu, "Concerns expressed by Chinese social media users during the COVID-19 pandemic: content analysis of Sina Weibo microblogging data," *Journal of medical Internet research*, vol. 22, no. 11, p. e22152, 2020, doi: <http://dx.doi.org/10.2196/22152>.
- [8] H. He, J. Chen, H. Jin, and S.-H. Chen, "Trading strategies based on K-means clustering and regression models," in *Computational Intelligence in Economics and Finance: Volume II*: Springer, 2007, pp. 123-134.
- [9] M. Nikou, G. Mansourfar, and J. Bagherzadeh, "Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms," *Intelligent Systems in Accounting, Finance and Management*, vol. 26, no. 4, pp. 164-174, 2019, doi: <http://dx.doi.org/10.1002/isaf.1459>.

- [10] U. Gurav and N. Sidnal, "Predict stock market behavior: role of machine learning algorithms," in *Intelligent Computing and Information and Communication: Proceedings of 2nd International Conference, ICICC 2017*, 2018: Springer, pp. 383-394, doi: http://dx.doi.org/10.1007/978-981-10-7245-1_38.
- [11] V. H. Shah, "Machine learning techniques for stock prediction," *Foundations of Machine Learning| Spring*, vol. 1, no. 1, pp. 6-12, 2007.
- [12] G. Sismanoglu, M. A. Onde, F. Kocer, and O. K. Sahingoz, "Deep learning based forecasting in stock market with big data analytics," in *2019 scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT)*, 2019: IEEE, pp. 1-4, doi: <http://dx.doi.org/10.1109/EBBT.2019.8741818>.
- [13] W. Huang, Y. Nakamori, and S.-Y. Wang, "Forecasting stock market movement direction with support vector machine," *Computers & operations research*, vol. 32, no. 10, pp. 2513-2522, 2005, doi: <http://dx.doi.org/10.1016/j.cor.2004.03.016>.
- [14] R. Jader and S. Aminifar, "Fast and accurate artificial neural network model for diabetes recognition," *NeuroQuantology*, vol. 20, no. 10, pp. 2187-2196, 2022.
- [15] W. Chen, C. K. Yeo, C. T. Lau, and B. S. Lee, "Leveraging social media news to predict stock index movement using RNN-boost," *Data & Knowledge Engineering*, vol. 118, pp. 14-24, 2018, doi: <http://dx.doi.org/10.1016/j.datak.2018.08.003>.
- [16] M. Saberikia, H. Beitollahi, R. Jader, and H. Farbeh, "Enhancing Performance in Mixed-Criticality Real-Time Systems Through Learner-Based Resource Management," *IET Cyber-Physical Systems: Theory & Applications*, vol. 10, no. 1, p. e70007, 2025, doi: <http://dx.doi.org/10.1049/cps2.70007>.
- [17] R. F. Jader, M. H. M. Abd, and I. H. Jumaa, "Signal modulation recognition system based on different signal noise rate using artificial intelligent approach," *Journal of Studies in Science and Engineering*, vol. 2, no. 4, pp. 37-49, 2022, doi: <http://dx.doi.org/10.53898/josse2022244>.
- [18] I. Kumar, K. Dogra, C. Utreja, and P. Yadav, "A comparative study of supervised machine learning algorithms for stock market trend prediction," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018: IEEE, pp. 1003-1007, doi: <http://dx.doi.org/10.1109/ICICCT.2018.8473214>.
- [19] K. Pesek, *Atherosclerotic cardiovascular disease*. IntechOpen, 2011.
- [20] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197-227, 2016, doi: <http://dx.doi.org/10.1007/s11749-016-0481-7>.
- [21] R. Jader and S. Aminifar, "An Intelligent Gestational Diabetes Mellitus Recognition System Using Machine Learning Algorithms," *Tikrit Journal of Pure Science*, vol. 28, no. 1, pp. 82-88, 2023.
- [22] M. R. Saberikia, H. Beitollai, R. Jader, and H. Farbeh, "Reliability and power consumption models of embedded systems with a machine learning-based optimization," *International Journal of Modeling, Simulation, and Scientific Computing*, p. 2541011, 2024, doi: <http://dx.doi.org/10.1142/S1793962325410119>.
- [23] A. Leach, C. Prout, and D. Dolata, "Applications of artificial intelligence techniques in conformational analysis," *Journal of Molecular Graphics*, vol. 6, no. 4, p. 215, 1988, doi: [http://dx.doi.org/10.1016/S0263-7855\(98\)80021-0](http://dx.doi.org/10.1016/S0263-7855(98)80021-0).
- [24] P. P. S. Kohli, S. Zargar, S. Arora, and P. Gupta, "Stock prediction using machine learning algorithms," in *Applications of Artificial Intelligence Techniques in Engineering: SIGMA 2018, Volume 1*, 2019: Springer, pp. 405-414, doi: http://dx.doi.org/10.1007/978-981-13-1819-1_38.
- [25] R. F. Jader, S. W. Kareem, and H. Q. Awla, "Ensemble deep learning technique for detecting MRI brain tumor," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, p. 6615468, 2024, doi: <http://dx.doi.org/10.1155/2024/6615468>.