Integration of Effective Models to Provide a Novel Method to Identify the Future Trend of Nikkei 225 Stocks Index

Jiang Zhu[1], Haiyan Wu[2]*
School of Finance, Jiangsu Vocational College of Finance & Economics, Huai'an 223003, Jiangsu, China[1]
School of Management-South China Business College, Guangdong University of Foreign Studies,
Guangzhou 510545, Guangdong, China[2]

Abstract*—The stock market refers to a financial market in which individuals and institutions engage in the buying and selling of shares of publicly listed firms. The valuation of stocks is influenced by the interplay between the forces of supply and demand. The act of allocating funds to the stock market entails a certain degree of risk, while it presents the possibility of substantial gains over an extended period. The task of predicting stock prices in the securities market is further complicated by the presence of non-stationary and non-linear characteristics in financial time series data. While traditional techniques have the potential to enhance the accuracy of forecasting, they are also associated with computational complexities that might lead to an elevated occurrence of prediction mistakes. This is the reason why the financial industry has seen a growing prevalence of novel methods, particularly in the stock market. This work introduces a novel model that effectively addresses several challenges by integrating the random forest methodology with the artificial bee colony algorithm. In the current study, the hybrid model demonstrated superior performance and effectiveness compared to the other models. The proposed model exhibited optimum performance and demonstrated a significant degree of effectiveness with low errors. The efficiency of the predictive model for stock price forecasts was established via the analysis of data obtained from the Nikkei 225 index. The data included the timeframe from January 2013 to December 2022. The results reveal that the proposed framework demonstrates efficacy and reliability in evaluating and predicting the price time series of equities. The empirical evidence suggests that, when compared to other current methodologies, the proposed model has a greater degree of accuracy in predicting outcomes.

***Keywords*—Financial market; stock future trend; Nikkei 225 index; random forest; artificial bee colony**

I. Introduction

Stocks are traded in a public market where companies list and raise capital by selling shares at set prices. The stability and security of the stock market are crucial for the functioning of national economies [1, 2]. The stock market is difficult to predict since there are a lot of variables at play. Stock performance is significantly impacted by political uncertainty that arises from political events [3], [4]. This connection often occurs as a result of investors' reactions to the policy uncertainty arising from the course of these events. Additionally, political developments and uncertainties influence investors' judgments about market timing and portfolio allocation across various markets.

Consequently, investors conduct two different kinds of research before purchasing a company. Fundamental analysis comes first. Investors consider things including the economy, industry performance, and the fundamental worth of equities. Second, investors use technical analysis to examine stock values and information produced by market activity, such as historical prices and volume [5]. Research on the conduct and functioning of stock markets have developed into essential due to the possible hazards involved [6]. Forecasting changes in stock prices is one of the most crucial responsibilities in this respect, as it provides investors with the knowledge they need to make wise decisions and minimize risks, in addition to aiding regulators in stabilizing financial markets.

Nonetheless, there can be serious concerns associated with inaccurate prediction outcomes and mysterious prediction methods [7]. The high volatility, non-stationarity, and non-linearity of stock price time series data further complicate the task of accurately forecasting stock prices in the securities market. Therefore, to reduce possible hazards, a trustworthy and persuasive prediction model must be created. It has long been used to forecast the stock market via the use of traditional techniques that involve studying fundamental and technical analysis. The ever-changing and dynamic character of the stock market presents challenges in conducting analysis.

The aforementioned attributes indicate that traditional statistical approaches may be inadequate in facilitating a comprehensive understanding of the stock market. Terms like artificial intelligence (AI) and machine learning (ML) might be confusing. The idea of artificial intelligence pertains to a computer system that can execute jobs that are normally performed by humans [8]. The notion of ML holds that computers can learn or make predictions using just their own experience and training without the need for outside programming [9]. This means that the system can make judgments based on information with little to no assistance from human [8]. Decision trees, first developed in the 1960s, are among the best techniques for data mining and are extensively used across many fields [10]. This is due to their simplicity, lack of ambiguity, and robustness, even when values are absent. It is possible to employ continuous or discrete variables as independent or target variables. Two approaches to analyzing decision trees can be taken when dealing with missing data: categorize missing values as a distinct category that can be examined with the other categories or use a pre-built decision tree model that sets the variable with many missing values as a target variable to make a prediction and replace these missing ones with the predicted value [11]. The random forest technique is used to reduce the overfitting risk that is often linked to the use of a single decision tree. Multiple decision trees are trained as part of the ensemble learning approach known as the random forest algorithm. The output of each tree is aggregated to determine the ultimate result of the algorithm [12]. Every decision tree produces its output on its own, and for regression tasks, the final prediction is derived from the average of the replies. The concept of random forests was first introduced by Breiman [13]. The model used by Oyebayo Ridwan Olaniran et al. [14] was utilized for high-dimensional categorization of genomic data. This model was used by Antoine Gatera et al. [15] to predict traffic accidents.

The integration and use of optimizers in conjunction with the chosen model have led to improvement in research outcomes, resulting in heightened accuracy of the obtained data. As a result, many optimization techniques, namely ant lion optimization (ALO) [16], grey wolf optimization (GWO) [17], battle royal optimizer (BRO) [18], mouth flame optimization (MFO) [19], Biogeography-based optimization (BBO) [20], Artificial bee colony (ABC) [21] have been presented. A mathematical model for resolving optimization issues that is based on studying and imitating the patterns of real bee foraging behavior is called the ABC algorithm. Three categories of honey bee agents work for the ABC in the colony: scouts, observers, and hired bees. In the ABC algorithm, there are two groups of bees with an equal number of bees each. One-half of them are called working bees, while the other half are called observer bees. The position of the food supply for the bee is seen as a solution in the ABC that needs its parameters optimized. The objective function of a problem, which is equivalent to the fitness value of the solution, is connected to the quality of the food supply. Put another way, locating the best answer is akin to the process of foraging used by bees to identify a quality food supply. The ABC algorithm's specifics are as follows. The first solutions are created at random and used by the bee agents as their food supply locations. Following initiation, the bee agents go through three main cycles of iterative changes: choosing viable solutions, updating the feasible solutions, and avoiding less-than-ideal solutions. The main contributions of the study are as follows:

- This research paper presents an innovative predictive model that combines the random forest and artificial bee colony algorithms. By capitalizing on sophisticated machine learning algorithms, the suggested model enhances the dependability of stock market forecasts through the reduction of error rates and the improvement of prediction accuracy.

- Due to the empirical validation of the proposed model using Nikkei 225 index data spanning a significant period of time, comparisons across various market conditions are limited. Through a comprehensive analysis of the model's performance across diverse dynamics, this study offers significant insights into its efficacy and resilience. This particular facet contributes to the overall comprehension of stock market prediction models based on machine learning and increases their practicality in real-life situations.

- The research's emphasis on assessing the vulnerabilities of models and rectifying shortcomings in interpretability, feature engineering, and external validation is of paramount importance in furthering the domain of stock market forecasting. Through a methodical examination of these deficiencies, the study makes a valuable contribution to enhancing the dependability and practicality of forecasting techniques that rely on machine learning.

The research assessed the reliability of various models, including RF, BRO-RF, and MFO-RF. The model selected for this article is ABC-RF recognized for its superior performance. Subsequently, the following section thoroughly examines all pertinent components of the investigation. Numerous analytical methods, including the RF model, assessment metrics, and optimizer approaches, were used to examine the data. The study's findings are presented and compared with those obtained using alternative methods in the third section. The last section offers a concise review of the findings of the research.

II. Literature Review

In recent times, there has been an increasing inclination towards utilizing machine learning algorithms for the purpose of forecasting stock market trends, with the intention of leveraging forthcoming price fluctuations and augmenting investor profitability. Agrawal [22] presented a stock market prediction system that employs nonlinear regression techniques based on deep learning. By conducting experiments on a wide array of datasets, such as ten years' worth of Tesla stock price data and data from the New York Stock Exchange, Agrawal establishes that the proposed method outperforms conventional machine learning approaches [22]. The methodology proposed by Petchiappan et al. [23] for forecasting the stock prices of media and entertainment companies substantially advanced this field of study. By employing machine-learning methodologies, particularly logistic and linear regression, they successfully construct a resilient prediction system tailored to the industry. Through a meticulous analysis of stock price data obtained from reputable media sources, their model provides investors with invaluable insights regarding profit optimization and loss mitigation. Petchiappan et al. [23] establish the effectiveness of their methodology by conducting extensive experiments, with a specific focus on its superiority in comparison to conventional approaches. Predicting stock market fluctuations remains a complex and demanding undertaking within the field of finance, owing to the ever-changing and multifaceted characteristics of stock prices. Sathyabama et al. [24] employ machine learning algorithms to forecast stock market transactions as a means of surmounting this obstacle. The research conducted by the authors’ places considerable emphasis on the impact that news and other external variables have on stock market trends. Moreover, this emphasizes the criticality of precise prognostic models in efficiently controlling market volatility. Sathyabama et al. [24] contribute to the existing body of knowledge by introducing an improved learning-based approach that integrates a Naïve Bayes classifier. Menaka et al. [25] made a scholarly contribution to this domain by conducting an exhaustive examination of machine learning algorithms that are employed in the prediction of stock prices across various stock exchanges. Menaka et al. [25] emphasized the adaptability of various machine-learning methodologies to construct accurate prediction models. These methodologies comprised boosted decision trees, support vector machines, ensemble methods, and random forests. To tackle the distinct obstacles presented by abrupt and capricious market fluctuations, Demirel et al. [26] directed their examination toward the companies comprising the Istanbul Stock Exchange National 100 Index. An assessment was made of the predictive performance of Support Vector Machines, Multilayer Perceptrons, and Long Short-Term Memory using daily data spanning a period of nine years [26]. The investigation of stock market forecasts continues to be substantial, given its extensive ramifications for international financial markets, shareholders, and enterprises. To tackle this obstacle, Tembhurney et al. [27] performed a comparative analysis of the performance of machine learning algorithms in forecasting the Nifty 50 stock market index. Tembhurney et al. [27] utilized the Python programming language to implement the Support Vector Machine and Random Forest algorithms for the purpose of training models with historical stock market data.

The effectiveness of machine learning algorithms in predicting stock market trends is illustrated in the literature review, which stands in contrast to traditional approaches. However, there are still notable shortcomings that persist. These include the lack of thorough examination of how models can be interpreted, the neglect of feature engineering, the dearth of external validation, and the inadequate evaluation of dynamic market conditions. Furthermore, there are limited comparisons made across different market conditions, and the assessment of model risks is insufficient. It is critical to address these deficiencies so as to enhance the reliability and applicability of machine learning-based stock market prediction models. Consequently, additional research is necessary to focus on the creation of models that are intuitive, robust, and adaptable, incorporating comprehensive risk assessment frameworks and capable of modifying to changing market conditions. This article focuses on the application of innovative approaches, particularly the combination of the random forest methodology and the artificial bee colony algorithm, to enhance the precision of stock market predictions in order to fill the deficiencies identified in the literature review. Additionally, the research improves the evaluation of model risks and addresses the scarcity of cross-market comparisons through an examination of Nikkei 225 index data spanning a significant period of time, from January 2013 to December 2022. In order to improve the dependability and practicality of machine learning-driven financial market forecasting, the objective of this study is to develop a stock market prediction model that is more intuitive, robust, and flexible.

III. Methods and Materials

A. Random Forest

Random forest regression is a kind of regression approach that is based on machine learning [28]. The RF algorithm, as described by Breiman [13], employs ensemble learning to enhance prediction accuracy by integrating multiple trees. This non-parametric data mining technique is capable of handling non-linear and non-additive relationships by utilizing recursive partitioning of the dataset to investigate the associations between a response variable and predictor variables, as highlighted by Wiesmeier et al. [29]. The fundamental constituent of RF is the decision tree, whereby the aggregation of numerous decision trees is used to mitigate the potential issue of over-fitting, as shown in Fig. 1. The training procedure for several decision trees is conducted in parallel, with each tree exhibiting modest variations owing to the incorporation of a random process inside the algorithm [30]. The RF technique additionally offers the projected significance of input parameters used in constructing the model.

The mean square error for an RF may be found using the following equation:

$$\mathrm{MSE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{k}=0}^{\mathrm{n}} \binom{\mathrm{n}}{\mathrm{k}} (\mathrm{Fi} - \mathrm{Yi})\mathrm{b}^2 \tag{1}$$

B. Battle Royal Optimizer

The process of identifying the most optimal solution among a set of feasible alternatives for a particular issue is sometimes referred to as optimization [18]. Optimization algorithms play a crucial part in several technical and commercial applications. Over the past few years, several optimization issues have been addressed via the use of metaheuristic algorithms, which draw inspiration from Darwin's theory of evolution [31]. Several algorithms, such as the gravitational search algorithm (GSA) [32] and water evaporation optimization (WEO) [33], draw inspiration from principles in physics. All algorithms aim to strike a delicate equilibrium between the processes of exploitation and exploration. In the year 2020, T.R. Farshi developed an optimization method known as BRO, which draws inspiration from the game strategy used in battle royale video games. The BRO starts by initializing a population with random individuals that are evenly distributed over the given spatial domain. Subsequently, every soldier or player employs a weapon to engage in combat by discharging projectiles at the closest adversary. The soldier who has a more advantageous position inflicts harm on another soldier. Furthermore, it is possible to represent all of these concepts numerically:

$$X_{dm,d} = X_{dm,d} + r\left(X_{b,d} - X_{dm,d}\right) \tag{2}$$

where, r is a randomly produced number that is spread out evenly in the range [0,1] and $X_{m,d}$ is the place of the hurt man in the d-dimensional space.

If a wounded soldier is still able to inflict harm on his opponent, the damage is reset to zero in the subsequent round. To provide improved exploration and convergence, a soldier is randomly regenerated from the probable issue space once their damage surpasses the threshold amount.

$$X_{d,m} = 0 \tag{3}$$

The returning soldier from a fatal combat zone is shown as:

$$X_{dm,d} = r(u_d - I_d) + I_d \tag{4}$$

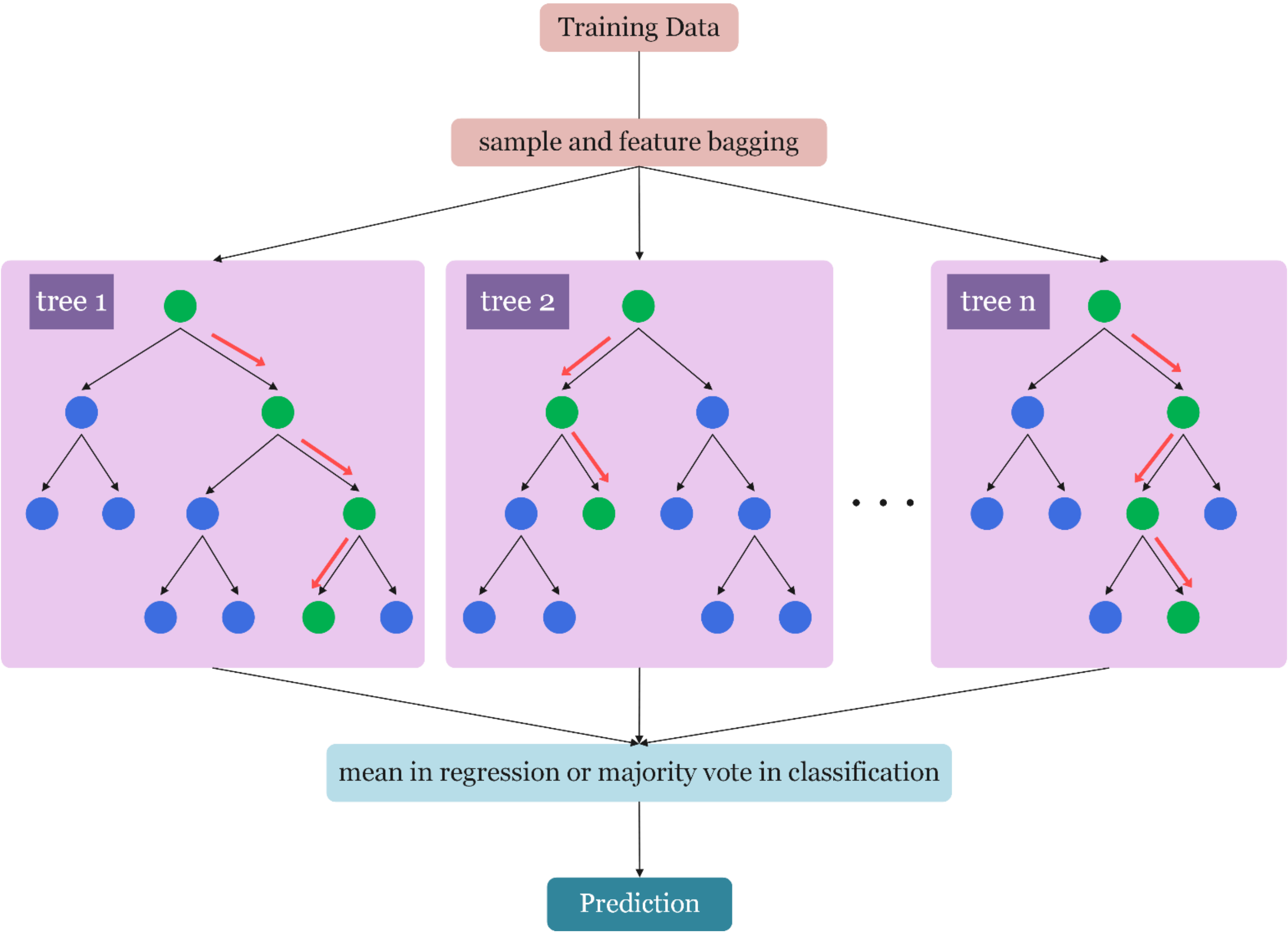
Training Data
sample and feature bagging
tree 1
tree 2
tree n
• • •
mean in regression or majority vote in classification
Prediction

Fig. 1. The random forest's architecture.

where $,u_d$ and I_d represent the dimension's upper and lower bounds, respectively. The search space continues to shrink in the direction of the optimal answer with each repetition. The quantity of iterations is associated with certain iteration as:

$$\begin{gathered}\text{delta } = \log 10(\text{ maxcircle }) \\ \text{delta } = \text{ delta } + \text{ round } \left(\frac{\text{delta}}{2}\right)\end{gathered} \tag{5}$$

where, maxcircle denotes the highest generation count.

Updates are made to the upper and lower boundaries as:

$$\begin{gathered}u_d = X_{\text{bes },d} + SD(X_d) \\ I_d = X_{\text{bes },d} - SD(X_d)\end{gathered} \tag{6}$$

The standard deviation of the whole population in dimension d is represented by $SD(X_d)$. The tuning of the random forest hyperparameters and the optimal values that were discovered through the use of the BRO optimizer are both detailed at the beginning of Table I.

TABLE I. HYPERPARAMETERS SETTING USING THE BRO ALGORITHM

Random Forest		BRO
Max depth	[10, 100, None]	60
Max features	[auto and sqr]	auto
Min samples leaf	[1, 4]	2
Min samples split	[2, 10]	3
Random state	[4, 24, 42, 64, 88]	24
Numbers of estimators	[200, 2000]	500

C. Moth-flame Optimization

In 2015, S. Mirjalili presented the stochastic optimization technique known as MFO [19]. Moths fly a great distance in a straight line by maintaining a constant angle concerning the moon. Nevertheless, the moth eventually converges on the artificial light after being caught in a spiral route around it [34]. By mimicking the moths' logarithmic spiral movement above the flame, the MFO algorithm determines the best solution. In the search space, a haphazard group of moths is first established. Their locations are updated in a spiral pattern concerning the flame, taking into account that the moth's movement should not be beyond the search space. It is possible to imagine that the moths are moving in all directions in a hyper ellipse around the flame. Each moth's location is updated in relation to its associated flame because the algorithm becomes stuck in local optima as a result of the moths' migration towards it. This lowers the likelihood of local optima stagnation and causes each moth to travel around distinct flames [34]. Every time the algorithm iterates in search of the optimal solution, the flame location is likewise changed, enhancing its exploration potential. Moths' migration to various flame positions in search space increases the degree of exploration but also reduces the capacity for exploitation. Finding a balance between exploitation and exploration is the primary goal of every optimization method. To increase the algorithm's exploitation potential, an adaptive approach for calculating the amount of flames is suggested. Throughout the iteration, the number of flames is adaptively reduced to guarantee that the moth adjusts its location to match the best-updated flame in the final set of iterations [34]. MFO is often used as:

$$M = \begin{bmatrix} CO_{1,1} & CO_{1,2} & \cdots & \cdots & CO_{1,h} \\ CO_{2,1} & CO_{2,2} & \cdots & \cdots & CO_{2,h} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ CO_{a,1} & CO_{a,2} & \cdots & \cdots & CO_{n,h} \end{bmatrix} \quad (7)$$

$$S = \begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & \cdots & S_{1,h} \\ S_{2,1} & S_{2,2} & \cdots & \cdots & S_{2,h} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{a,1} & S_{a,2} & \cdots & \cdots & S_{2,h} \end{bmatrix} \quad (8)$$

In the current context, the variable "h" denotes the number of dimensions, whereas the variable "a" indicates the number of moths.

The process of global optimization is carried out via the use of the three-step MFO technique.

$$MFO = (I, F, T) \quad (9)$$

The term "function" refers to a mathematical concept that describes the relationship between a set of The letter "I" is used to represent a distinct mathematical function, while the sign "F" is used to describe the flight trajectory of a moth as it explores its surroundings in pursuit of appropriate habitat. Furthermore, the sign T is used to denote the specific factors that dictate the cessation of the moth's flight.

$$X_i = t(C_i, S_j) \quad (10)$$

The formula used in this particular situation incorporates the twisting function, written as t, the quantity of the i-th moths marked as C_i, and the quantity of the j-th flames designated as S_j.

$$S(C_i, S_j) = Z_i \cdot e^{bt} \cdot cos(2\pi t) + S_j \, (11)$$

the variable Z_i denotes the spatial distance between the moth and the flame. The parameter b is a constant within the scope of this research. Furthermore, the variable t represents a stochastic quantity drawn from the closed interval [-1,1].

$$Zi = |S_j - X_i| \quad (12)$$

A description of the tuning of the random forest hyperparameters and the optimal values that were discovered through the use of the MFO optimizer can be found in Table II.

TABLE II. HYPERPARAMETERS SETTING USING THE MFO ALGORITHM

Random Forest		MFO
Max depth	[10, 100, None]	50
Max features	[auto and sqr]	auto
Min samples leaf	[1, 4]	1
Min samples split	[2, 10]	4
Random state	[4, 24, 42, 64, 88]	64
Numbers of estimators	[200, 2000]	200

D. Artificial Bee Colony

The ABC strategy, presented by Karaboga and Basturk [21], is a meta-heuristic optimization method that operates on a population-based approach. The modification mentioned in reference [35] is especially intended for discrete optimization situations. The ABC algorithm is derived from the fundamental search concepts that are based on the intelligent foraging behavior shown by honeybee swarms. The algorithm categorizes foraging bees into three distinct groups based on their behavior and responsibilities: employed bees, observer bees, and scout bees. Bees' exhibit collective organization to optimize the accumulation of nectar, which serves as their primary energy source, inside the food storage located in their hives, as seen in Fig. 2. This is achieved via the use of suitable division of labor strategies [36]. The foraging bees are in charge of taking advantage of food sources by collecting and transporting them back to their hives, often exploring sites that other foragers have previously visited. The observer bees situated inside their hives acquire knowledge of the foraged food sources via the communication of employed bees, which is conveyed through their dance behavior upon returning to the hives. Subsequently, the observer bees choose to visit the food sources based on the perceived quality of those food sources, as shown by the length of the dances performed. The scout bees explore novel food sources by venturing in a direction that is chosen randomly, as the summary of this process is shown in Fig. 3. The scout and observer bees are often denoted as unoccupied bees, undergoing a transformation into occupied bees subsequent to their identification of a novel food source during their foraging activities [36].

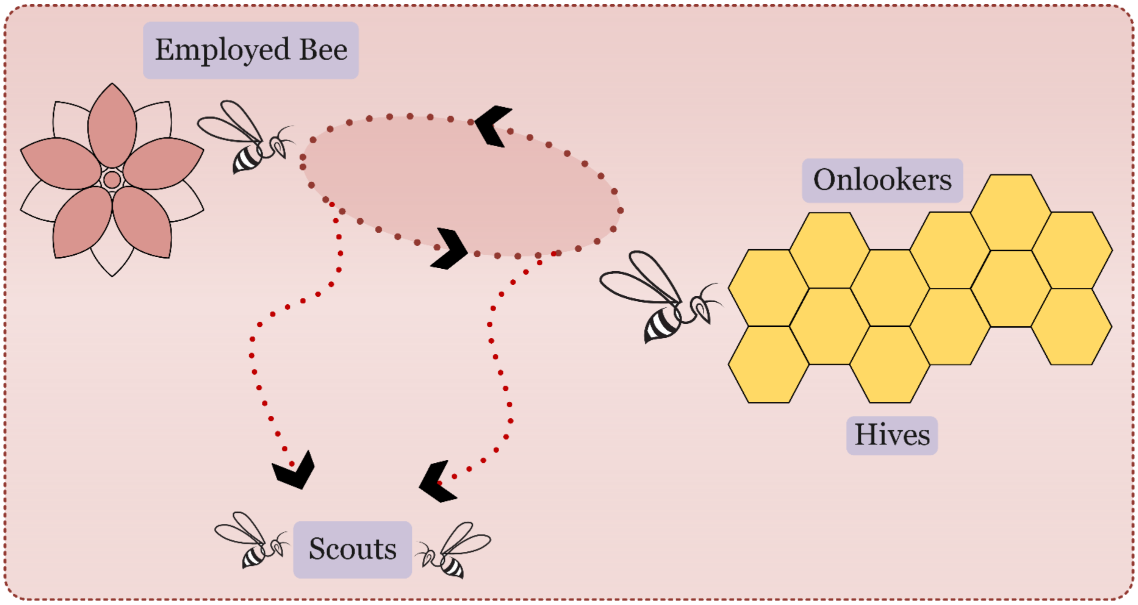

Fig. 2. The artificial bee colony optimizer's operational mechanism.

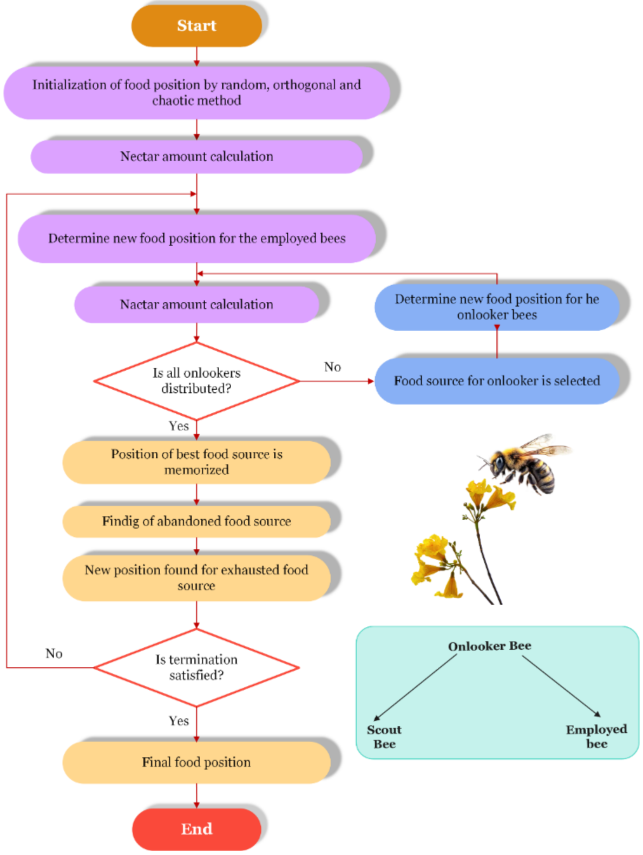
Start
Initialization of food position by random, orthogonal and chaotic method
Nectar amount calculation
Determine new food position for the employed bees
Nactar amount calculation
Determine new food position for he onlooker bees
Is all onlookers distributed?
No
Food source for onlooker is selected
Yes
Position of best food source is memorized
Findig of abandoned food source
New position found for exhausted food source
No
Is termination satisfied?
Yes
Final food position
End
Onlooker Bee
Scout Bee
Employed bee

Fig. 3. Flow diagram of ABC.

The initial food sources are expressed by applying a random solution vector boundary value.

$$x_{i,j} = x_j^{min} + rand(0,1)\left(x_j^{max} - x_j^{min}\right) \tag{13}$$

where, $j = 1, \ldots D, SN$, and $i = 1, \ldots SN$. D represents the parameters that need to be optimized, SN is the number of solutions, and x_j^{min} and x_j^{max} denote the lower and upper bounds of the j th parameter, respectively. Once the food sources have been started, the fitness value of each will be determined using the following formula.

$$\text{fit}_{\ i} = \frac{1}{1+obj \cdot fun_i} \tag{14}$$

where, $obj \cdot fun_i$ indicates the intentional action. SN provides precisely the same number of working bees and observers as there are answers. A single food source is equal to one active bee. Working bees and observers search for nearby food sources and adjust their location depending on the following equation to come up with fresh solutions.

$$v_{ij} = x_{ij} + r_{ij}\left(x_{ij} - x_{kj}\right) \tag{15}$$

where, j, k, and S are randomly selected, and k, i, and r_{ij} are random integers in the range $[-1,1]$. It's used to manage various communities and recalculate the fitness value of the new solution to see which of the v_{ij} and x_{ij} fitness values are bigger. Fitness is a particular probability that observer bees consider when choosing food sources, and they calculate it using the following formula.

$$p_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \tag{16}$$

where, the quantity of nectar in the relevant food supply is related to fit_i, the fitness value of the solution. The hired bees will turn into scout bees and go on a haphazard quest if they investigate the available food supply for longer than the upper limit. Using the ABC optimizer, the optimal values for the random forest hyperparameters were determined, and Table III provides a description of the tuning process.

TABLE III. HYPERPARAMETERS SETTING USING THE ABC ALGORITHM

Random Forest		ABC
Max depth	[10, 100, None]	80
Max features	[auto and sqr]	auto
Min samples leaf	[1, 4]	2
Min samples split	[2, 10]	2
Random state	[4, 24, 42, 64, 88]	42
Numbers of estimators	[200, 2000]	300

IV. GATHERING AND PREPARING DATA

Several factors should be considered while doing a comprehensive examination of a company, including the trading volume and the Open, High, Low, and Close (OHLC) prices for a certain period of time. Data on the Nikkei 225 stock performance from the start of 2013 to the end of 2022 was acquired for this particular research. The dataset included details on the OHLC prices and trading volume for each day throughout the specified time frame. An extensive examination of the data landscape was conducted as part of the first step to spot any anomalies, outliers, or discrepancies that might cast doubt on the accuracy of the findings. The dataset was cleaned and prepared many times after the research was finished. Scaling and normalizing were only two of the numerous methods used in the procedures to reduce error rates and encourage consistency in the training outputs. To maximize the models' functionality, two sets of prepared data were made. As observed in Fig. 4, a partitioning method was used in this study, allocating 80% of the dataset for training and the remaining 20% for validation and testing. The main objective of this division was to strike the ideal balance between the need for a sizable amount of data to train the model and the demand for a vast and unknown dataset to perform extensive testing and validation.

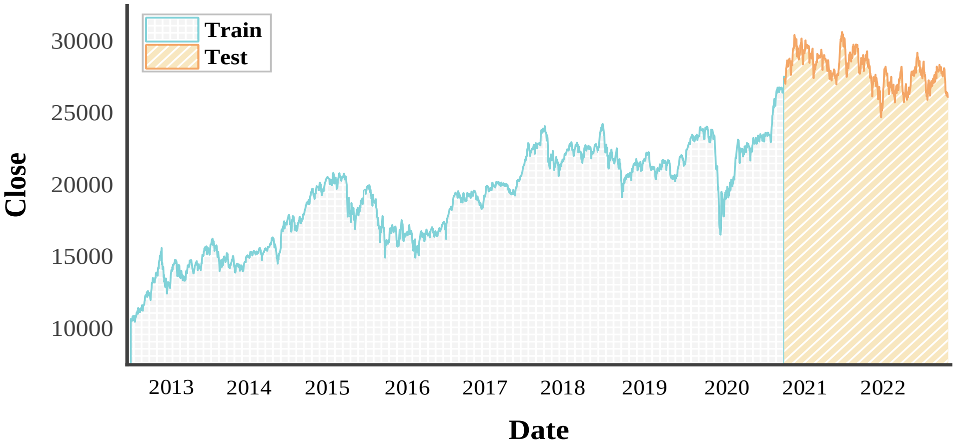

Fig. 4. Splitting data into testing and training sets.

V. Assessment Metrics

The evaluation of the accuracy of the future forecast was conducted by using a variety of performance measures. The carefully chosen metrics provide a thorough evaluation of the reliability and precision of the predictions. Various factors were taken into account throughout the evaluation process. The mean absolute percentage error (MAPE), mean absolute error (MAE), coefficient of determination (R^2), which measures the proportion of the dependent variable's variability that can be explained by the independent variable, and mean square error (MSE) is employed to calculate the average absolute discrepancy between the predicted and observed values. These strategies provide valuable help and significantly enhance the process of assessing the accuracy of forecasting models.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{17}$$

$$MSE = \frac{1}{N}\sum_{k=0}^{n}\binom{n}{k}(Fi - Yi)b^2 \tag{18}$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{19}$$

$$MAPE = \left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|\right) \times 100 \tag{20}$$

VI. Results and Discussions

A. Statistic Values

This inquiry phase includes Table IV, which offers a comprehensive description of the statistical data in the dataset. The information is easier to grasp when the OHLC price and volume statistics are included in the table. To conduct a comprehensive and precise examination of the data, statistical metrics like as the count, mean, minimum, maximum, standard deviation (Std.), 50%, and variance may be used.

TABLE IV. A Statistical Summary of the Concerned Data Set is Provided

	count	mean	Std.	min	50%	max	variance
Open	2442	20813.83	4765.013	10405.67	20538.9	30606.15	22705344
High	2442	20926.76	4777.106	10602.12	20632.72	30795.78	22820737
Low	2442	20690.79	4748.074	10398.61	20451.26	30504.81	22544211
Volume	2442	3730.003	1985.287	0	3180	19840	3941363
Close	2442	20812.22	4763.784	10486.99	20559.85	30670.1	22693641

VII. Compare and Analyses

This research's primary objective is to identify and assess the best hybrid algorithm for stock price prediction. The creation of forecasting models and a deep comprehension of the many factors influencing stock market trends serve as the study's cornerstones. The primary objective is to provide analysts and investors with valuable insights to enable them to make informed and prudent investment choices. The performance of each model is thoroughly analyzed in Fig. 5, 6 and Tables V and VI. A comprehensive evaluation of each model's efficacy is also provided.

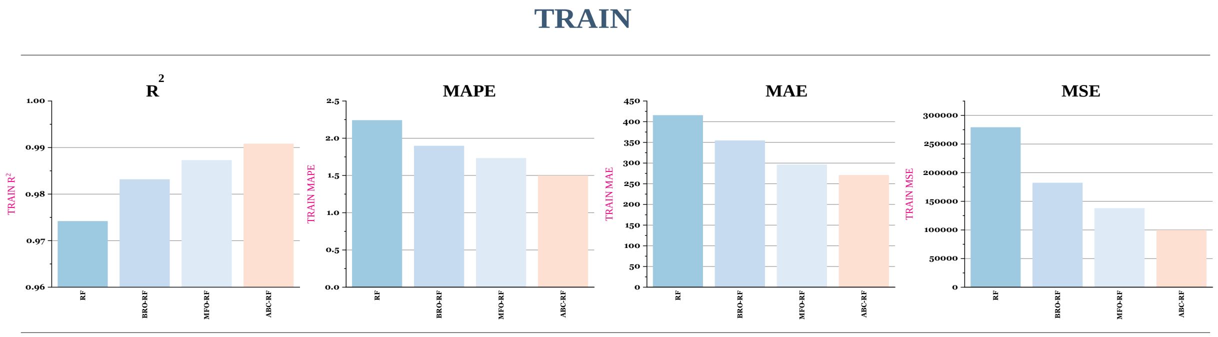

Fig. 5. The suggested model's training outcomes for the R^2, MAPE, MAE, and MSE.

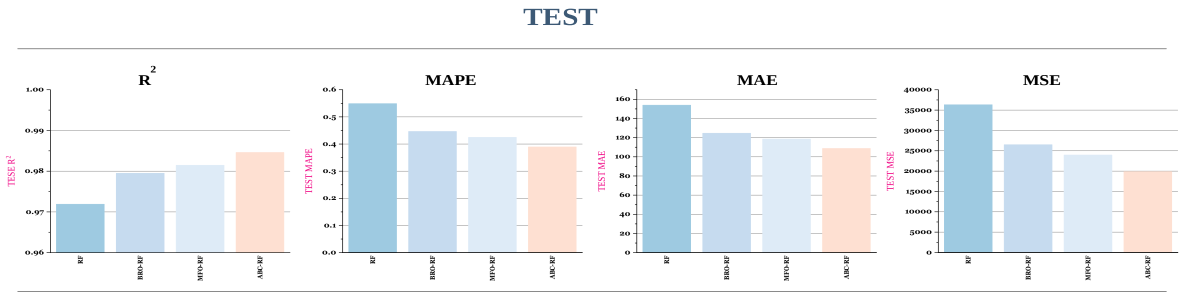

Fig. 6. The suggested model's testing outcomes for the R^2, MAPE, MAE, and MSE.

TABLE V. AN ESTIMATE OF THE MODELS' ASSESSMENT RESULTS

	TRAIN SET				**TEST SET**			
	R^2	MAPE	MAE	MSE	R^2	MAPE	MAE	MSE
RF	0.974	2.24	415.17	278883	0.972	0.55	153.92	36337
BRO-RF	0.983	1.90	354.19	181946	0.979	0.45	124.67	26519
MFO-RF	0.987	1.73	295.96	137611	0.981	0.42	118.58	23969
ABC-RF	0.991	1.50	270.41	99445.4	0.985	0.39	108.74	19908

TABLE VI. AN ESTIMATE OF THE MODELS' ASSESSMENT RESULTS FOR THE S&P 500 INDEX

	TRAIN SET				TEST SET			
	R^2	MAPE	MAE	MSE	R^2	MAPE	MAE	MSE
RF	0.9656	3.42	85.38	9273.03	0.9653	1.11	46.57	3177.56
BRO-RF	0.9833	2.42	53.69	4491.54	0.9805	0.74	30.27	1782.87
MFO-RF	0.9794	1.91	48.53	5554.27	0.9776	0.80	33.16	1970.73
ABC-RF	0.9907	1.90	45.84	2491.17	0.9885	0.59	24.32	1050.46

Four commonly used metrics were employed to evaluate the data analysis: MSE, MAPE, MAE, and R^2. It is generally agreed upon that the aforementioned metrics provide a thorough evaluation of the overall efficacy, accuracy, and reliability of the analysis. The RF model's efficacy has been evaluated using the MAE, MSE, R^2, and MAPE criteria, both in the presence and absence of an optimizer. It will be able to get a better grasp of the model's functionality via this approach and provide opinions based on the information it has gained. After analyzing the training and test sets, it was seen that in the absence of the optimizer, the RF model produced R^2 values of 0.974 and 0.972 for the training and testing sets, respectively. In comparison, the MAE values for training and testing were 415.17 and 153.92, while the MSE values were 278883 and 36337. The MAPE values for the training and testing sets were 2.24 and 0.55, respectively. Using optimizers significantly boosted the RF model's efficiency. The results of using the BRO optimizer have been significantly improved, as can be seen by looking at the drop in the MAPE value to 1.90 for the training dataset and 0.45 for the testing dataset, as well as the R^2 values for training and testing were 0.983 and 0.979 for MAE and MSE values, which fell to 354.19 and 181946 for training and 124.67 and 26519 for testing. The MFO-RF model performed better than the BRO-RF model, according to a comparative study that was done between the two models. During training and testing, the MFO-RF model's R^2 values were determined to be 0.987 and 0.981, respectively. It's important to understand that the MSE values for training and testing dropped to 137611 and 23969. The MAE and MAPE values also decreased to 295.96 and 1.73 for training and 118.58 and 0.42 for testing, respectively. The results of this research show that the MFO-RF model performs better than the BRO-RF model in terms of efficacy. The noteworthy R^2 values of 0.991 and 0.985 obtained during training and testing, respectively, illustrate the efficacy of the ABC-RF model. The ABC-RF model performed better than the other models; it showed the lowest MAPE values, 1.50 for training and 0.39 for testing, and MAE values, which dropped to 270.41 for training and 108.74 for testing, and MSE value for testing was 19908. The findings described above indicate the high degree of accuracy and dependability that the ABC-RF model exhibits, proving its usefulness for the intended purpose.

In comparison to the RF, BRO-RF, and MFO-RF models, the ABC-RF model consistently achieves the highest scores or lowest error values across all evaluation metrics, including R^2, MAPE, MAE, and MSE. The consistent pattern observed highlights the ABC-RF model's exceptional predictive accuracy and capacity for generalization. Significantly, the superiority in performance of the model persists from the training set to the test set, suggesting that it is resistant to overfitting. The robustness of the ABC-RF model indicates that it effectively captures latent data patterns and relationships, resisting the influence of noise. As a result, its capability to generalize to unseen data is enhanced. Moreover, the efficacy of the ABC-RF model in various market environments is apparent, as evidenced by its performance on both the S&P 500 and Nikkei 225 indices. This implies that the performance of the system is not limited to particular datasets or markets, but is rather a result of its rigorous optimization and modeling methodology. The application of the Artificial Bee Colony optimization method almost certainly plays a substantial role in the superior performance of the ABC-RF model. This optimization method has gained recognition for its effectiveness in investigating search spaces and identifying solutions of superior quality. It has the potential to surpass the methods utilized in BRO-RF and MFO-RF. Furthermore, ABC-RF achieves an admirable equilibrium between intricacy and efficacy, as demonstrated by its reduced error metrics and increased R^2 values in comparison to alternative models. This suggests that the model effectively utilizes the benefits of Random Forests while simultaneously optimizing parameters to reduce errors and improve predictive precision. In brief, the ABC-RF model exhibits several advantages over its competitors (RF, BRO-RF, and MFO-RF): superior predictive accuracy, robustness against over fitting, consistency across diverse market indices, efficient optimization, and the capacity to strike a balanced equilibrium between complexity and performance. The results of this study highlight the effectiveness of utilizing the Artificial Bee Colony optimization method to enhance the performance of Random Forest models when attempting to forecast stock prices.

Extensive research has shown the dependability of the ABC-RF model as a reliable instrument for accurately forecasting stock prices. By comparing the Nikkei 225 index curves with the analogous curves shown in Figs. 7, 8, one may assess the efficacy of the model. The ABC-RF model performs better than models like RF, BRO-RF, and MFO-RF when it comes to stock price forecasting. The ABC-RF model combines the random forest with the artificial bee colony technique to estimate stock prices, according to a thorough analysis of the model's efficacy. The RF technique lowers stock price volatility and improves the accuracy of future trend estimates, both of which boost the accuracy of the model. One of the characteristics that distinguished the ABC-RF model from the others was its ability to learn from previous datasets. A model has to be able to learn from prior datasets and adjust its projections in response to changing market circumstances to predict stock prices with any level of accuracy. In conclusion, due to its accuracy, reliability, and ability to draw conclusions from historical datasets, the ABC-RF model is a highly helpful tool for stock price prediction. For those looking to conclude profitable stock market trades, it is the preferred option because of its utilization of the RF algorithm and ABC optimizer, as well as its adaptability to shifting market circumstances.

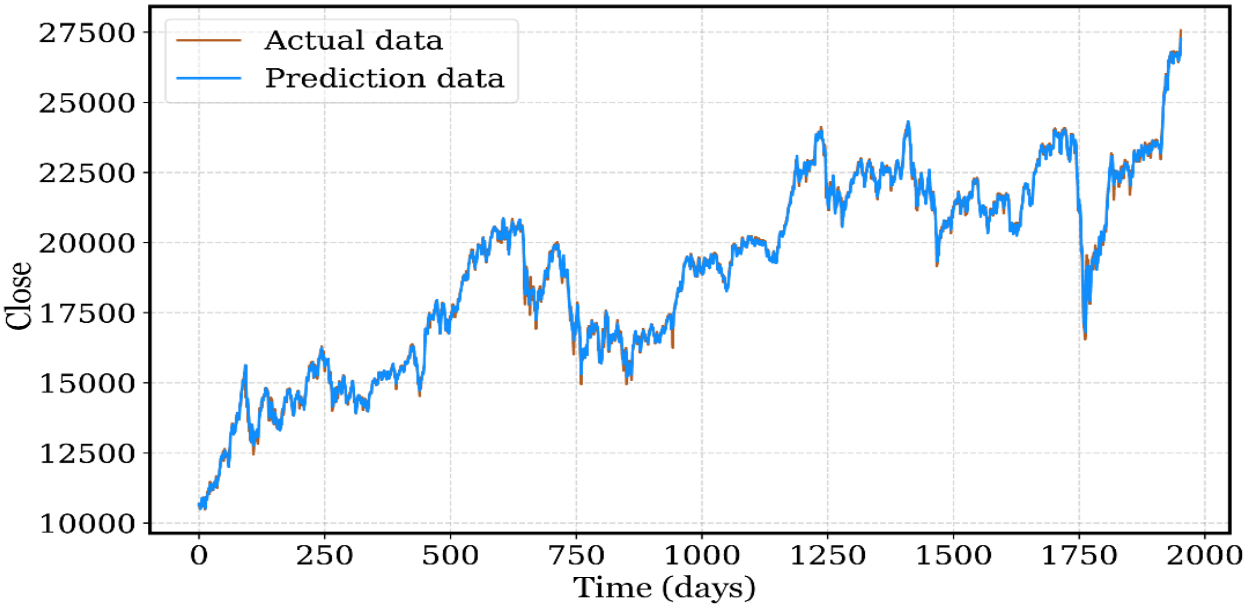

Fig. 7. The forecasting graph created during the training phase by using the ABC-RF approach.

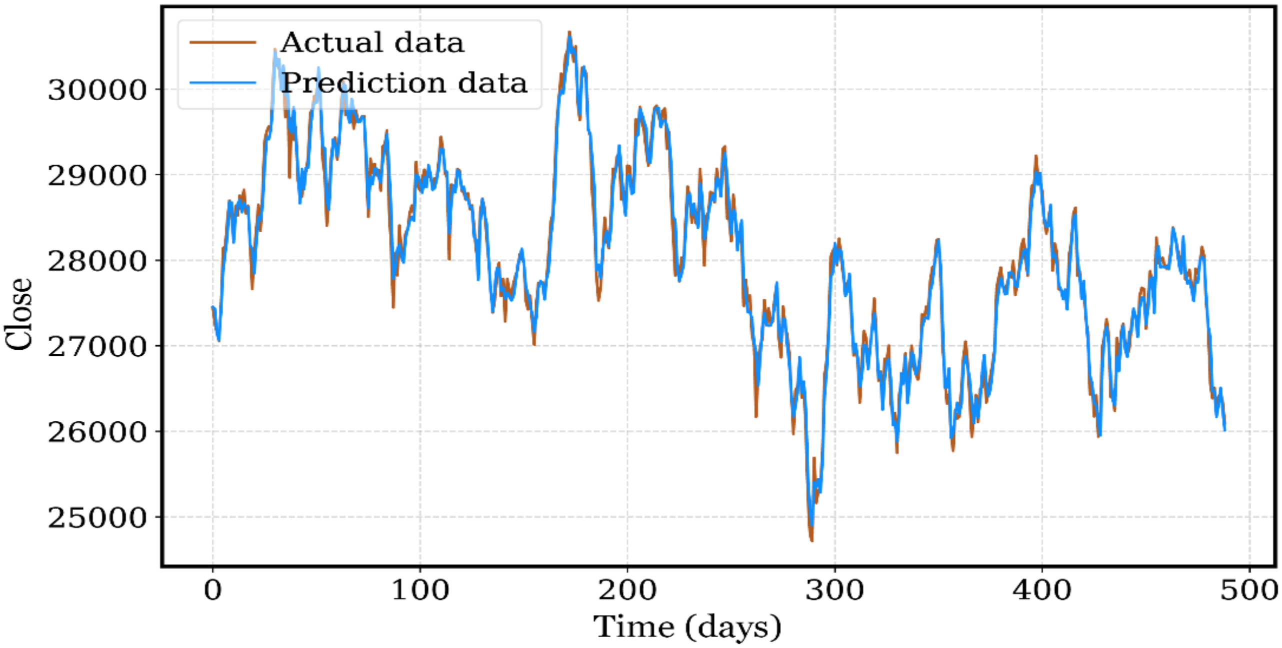

Fig. 8. The forecasting graph created during the testing phase by using the ABC-RF approach.

VIII. Conclusion

The stock market exhibits a significant degree of volatility. Nevertheless, it provides investors with significant potential to increase the value of their investments. One approach to do this is by using various visual aids such as charts, graphs, and balance statements of corporations. Alternatively, individuals have the option to use a Machine Learning Algorithm to do the task on their behalf. The model has the capability to efficiently analyze historical data, trend lines, charts, and other relevant information and provide informed recommendations for future actions. Machine Learning technology has been deemed groundbreaking and has shown its efficacy for several investors. The multitude of complex aspects that influence stock price prediction may make the development of reliable and accurate prediction models difficult. A deep comprehension of the non-linear and volatile aspects of the market is necessary to provide reliable projections. Fortunately, the ABC-RF model provides a workable solution to these problems and has shown to be very accurate and reliable. This research evaluated the performance of many stock price prediction models, including RF, BRO-RF, and MFO-RF. The RF parameters were optimized using hyperparameter optimization methods, such as BRO, MFO, and ABC.

Nevertheless, the ABC optimizer approach produced superior outcomes when paired with RF. The dataset utilized in this analysis consisted of OHLC price and volume data for the Nikkei 225 index from the beginning of 2013 to the end of 2022. The experiment's findings show how accurate and dependable the ABC-RF model is in estimating stock values.

- As part of the research process, a comparison study with several other models was carried out to evaluate the accuracy and predictive potential of the ABC-RF model. Based on the data gathered, it can be said that the ABC-RF model consistently performed better than the other models. The calculated R^2 score of 0.985 shows how accurate the prediction models are. The model's predictions seemed to be quite accurate, with an observed MSE score of 19908 and an MAE value of 108.74 throughout the testing process. With a 0.39 MAPE score, the model showed a constant capacity for generating trustworthy forecasts. The ABC-RF model demonstrated greater accuracy and effectiveness in relation to the other models being studied.

The ABC-RF model provides investors with valuable insights to facilitate educated investing decision-making and serves as an effective instrument for stock price prediction. The study is limited by its use of historical data from the Nikkei 225 index, which may not capture all important market conditions or unexpected events. As a result, the conclusions may not be applicable to other markets. In addition, although the suggested ABC-RF model exhibited improved performance, its intricacy may impede comprehensibility for investors lacking extensive knowledge in machine learning, presenting obstacles for practical use. The model's assumptions of stationarity may not adequately account for the non-stationary characteristics of stock market dynamics, which could lead to a decrease in its accuracy when anticipating abrupt shifts or structural changes. Furthermore, the utilization of the model may be hindered for certain users due to the excessive computational resources and time needed, particularly when dealing with extensive datasets. Despite attempts to optimize parameters, there is still a possibility of overfitting, which highlights the need for rigorous validation methodologies and additional testing on out-of-sample data. Ultimately, the accuracy of the model's predictions can be affected by external factors like geopolitical events or market shocks, which are difficult to adequately include, thereby compromising its reliability and robustness. Subsequent research endeavors may include the following: expanding the scope of the study to encompass diverse markets in order to evaluate the adaptability of the ABC-RF model; enhancing the interpretability of the model for non-expert users; investigating dynamic modeling approaches to more accurately capture market fluctuations; optimizing computational resources and scalability; integrating risk management techniques; investigating ensemble learning methods; and designing mechanisms to provide real-time forecasts and updates. The aforementioned endeavors seek to improve the model's capacity to generalize, be utilized, be accurate, and be robust. As a result, they support well-informed investment decision-making and tackle the ever-changing complexities of the financial markets.

Funding

This work was supported by Projects of key colleges and universities in Guangdong province: 2019WZDXM027, Guangzhou philosophy and Social Sciences Co-construction project: 2021GZGJ83, Accounting scientific research project of Guangdong Provincial Finance Department: 9-22, Special project in key fields of colleges and universities in Guangdong province: 2020ZDZX3009, Feature innovation project of colleges and universities in Guangdong province: 2020KTSCX163, Feature innovation project of colleges and universities in Guangdong province :2018KTSCX256 and Guangdong baiyun university key project: 2019BYKYZ02

References

[1] S. Claessens, J. Frost, G. Turner, and F. Zhu, “Fintech credit markets around the world: size, drivers and policy issues,” BIS Quarterly Review September, 2018.

[2] W. Li et al., “The nexus between COVID-19 fear and stock market volatility,” Economic research-Ekonomska istraživanja, vol. 35, no. 1, pp. 1765–1785, 2022.

[3] J. W. Goodell, R. J. McGee, and F. McGroarty, “Election uncertainty, economic policy uncertainty and financial market uncertainty: a prediction market analysis,” J Bank Financ, vol. 110, p. 105684, 2020.

[4] B. Kelly, Ľ. Pástor, and P. Veronesi, “The price of political uncertainty: Theory and evidence from the option market,” J Finance, vol. 71, no. 5, pp. 2417–2480, 2016.

[5] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, “Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques,” Expert Syst Appl, vol. 42, no. 1, pp. 259–268, 2015.

[6] Z. Wang et al., “Measuring systemic risk contribution of global stock markets: A dynamic tail risk network approach,” International Review of Financial Analysis, vol. 84, p. 102361, 2022.

[7] Z. Li, W. Cheng, Y. Chen, H. Chen, and W. Wang, “Interpretable click-through rate prediction through hierarchical attention,” in Proceedings of the 13th International Conference on Web Search and Data Mining, 2020, pp. 313–321.

[8] N. Ashfaq, Z. Nawaz, and M. Ilyas, “A comparative study of Different Machine Learning Regressors For Stock Market Prediction,” 2021. doi: 10.48550/arxiv.2104.07469.

[9] V. U. Kumar, A. Krishna, P. Neelakanteswara, and C. Z. Basha, “Advanced prediction of performance of a student in an university using machine learning techniques,” in 2020 international conference on electronics and sustainable communication systems (ICESC), IEEE, 2020, pp. 121–126.

[10] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, vol. 2. Springer, 2009.

[11] Y.-Y. Song and Y. Lu, “Decision tree methods: applications for classification and prediction.,” Shanghai Arch Psychiatry, vol. 27, no. 2, pp. 130–135, Apr. 2015, doi: 10.11919/j.issn.1002-0829.215044.

[12] P. Skoda and F. Adam, Knowledge Discovery in Big Data from Astronomy and Earth Observation: Astrogeoinformatics. Elsevier, 2020.

[13] L. Breiman, “Random forests,” Mach Learn, vol. 45, pp. 5–32, 2001.

[14] O. R. Olaniran and M. A. A. Abdullah, “Bayesian weighted random forest for classification of high-dimensional genomics data,” Kuwait Journal of Science, vol. 50, no. 4, pp. 477–484, 2023, doi: 10.1016/j.kjs.2023.06.008.

[15] A. Gatera, M. Kuradusenge, G. Bajpai, C. Mikeka, and S. Shrivastava, “Comparison of random forest and support vector machine regression models for forecasting road accidents,” Sci Afr, vol. 21, p. e01739, 2023, doi: 10.1016/j.sciaf.2023.e01739.

[16] S. Mirjalili, “The ant lion optimizer,” Advances in engineering software, vol. 83, pp. 80–98, 2015.

[17] S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey wolf optimizer,” Advances in engineering software, vol. 69, pp. 46–61, 2014.

[18] T. Rahkar Farshi, “Battle royale optimization algorithm,” Neural Comput Appl, vol. 33, no. 4, pp. 1139–1157, 2021.

[19] S. Mirjalili, “Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm,” Knowl Based Syst, vol. 89, pp. 228–249, 2015.

[20] D. Simon, “Biogeography-based optimization,” IEEE transactions on evolutionary computation, vol. 12, no. 6, pp. 702–713, 2008.

[21] D. Karaboga and B. Basturk, “A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm,” Journal of global optimization, vol. 39, pp. 459–471, 2007.

[22] S. C. Agrawal, “Deep learning based non-linear regression for Stock Prediction,” IOP Conference Series: Materials Science and Engineering ; volume 1116, issue 1, page 012189 ; ISSN 1757-8981 1757-899X, 2021, doi: 10.1088/1757-899x/1116/1/012189.

[23] M. Petchiappan and J. Aravindhen, “Comparative Study of Machine Learning Algorithms towards Predictive Analytics,” Recent Advances in Computer Science and Communications ; volume 16, issue 6 ; ISSN 2666-2558, 2023, doi: 10.2174/2666255816666220623160821.

[24] S. Sathyabama, S. C. Stemina, T. SumithraDevi, and N. Yasini, “Intelligent Monitoring and Forecasting Using Machine Learning Techniques,” Journal of Physics: Conference Series ; volume 1916, issue 1, page 012175 ; ISSN 1742-6588 1742-6596, 2021, doi: 10.1088/1742-6596/1916/1/012175.

[25] A. Menaka, V. Raghu, B. J. Dhanush, M. Devaraju, and M. A. Kumar, “Stock Market Trend Prediction Using Hybrid Machine Learning Algorithms,” International Journal of Recent Advances in Multidisciplinary Topics; Vol. 2 No. 4 (2021); 82-84 ; 2582-7839, Feb. 2021, [Online]. Available: https://journals.ijramt.com/index.php/ijramt/article/view/643

[26] U. Demirel, H. Cam, and R. Unlu, “Predicting Stock Prices Using Machine Learning Methods and Deep Learning Algorithms: The Sample of the Istanbul Stock Exchange,” 2021, [Online]. Available: https://hdl.handle.net/20.500.12440/3191

[27] P. M. Tembhurney and S. Pise, “Stack Market Prediction Using Machine Learning (ML) Algorithms,” International Journal for Indian Science and Research Volume-1(Issue -1) 08, Feb. 2022, [Online]. Available: https://zenodo.org/record/6787069

[28] P. Jain, A. Choudhury, P. Dutta, K. Kalita, and P. Barsocchi, “Random forest regression-based machine learning model for accurate estimation of fluid flow in curved pipes,” Processes, vol. 9, no. 11, p. 2095, 2021.

[29] M. Wiesmeier et al., “Estimation of total organic carbon storage and its driving factors in soils of Bavaria (southeast Germany),” Geoderma Regional, vol. 1, pp. 67–78, 2014.

[30] D. Chen, N. Chang, J. Xiao, Q. Zhou, and W. Wu, “Mapping dynamics of soil organic matter in croplands with MODIS data and machine learning algorithms,” Science of the Total Environment, vol. 669, pp. 844–855, 2019.

[31] H. A. Abbass, C. S. Newton, and R. Sarker, Heuristic and optimization for knowledge discovery. IGI Global, 2001.

[32] E. Rashedi, H. Nezamabadi-Pour, and S. Saryazdi, “GSA: a gravitational search algorithm,” Inf Sci (N Y), vol. 179, no. 13, pp. 2232–2248, 2009.

[33] A. Kaveh and T. Bakhshpoori, “Water evaporation optimization: a novel physically inspired optimization algorithm,” Comput Struct, vol. 167, pp. 69–85, 2016.

[34] A. Sharma et al., “Improved moth flame optimization algorithm based on opposition-based learning and Lévy flight distribution for parameter estimation of solar module,” Energy Reports, vol. 8, pp. 6576–6592, 2022, doi: https://doi.org/10.1016/j.egyr.2022.05.011.

[35] M. H. Kashan, N. Nahavandi, and A. H. Kashan, “DisABC: a new artificial bee colony algorithm for binary optimization,” Appl Soft Comput, vol. 12, no. 1, pp. 342–352, 2012.

[36] D. Karaboga and B. Basturk, “On the performance of artificial bee colony (ABC) algorithm,” Appl Soft Comput, vol. 8, no. 1, pp. 687–697, 2008.