Review and Analysis of Financial Market Movements: Google Stock Case Study

Yiming LU
School of Continuing Education, Shijiazhuang College of Applied Technology, Shijiazhuang, 050081, China

***Abstract*—A financial marketplace where shares of companies with public listings are bought and sold is called the stock market. It serves as a gauge of a nation's economic health by taking into account the operations of individual businesses as well as the general business climate. The relationship between supply and demand affects stock prices. Though it might be dangerous, stock market investing has the potential to provide large rewards in the long run. Together with increased prediction accuracy, optimization techniques such as Biogeography-based optimization (BBO), Artificial bee colony algorithm (ABC) and Aquila Optimization (AO) Algorithm further enhance the Extreme gradient boosting (XGBoost) ability to adapt to changing market conditions. The results were 0.955, 0.966, 0.972, and 0.982 for XGBoost, BBO-XGBoost, ABC-XGBoost, and AO-XGBoost, in that order. The performance difference between AO-XGBoost and XGBoost shows how combining with the optimizer may enhance the model's performance. By comparing the output of many optimizers, the most accurate optimization has been determined to be the model's main optimizer.**

Keywords—Stock future trend; financial market; investment; machine learning algorithms; Google stock

I. Introduction

A. Background knowledge

The task of forecasting stock prices is difficult [1], [2], [3], [4]. The long-term unpredictability in the market presents a significant challenge. The conventional theory of markets posits that stock valuations are inherently random and lack any discernible pattern. Nevertheless, contemporary technical analysis has provided evidence to support the notion that historical data largely influences stock prices, hence emphasizing the importance of understanding patterns of movement in order to make accurate projections [5]. Various economic factors, including political events, general economic circumstances, commodity price indicators, investor sentiment, movements in other stock markets, and investor behavior, also exert an influence on the categorization and fluctuations of stock market entities [6]. Stock values are determined through the assessment of market capitalization, which is the total worth of a company's outstanding shares. Additionally, statistical data can be derived by employing a range of technical variables [7]. Stock indexes are often constructed using the prices of equities with significant market investments, serving as a means to assess the economic conditions of various nations. Empirical evidence suggests that stock market capitalization exerts a positive impact on a country's economic growth [8]. Investments pose inherent risks due to the volatile nature of stock price fluctuations. Furthermore, ascertaining the market standing of governments is sometimes a complex task. Due to the inherent characteristics of stock values, namely their dynamic, non-parametric, and non-linear nature, statistical models often encounter challenges in accurately predicting precise values and trends, hence diminishing their efficacy [9], [10]. Machine learning (ML) is often regarded as the most powerful technology due to its ability to leverage diverse algorithms to improve performance in specific case studies. ML is commonly acknowledged for its significant capability to identify trustworthy data and detect patterns within datasets [11], [12], [13], [14], [15]. It is imperative to acknowledge the challenging and intricate nature of projecting stock prices. Although machine learning algorithms cannot be deemed as reliable sources for generating precise forecasts, they have significantly impacted stock prediction and financial markets through many means. This paper examines the effects and contributions of ML techniques in the domain of stock prediction. ML algorithms possess the capability to analyze historical stock price data in order to identify patterns and trends that may not be readily discernible to human analysts. The individuals possess the capability to discern complex patterns inside the dataset and employ them in order to predict forthcoming fluctuations in prices [16]. Various techniques of machine learning are employed for the purpose of prediction. The model employed in this study is an Extreme gradient boosting (XGBoost). Chen et al. (2015) [17] developed XGBoost, which is effective at building boosted trees, performing parallel operations, and resolving issues related to both regression and classification. The primary aim of the method is to optimize the value of the objective function by the repeated combination of weak base learning models into a stronger learner [18]. The residual is utilized to adjust the preceding prediction so that the designated loss function may be optimized at each gradient boosting cycle. The optimizers provided for the hyperparameters optimizer of the model are Biogeography-based optimization (BBO) [19], [20], [21], Artificial bee colony algorithm (ABC) [22], [23], and Aquila Optimizer (AO) [24]. The concept of species movement based on habitat appropriateness is the foundation of biogeography-based optimization or BBO. Therefore, a solution is like a habitat for an optimization issue. A crowded environment, where the circumstances for living species are better than in other habitats, is a better answer for the population. The environment where living things struggle is the worst answer for the people. Because they share properties, the superior solutions draw in the inferior ones. Another optimization method used is ABC; three groups of bees make up the artificial bee colony in the ABC algorithm: employed bees, bystanders, and scouts. In the first half of the colony are the working artificial bees, while in the second half are the observers. There is one working bee for each food source. Stated differently, the quantity of food sources and the number of bees engaged are the same. An employed food source bee transforms into a scout. Another optimization method which has the best result, Aquila has a dark brown plumage, with a distinct golden-brown hue seen on the posterior region of its neck. Aquila has outstanding velocity and dexterity. Furthermore, the Aquila has robust lower extremities and formidable talons. This feature facilitates the capture of diverse prey. Aquila has been identified as an adult deer assailant. Aquila builds huge nests in elevated locations like as mountains or other elevated terrains. Aquila has exceptional cognitive abilities and demonstrates remarkable proficiency in hunting people.

B. Contributions and Novelties

This study offers valuable insights into the field of stock market prediction by seeking to enhance prediction accuracy through the integration of machine learning methodologies, specifically XGBoost. The model showcases improved flexibility in response to dynamic market conditions by integrating optimization techniques such as BBO, ABC, and AO. By conducting a thorough comparison of these optimization techniques, the study systematically determines that AO-XGBoost is the most effective model, demonstrating exceptional performance in predicting stock market values. This advantageous model shows potential for incorporation into dynamic trading systems, offering lucrative suggestions for traders and investors. The effectiveness of the proposed model is further confirmed through empirical validation using real-world data from Google stock prices, which shows minimal inaccuracies in forecasting stock prices. The study's well-defined performance evaluation highlights the dependability and resilience of the AO-XGBoost model, providing practical implications for decision-making in financial markets.

II. Literature Review

A. Related Works

Agrawal [25] focuses on stock market forecasting using machine learning algorithms, aiming to accurately estimate future stock values. The efficient market hypothesis suggests that stock prices depend on available information, making accurate prediction crucial for investor decision-making. The paper introduces a deep learning-based non-linear regression method for stock price prediction, evaluated on Tesla and New York Stock Exchange datasets from 2010 to 2020. The results indicate superior performance compared to existing machine learning approaches [25].

Hong et al. [26] employs Bidirectional Long Short-Term Memory (BLSTM), considered more accurate than unidirectional LSTM, to forecast near-future stock prices. The dynamic nature of the stock market is likened to a living creature, requiring continuous input and analysis of data for accurate predictions through consistent monitoring with BLSTM.

Wen et al. [27] presents a stock price forecasting model using Principal Component Analysis (PCA) and Long Short-Term Memory (LSTM). PCA reduces data correlation and dimensionality, while LSTM accurately forecasts stock prices. Experimental results on Pingan Bank data show improved accuracy compared to traditional models.

Simon et al. [28] focuses on using Artificial Neural Networks (ANN) as a dominant technique for accurate predictions. It reviews various ANN models and enhancement techniques, exploring research strategies to improve accuracy in stock market prediction.

Krollner et al. [29] interest in using machine learning for stock market prediction stems from its profit potential. Krollner et al. explores a less-explored aspect by applying ANNs in financial data mining for risk management. The goal is to leverage advances in stock market forecasting to develop hedging strategies safeguarding portfolios during market downturns. Simulation results demonstrate that ANNs offer a flexible and effective decision support tool for risk management in the Australian stock market.

B. Gaps and Fulfillment

Insufficient attention has been given to the examination of optimization techniques aimed at improving the efficacy of ML models in the context of stock price prediction within the current body of research. Many studies have primarily concentrated on experimental findings derived from specific datasets. However, there exists a dearth of comprehensive assessments encompassing diverse datasets that encompass various market conditions. To improve the performance of ML models, this research suggests integrating optimization techniques such as ABC, BBO, and AO. These techniques enhance the configuration of model parameters and enhance the precision of predictions. In order to make up for the insufficient assessment of real-world data, we conduct thorough testing on datasets that represent Google stock. The proposed model's generalizability and robustness can be guaranteed through this.

III. Methodology

A. Extreme Gradient Boosting Regression Method

A method of machine learning that uses gradient-boosting decision trees to provide regression and classification, which has been demonstrated in Fig. 1. To guarantee thorough coverage, the approach generates a sequence of weak learners, often in the form of classification regression trees. The model generates the final iteration of the regression model by averaging the weighted summation of the learners once the training procedure is complete. Regression trees are produced using the regularized and improved XGBoost regression algorithm, which helps predict continuous numerical target variables with accuracy. This approach follows the guidelines of gradient boosting, in which a new learner is introduced at construction time depending on the weak learner iteration's residual error. To reduce the overall model error, a gradient is used to generate the new learner. The XGBoost method is well known for its low computational complexity, fast speed, and great accuracy. Assume that the training sample data set D's class labels $\boldsymbol{y_i}$ and samples $\boldsymbol{x_i}$ are represented as follows:

$$D = \{(X_1, Y_1), (X_2, Y_2) \dots (X_N, Y_N)\} \tag{1}$$

Consequently, the i^{th} sample's prediction function may be shown as follows:

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i), f_k \in F \quad (2)$$

Here, $f_k(xi)$ is the discriminating function of the $\boldsymbol{K}$th tree for the ith data, and $\boldsymbol{F}$ is the resilient structure for the $\boldsymbol{k}$ choice tree model integrating. The enhanced tree model that XGBoost uses anticipates a value using a starting tree, calculates the value's deviation from the real value, and then inserts a tree to get the deviation.

when $k = 0, \hat{y}_i^{(0)} = 0$; when $k = 1, \hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)$; when we add t trees, then

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + \eta f_t(x_i), 0 < \eta < 1 \quad (3)$$

where, η is the learning rate.

The calculation of the objective function for XGBoost with the Mean Squared Error (MSE) loss function is as follows:

$$\text{Objective} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{t=1}^{T} \Omega(f_t) \quad (4)$$

Preventing overfitting is achieved by regularizing the second term, which is the mean squared error between the true labels y_i and the predictions $\hat{y}_i$.

B. Biogeography-Based Optimization

The BBO algorithm works by modeling the movement of various species, which depends on how well-suited their specific surroundings are, as Fig. 2 illustrates. This process makes it easier to examine the complex relationships that exist between different species and their environments, which in turn leads to more accurate and trustworthy predictions about ecological patterns. Within the framework of an optimization issue, it is possible to argue that a solution resembles a habitat. Creating densely populated habitats that offer more favorable living circumstances for a wider variety of species than other habitats is a useful tactic for handling population issues. The least desirable situation for a population is one in which living things must deal with significant issues. Because of this, the better solutions have the capacity to make the worse solutions pay attention by virtue of their shared characteristics. The sharing of attributes is accomplished by the application of the defined operators. Emigration rates show that migration is the process by which people move from a less favorable environment to a more favorable one. The assessment of a species' entrance into a certain geographic area is known as the immigration rate. When comparing a less favorable option to a more favorable one, the expected rate of migration is predicted to be higher.

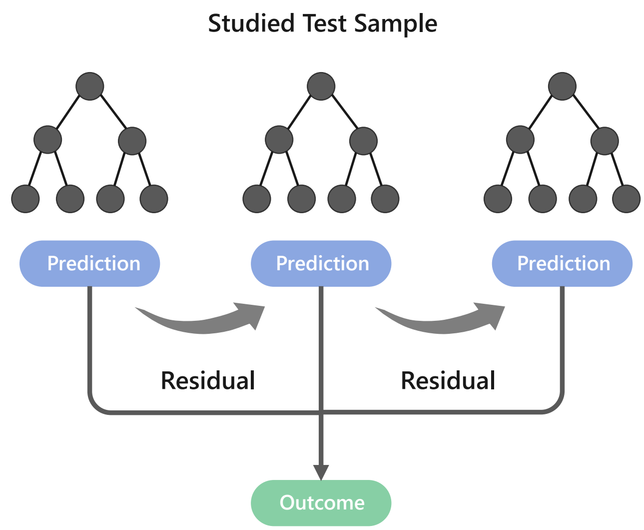

Fig. 1. The overall structure of the XGBoost.

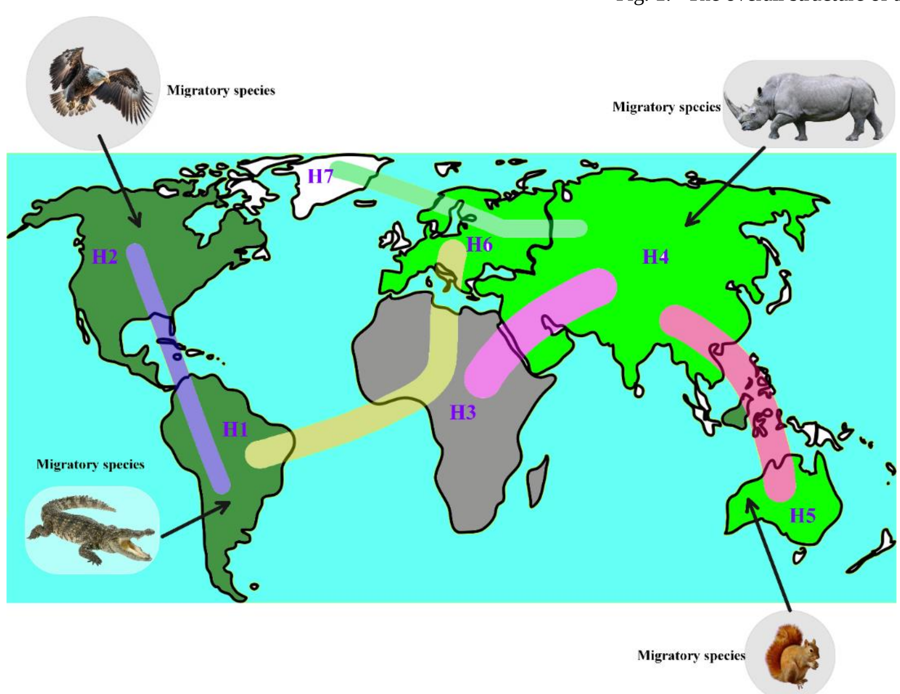

Fig. 2. Visual representation of the structure of BBO, habitats and migrations of species.

Conversely, the rate of migration pertains to the quantitative evaluation of the population size of individuals within a particular species who relocate from their habitat. As a result, it is anticipated that the rate of immigration will be higher when considering a poor solution in comparison to an optimal one. The BBO has employed the straight paths described by Eq. (5). The population size has been determined to be 100, and the epoch has been set to 500.

$$\mu_k = \frac{E \times k}{n} \lambda_k = I\left(1 - \frac{k}{n}\right) \tag{5}$$

where, μ_k is the migration amount of kth habitat, λ_k represents migration amount of kth habitat, I donates the maximum immigration amount, E is the maximum emigration rate, and n is the maximum amount of species that a habitat can support.

K: Number of species count.

C. Artificial Bee Colony

The ABC method is an optimization technique that draws inspiration from the intricate search and foraging activity of honey bees. It is commonly employed to address complicated issues with multiple dimensions and several modes in various domains of life. The ABC algorithm incorporates a nectar collection system within the honey bee colony, comprising nectar sources, employed bees, and unemployed bees. The unemployed bees are further categorized as bystanders and scouts. The task of the employed bees is to locate new food sources close to established ones. The observers must wait in the dance area and make a probabilistic decision regarding whether to select the food sources discovered by the employed bees, and the scouts must locate new food sources randomly [30]. During the process of nectar collection, certain workers within the beehive exhibit scouting behavior by engaging in continuous and indiscriminate search activities for food sources in close proximity to the hive. Subsequently, the researchers discovered two distinct food sources, labeled A and B. Following the observation of the waggle dance, a subset of the prospective worker bees transitioned into the role of hired bees and commenced their exploration of the aforementioned food sources. When a scout bee discovers a food source that surpasses a specific nectar threshold, it transitions into an employed bee. This bee subsequently engages in nectar collection and afterward returns to the beehive to deposit the gathered honey into the nectar store. When a scout bee discovers a food source that surpasses a specific nectar threshold, it transitions into an employed bee. This bee subsequently engages in nectar collection and afterward returns to the beehive to deposit the gathered honey into the nectar store. When a food supply isn't replenished many times, the worker bee transforms into a scout in search of fresh food. A detailed explanation of the mathematical models related to the several stages of the ABC algorithm can be found in the following section, and the visual representation of bees are shown in Fig. 3.

For starting food sources, the following equation is shown using a random solution vector boundary value. The population size is determined to be 100 and the epoch is 500.

$$x_{i,j} = x_j^{min} + rand(0,1)\left(x_j^{max} - x_j^{min}\right) \tag{6}$$

where, $i = 1, \ldots SN, j = 1, \ldots D, SN$ represents the number of solutions to be optimized, and D signifies the parameters to be optimized, where $x_j^{\min}$and $x_j^{\max}$signify the lower and upper limits of the $j\ th$ parameter, respectively. The aforementioned formula must be employed to ascertain the fitness value of each food source subsequent to the initiation of said food sources.

$$\text{fit}_{\ i} = \frac{1}{1 + obj \cdot fun_i} \tag{7}$$

$obj \cdot fun_i$, in this context, pertains to the intentional conduct.

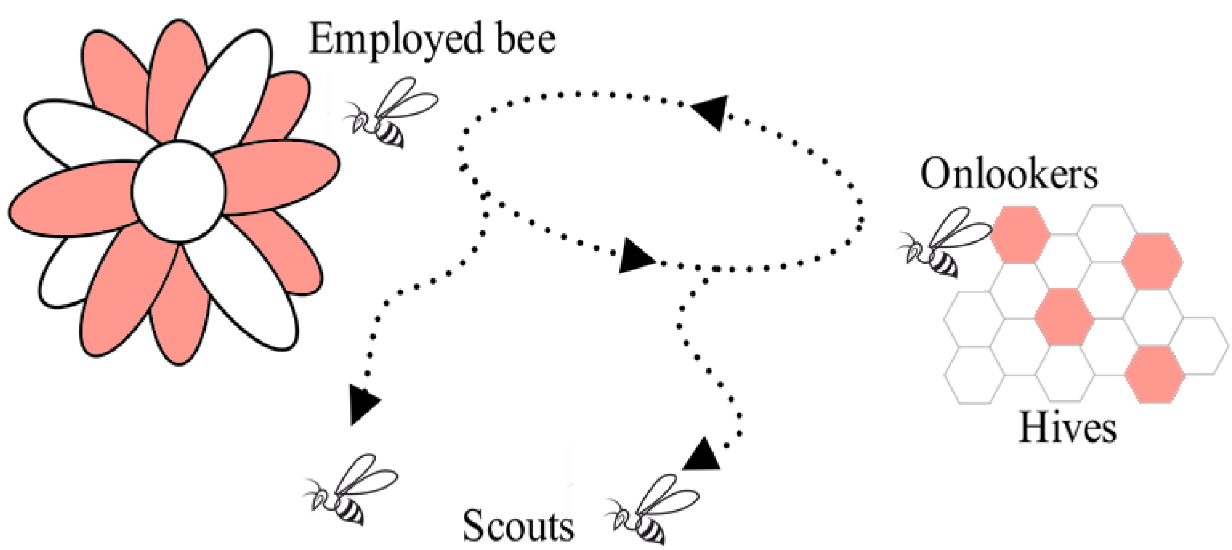

Fig. 3. The visual representation of the bees.

1) Employed bee stage: The quantity of responses matches the quantity of working bees and observers that SN has provided. Each active bee is supported by a single food source. By using Eq. (8) as a means to generate novel solutions, both observers and working bees engage in the exploration of proximate food sources and subsequently adjust their respective locations.

$$v_{ij} = x_{ij} + r_{ij}(x_{ij} - x_{kj}) \tag{8}$$

where, $k \neq i; r_{ij}$ is a randomly generated integer between [-1,1] and $j \in \{1,2,\dots,D\}$ and $k \in \{1,2,\dots,S\}$ are chosen at random. It is used to police a variety of communities. The fitness value of the new solution should be recalculated; subsequently, a comparison between the fitness values of $\boldsymbol{v_{ij}}$ and $\boldsymbol{x_{ij}}$ should be conducted, and the greater value should be selected.

2) Onlooker bee stage: As indicated by Eq. (9), observer bees evaluate a specific probability and measure of fitness while selecting food sources.

$$p_i = \frac{f_{it_i}}{\sum_{n=1}^{SN} fit_n} \tag{9}$$

where, fit_i is the solution's fitness value, which is connected to the food source's relevant nectar level. Eq. (7) illustrates that when sources of food are more appropriate, people are more inclined to choose them. The employed step goes to each food location and examines it, but the observer stage only applies Eq. (8) to the food sources chosen to produce fresh results. The spectator stage and the engaged stage are the same.

3) State – Scout: If the current food supply has been seen for a duration beyond the predetermined threshold, the employed bees will undergo a transformation into scout bees and start a stochastic exploration, as outlined in Eq. (6).

D. Aquila Optimization Algorithm

The AO was introduced in the year 2021 [24], as illustrated in Fig. 4, and the structure of the method is shown in Fig. 5. The following are the stages of the AO algorithm for establishing:

1) First-class vertical dives include: The Aquila species employs a strategic approach in identifying its prey area. It first undertakes a high-altitude flight to ascertain the optimal hunting region within the global context. This enables Aquila to effectively narrow down its search space and seek the most favorable solution. The procedure is seen in Eq. (10). The population size has been determined to be 100 individuals and the epoch has been set to 500.

$$\begin{cases} Z_1(t+1) = Z_{best}(t) \times \left(1 - \frac{t}{T}\right) \\ +\left(Z_M(t) - Z_{best}(t) \times rand\right) \\ Z_M(t) = \frac{1}{N}\sum_{i=1}^{N} Z_i(t), \\ \forall j = 1,2,\dots\dots, Dim \end{cases} \tag{10}$$

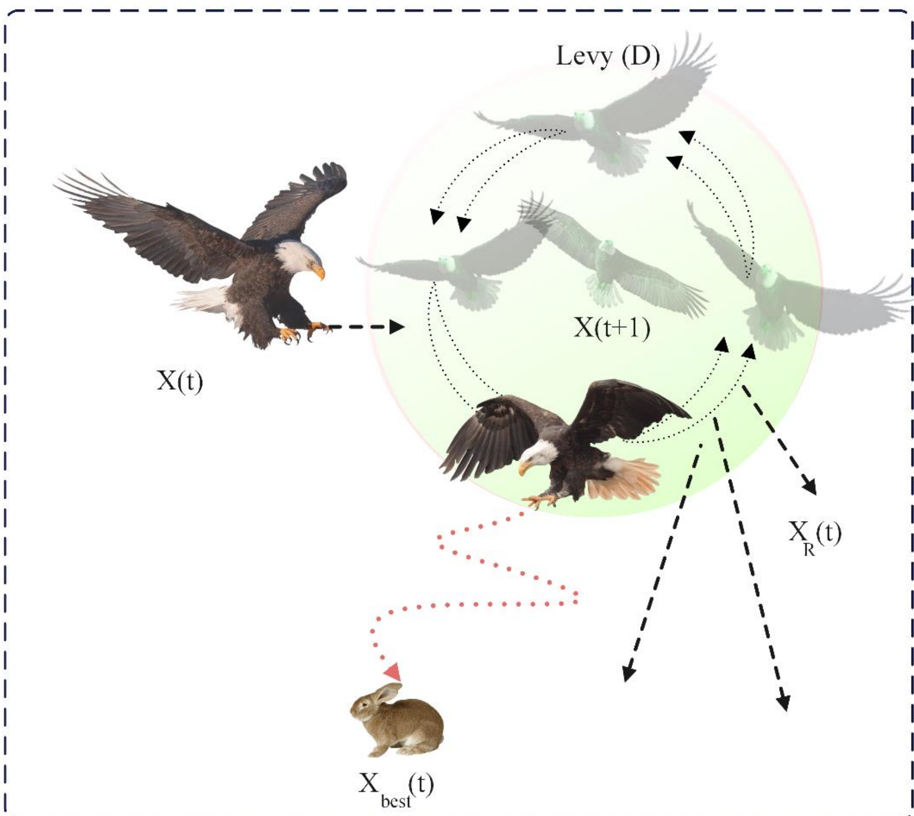

Fig. 4. Aquila optimization algorithm.

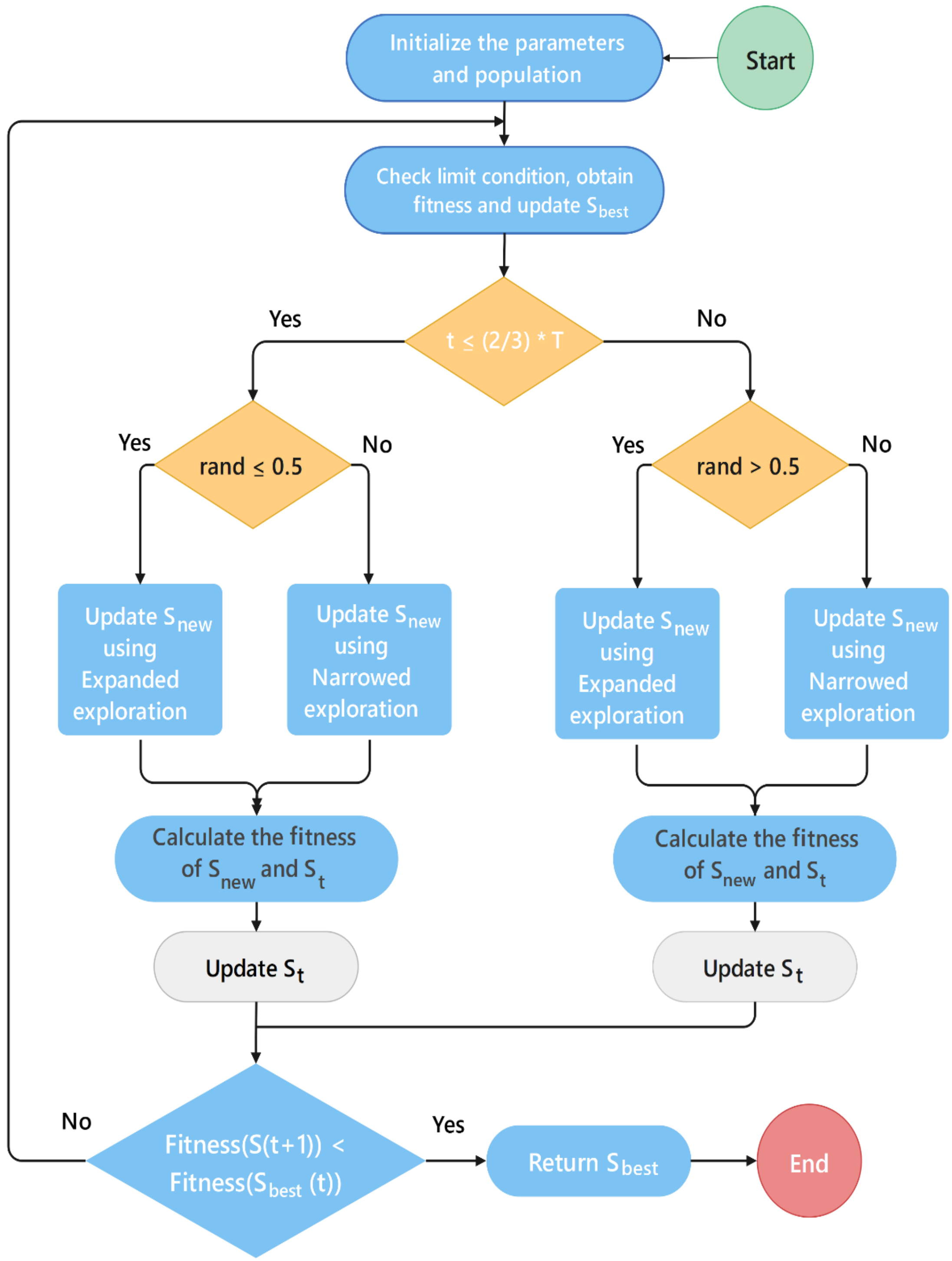

Fig. 5. Flowchart of the aquila optimization algorithm.

$Z(t+1)$ is the formula's generation $t+1$ solution, which was produced using the search procedure. The best strategy is Z_1, $Z_{\text{best}}(t)$, showing the position of the nearby prey target. Number t denotes the current iteration. T stands for the maximum number of iterations that may be carried out. Z(t) indicates the current solution's mean location at the t-th iteration. An integer between 0 and 1 is referred to as a rand. The subsequent swift glide attack: The eagle descends to a height where it can detect the prey region. It then hovers above the area it plans to hunt or investigate in search of the best solution.

2) Using formula (11):

$$\begin{cases} Z_2(t+1) = Z_{best}(t) \times Z(D) \\ +Z_R(t) + (y - z) \times rand \\ L(D) = s \times \dfrac{\mu \times \sigma}{|\nu|^{\frac{1}{\beta}}} \end{cases} \tag{11}$$

The random solution, denoted as $\boldsymbol{Z_R(t)}$, is a variable that takes on values between 1 and N, D represents the dimensional space, while L(D) refers to the hunting flight distribution function.

3) The third category of flights at low altitudes: Once the prey region has been precisely identified and the Aquila is prepared for landing and initiating an assault, it will assess the response of the prey inside the designated target area using a low-altitude and gradual descending approach, gradually closing in on the intended target. The procedure is demonstrated in Eq. (12).

$$\begin{aligned} Z_3(t+1) = \left(Z_{best}(t) - Z_M(T)\right) \times \alpha \\ - rand + \Big((U_b - L_b) \\ \times rand + L_b\Big) \times \delta \end{aligned} \tag{12}$$

The adjustment parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ are set at a lesser value of 0.1. The variables $\boldsymbol{L_b}$ and $\boldsymbol{U_b}$ denote the bottom and upper limits, correspondingly, of the given issue.

4) The fourth kind of walking capture: Upon the Aquila's approach to the designated target during the same day, it engages in an aerial assault on the prey, using its superior positioning and swiftly converging on its intended objective. The procedure is shown in Eq. (13).

$$\begin{cases} Z_4(t+1) = Q_F \times Z_{best}(t) - \\ \left(G_1 \times Z(t) \times rand\right) - \\ -G_2 \times L(D) + rand \times G_1 \\ Q_F(t) = \dfrac{2 \times rand - 1}{t^{(1-T)2}} \\ G_1 = 2 \times rand - 1 \\ G_2 = 2 \times \left(1 - \dfrac{t}{T}\right) \end{cases} \tag{13}$$

In this context, $\boldsymbol{c_F}$ denotes the quality function used to optimize the search strategy. $\boldsymbol{p_1}$ corresponds to the diverse movements executed by the Aquila throughout its prey-hunting activities, while $\boldsymbol{p_2}$ represents the flight slope of the Aquila during the hunting process. $\boldsymbol{Z(t)}$ denotes the current solution at the $\boldsymbol{t}$th iteration.

E. Dataset Description

Ensuring that the raw data is of exceptional quality is an essential first step towards gaining meaningful information. Data preparation is a necessary first step in order to do this. It involves many activities, including deleting irrelevant data, standardizing it for shared use, and organizing it to make it simple to retrieve important information. This is particularly relevant for large-scale data efforts when data quality is more important than quantity. Tasks related to data preparation may also involve encoding categorical data and scaling, standardizing, normalizing, and cleaning data in compliance with industry standards. By completing these procedures, analysts may improve the accuracy and reliability of the insights they get from the data. Min-Max scalers were used to scale and normalize the data, remove any possible inconsistencies, and measure the null, missing, and unknown values as part of the project's data pre-processing step. This method is shown using data from the Google stock. This data covers the period from 2015 to the middle of 2023 and goes through a number of preliminary procedures, including normalization.

Google stock data spanning a substantial timeframe is easily accessible and frequently examined, guaranteeing the dataset's resilience and superior quality. The availability of the hybrid AO-XGBoost model enables the training and testing processes. Google is widely recognized as a prominent technology corporation, and its stock performance serves as a reliable indicator of prevailing patterns within the technology industry. The examination of Google stock can yield significant insights into the intricacies of the technology market, thereby facilitating the development of a hybrid model centered on the prediction of stock market trends. Technology stocks, such as Google, frequently demonstrate elevated levels of volatility and substantial prospects for growth. Through the examination of Google stock data, the hybrid model is capable of capturing and acquiring knowledge from the intricate patterns and trends that are inherent in unpredictable markets. This is essential for improving the model's ability to make accurate predictions. Investors and analysts show significant interest in Google's stock performance due to its widespread recognition and following. Utilizing Google stock data to develop a hybrid AO-XGBoost model can effectively engage stakeholders who possess knowledge and interest in the company, thereby enhancing the model's attractiveness and pertinence. The prominent position of Google as a dominant player in the technology industry presents a valuable occasion to evaluate the efficacy of the hybrid AO-XGBoost model in comparison to a widely recognized stock. The validation of the model's efficacy and dependability in forecasting stock market fluctuations can be achieved by juxtaposing its predictions with the actual stock performance of Google. The hybrid AO-XGBoost model for stock market prediction can benefit from utilizing Google stock data for analysis and model development.

Several charts in Fig. 6 depict histograms illustrating the distribution of data points in specific bins. These bins are evenly distributed along the x-axis, with each bar's height indicating the number of data points in the corresponding bin. Employing histograms proves valuable for scrutinizing data distribution and recognizing patterns. Examining the frequency of each variable over time aids in understanding how the data changes and responds to external factors. The graphical representation of the data in these charts is easily interpretable, helping users make informed decisions based on the presented information.

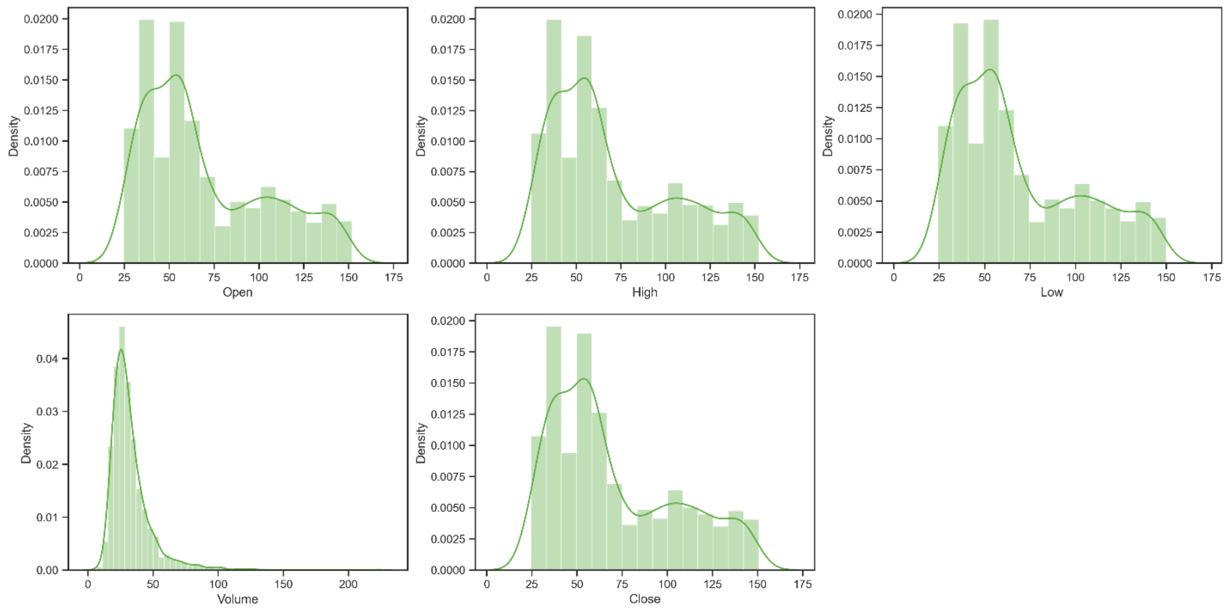

Fig. 6. Display density of variables.

Feature scaling is the process of rescaling numerical features in a dataset to a specified range, usually between zero and one. It is sometimes referred to as data preparation or Min-Max normalization. As all the characteristics are brought to the same scale, the goal is to preserve the relative relationships between the values. This could be especially important for ML algorithms that depend on the amount of input data.

$$\text{XScaled} = \frac{(X - Xmin)}{(Xmax - Xmin)} \quad (14)$$

where, X denotes the feature's initial value that you wish to normalize, X_{min} denotes the feature's lowest value in the dataset, and X_{max} denotes the feature's maximum value in the dataset.

1) Candlestick description: The Candlestick chart originated in the 16th century and was created by a Japanese rice merchant who dealt with financial instruments. The chart is a hybrid of a line chart and a bar chart, where each bar visually reflects the price range within a certain time period. It is mostly used in the technical analysis of patterns in stock and currency prices. A candlestick is constructed using the open, high, low, and closing prices of the day as Fig. 7 describes the overall structure. A full candlestick is drawn when the open price is higher than the closing price.

2) Statistical result: The statistical properties are shown in Table I and include variance, skewness, kurtosis, mean, minimum, maximum, and standard deviation (Std.).

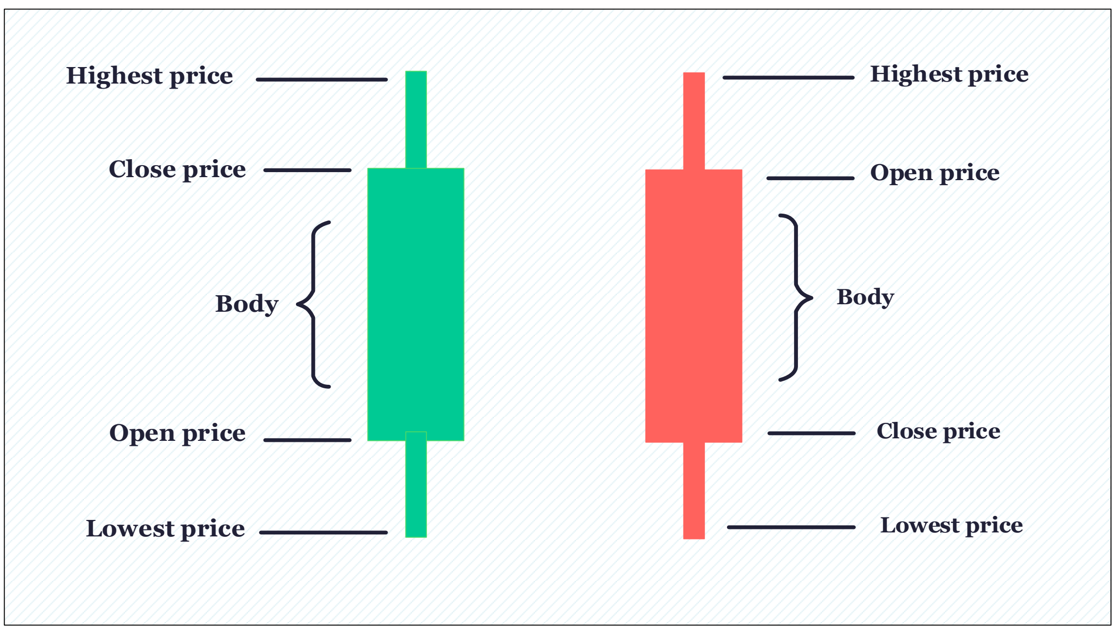

Fig. 7. Candlestick description.

TABLE I. STATISTICAL RESULTS OF THE PRESENTED DATASET

	Open	High	Low	Volume	Close
Mean	70.04	70.80	69.33	32.59	70.09
Std.	34.54	34.97	34.15	15.60	34.55
Minimum	24.65	24.72	24.31	6.93	24.55
Maximum	151.85	152.10	149.88	223.29	150.70
Skewness	0.74	0.73	0.74	2.87	0.73
kurtosis	-0.62	-0.65	-0.61	16.57	-0.63
variance	1193.42	1223.37	1165.98	243.54	1194.32

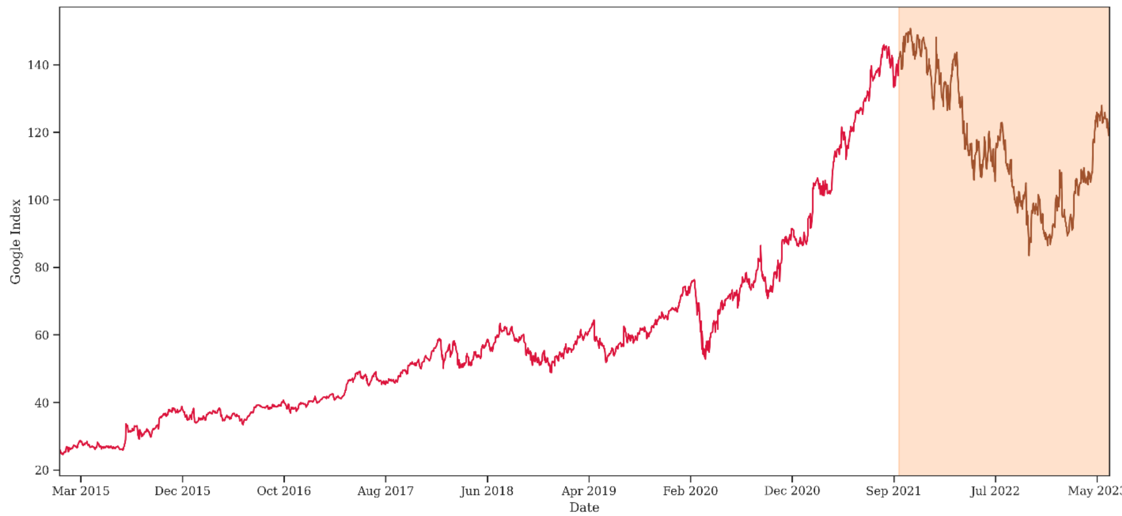

Fig. 8. Separation of the dataset into train and test.

The closing price data, which has been split into training and testing zones, is shown in Fig. 8. This approach guarantees data accuracy while assisting customers in getting accurate insights.

F. Evaluation Metrics

The following mathematical formulae were used to determine the performance metrics used in the current work: mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared error (MSE), and coefficient of determination (R^2).

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{16}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{17}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{18}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{19}$$

In which, $\bar{y}$ is the mean value, y_i indicates the true value, and $\hat{y}_i$ provides the projected value. n denotes the stock series data length.

G. Detailed Description

The data is gathered from 2015 to 2023. The unprocessed data is standardized to ensure that all characteristics are on a same scale. This aids in mitigating the algorithm's inclination towards favoring a certain trait. The normalized data is partitioned into two distinct sets: an 80% training set and a 20% test set. The training set is used to instruct the machine learning model, whilst the test set is employed to assess its performance. The XGBoost method is selected as the ML model for forecasting the closing price. XGBoost is a robust ensemble approach based on trees that has shown effectiveness in a wide range of prediction applications. The hyperparameters of the XGBoost model are fine-tuned to enhance its performance. Hyperparameters are the configuration options that govern the behavior of a model, and identifying the optimal combination of hyperparameters may have a substantial influence on the accuracy of the model. The performance of the trained AO-XGBoost model is assessed using four assessment measures. The AO-XGBoost model, which has been trained, is used to forecast the future closing price of Google stock. The forecast is derived from the analysis of past data and the acquired trends.

IV. RESULT AND DISCUSSION

A. Tuning of the Hyperparameters

Table II provides a comprehensive overview of the hyperparameter configurations for XGBoost, utilizing three distinct optimization algorithms: AO, ABC, and BBO. Each row in the table represents a distinct hyperparameter, while each column corresponds to a particular optimizer. This hyperparameter specifies the quantity of boosting rounds or trees to construct. The range for all three optimizers (AO, ABC, BBO) is 100 to 5000. Each optimizer employs a distinct value within this range: 500 for AO, 500 for ABC, and 300 for BBO. Gamma serves as a regularization parameter that governs the level of complexity exhibited by the trees. To establish a new partition on a leaf node, the smallest loss reduction is required. The gamma values range from 0 to 10. The chosen values for each optimizer are 4.967375 for AO, 7.071759 for ABC, and 5.612515 for BBO. The L1 regularization term on weights is referred to as a hyperparameter. It promotes the reduction of weight vectors. The range of values for the variable "reg alpha" is between 0 and 5. The chosen values for each optimizer are as follows: 0.457479 for AO, 0.646517 for ABC, and 0.541563 for BBO. The term "Reg lambda" refers to the L2 regularization term applied to the weights. The term is also referred to as the Ridge regularization term. The permissible values for the regularization parameter lambda range from 0 to 1. The chosen values for each optimizer are as follows: 0.799153 for AO, 0.677998 for ABC, and 0.87415 for BBO. This hyperparameter adjusts the weight or influence of each tree in the model. It is alternatively referred to as eta. The learning rate is bounded between 0.0001 and 1. The selected values for each optimizer are 0.001 for AO, 0.0001 for ABC, and 0.01 for BBO.

B. *Comparative Analysis*

To assess the effectiveness of the presented models, a range of common measures was employed. These measures included MSE, MAPE, R^2, and MAE. These metrics offer a comprehensive overview of the prediction precision of the methods. A detailed summary of the performance metrics for four models, namely XGBoost, BBO-XGBoost, ABC-XGBoost, and AO-XGBoost, is presented in Table III. The development and evaluation of these models were based on historical stock price data for a Google stock market, spanning from the start of 2015 to the middle of 2023. This dataset was chosen to provide a comprehensive evaluation of the model's performance over a period of several years.

The results shown in Table III indicate that the AO-XGBoost model outperforms the other models in terms of forecasting accuracy; the model's success can be seen in the relatively low values for MSE, MAPE, and MAE, which show that the model was able to capture the complex temporal patterns and correlations found in stock price data. These outcomes imply that the AO-XGBoost model is a potentially useful tool for forecasting future market trends and helping investors make well-informed decisions, as it has the closest estimation curve to the actual index shown in Fig. 9 and Fig. 10. Through a comparative analysis of the four models represented by Table III, it is evident that the AO approach yielded the greatest results when it came to optimizing the model's hyperparameters among the optimization techniques utilized. BBO and ABC were the next top-performing techniques. The classification of the BBO-XGBoost, ABC-XGBoost, and AO-XGBoost models' findings is 0.966, 0.972, and 0.982, indicating an improvement in the model's performance.

These results have many practical implications for investors and financial institutions. Owing to its improved performance, AO-XGBoost could be a helpful tool for short-term stock price prediction, empowering traders to make more informed trading choices. The feature importance study also emphasizes how important it is to include historical data and sentiment analysis in prediction models. Given that these models rely on previous data, stock markets may be impacted by a variety of unanticipated events. Future research may look at using real-time data and external factors, including economic indicators and current events, to further increase prediction accuracy.

In conclusion, our research demonstrates the usefulness of machine learning models, specifically XGBoost, for stock market forecasting. Investors who understand the importance of certain features and model performance metrics might lower their risks in the erratic world of stock trading. Fig. 11 and Fig. 12 show the models' training and testing outcomes for each model.

TABLE II. SETTING OF THE HYPERPARAMETERS

XGBoost		AO	ABC	BBO
Numbers of estimators	[100, 5000]	500	500	300
gamma	[0,10]	4.967375	7.071759	5.612515
Reg alpha	[0, 5]	0.457479	0.646517	0.541563
Reg lambda	[0, 1]	0.799153	0.677998	0.87415
Learning rate	[0.0001, 1]	0.001	0.0001	0.01

TABLE III. THE OUTCOMES OF THE PERFORMANCE CRITERIA FOR MODELS

MODEL/Metrics	**TRAIN SET**				**TEST SET**			
	R^2	***MAPE***	***MAE***	***MSE***	R^2	***MAPE***	***MAE***	***MSE***
XGBoost	0.960	3.225	2.729	29.045	0.955	2.719	3.182	14.877
BBO-XGBoost	0.982	3.748	2.427	13.105	0.966	2.314	2.679	11.166
ABC-XGBoost	0.985	2.054	1.702	11.161	0.972	2.150	2.438	9.213
AO-XGBoost	0.990	3.007	1.803	7.250	0.982	1.721	1.994	5.946

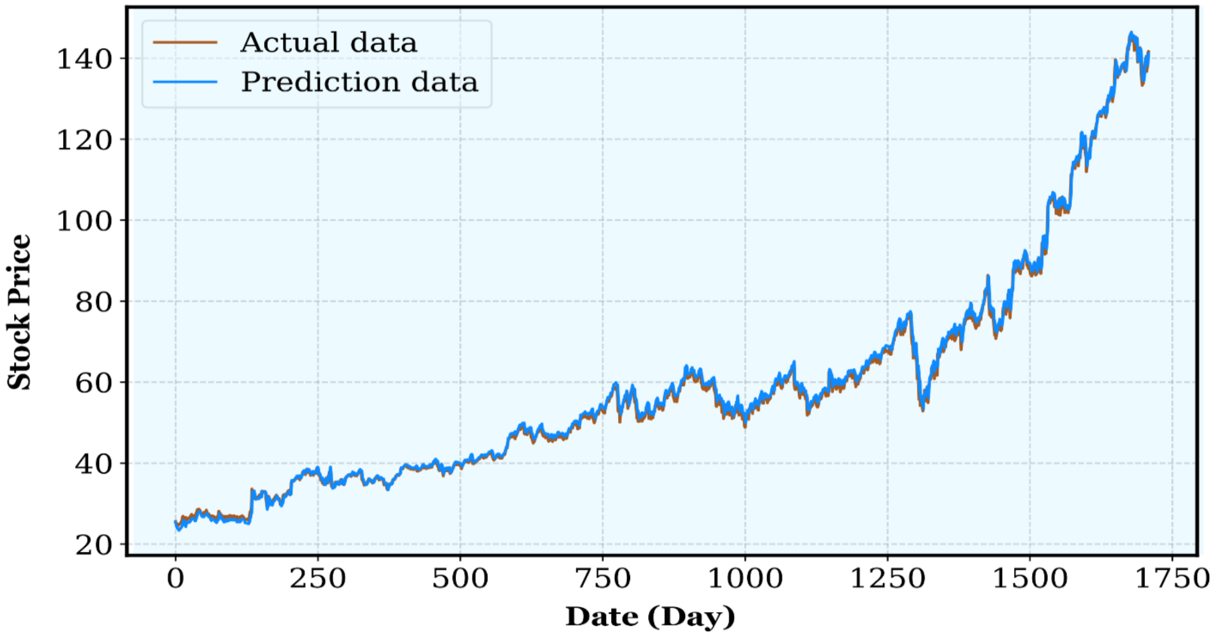

Fig. 9. AO-XGBoosts performance during training in contrast to the other models.

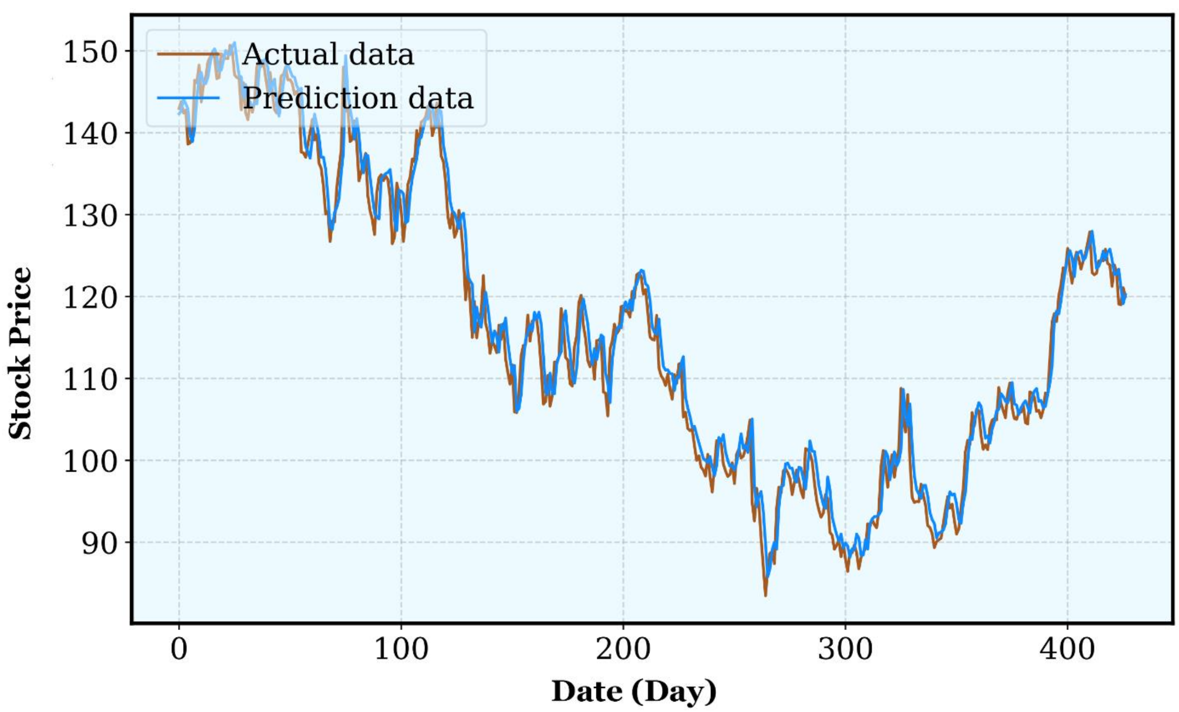

Fig. 10. AO-XGBoost's performance during testing in contrast to the other models.

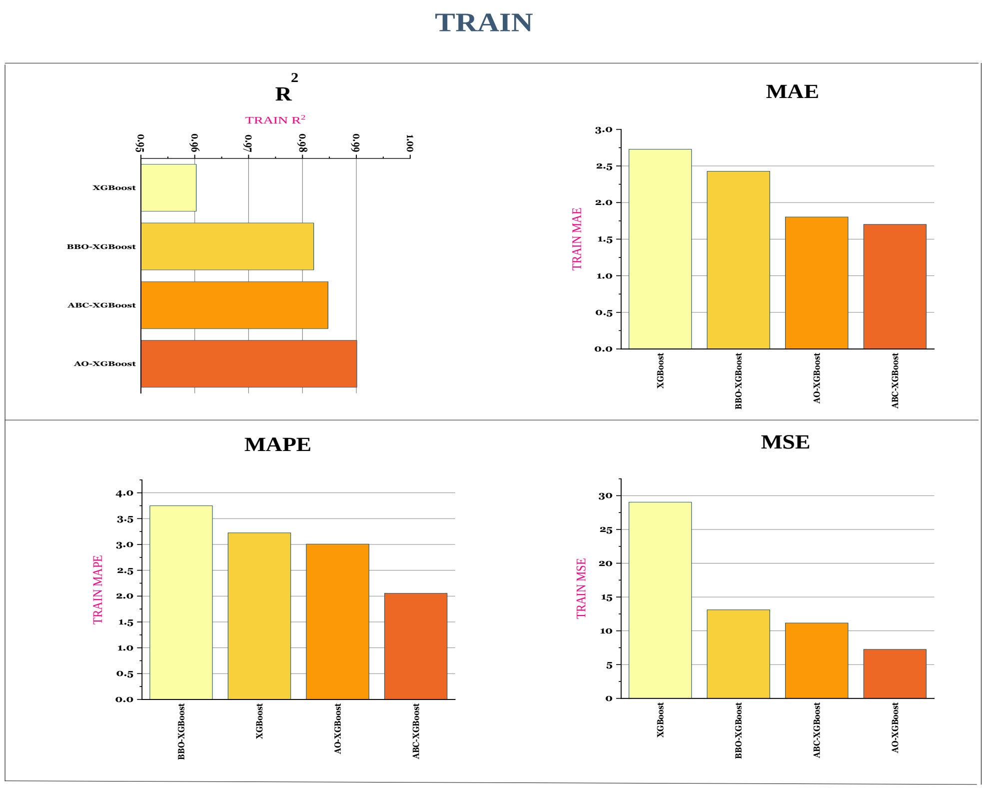

Fig. 11. The train's assessment criteria's outcome.

Fig. 12. The test's assessment criteria's outcome.

TABLE IV. THE MODEL COMPARISON ANALYSIS TO PRIOR STUDIES

References	**Method**	R^2
[31]	CEEMDAN-LSTM	0.9031
	LSTM	0.6896
	EMD-SC-LSTM	0.9111
	EMD-LSTM	0.8703
	SC-LSTM	0.6871
	CEEMDAN-SC-LSTM	0.9206
[32]	MLS-LSTM	0.95
	SVM	0.93
	Linear regression	0.73
Present work		0.982

The results presented in Table IV indicate that the AO-XGBoost model exhibits superior performance in the field of stock price prediction when compared to other evaluated models. Our study revealed a remarkable R^2 value of 0.982, which outperformed other evaluated models including CEEMDAN-LSTM, LSTM, EMD-SC-LSTM, EMD-LSTM, SC-LSTM, CEEMDAN-SC-LSTM, SVM, Linear Regression, and MLS-LSTM. This observation suggests a higher degree of correlation between the projected and observed stock prices, implying that the model exhibits greater efficacy in capturing the fundamental patterns within the data. The AO-XGBoost model acquires the ability to dynamically modify its parameters during training by incorporating adaptive optimization techniques. This capability enables the system to adjust to the intricacies of the data and enhance its capacity to manage fluctuations and intricacies within the stock market. The utilization of the XGBoost ensemble learning algorithm is effective in capturing nonlinear relationships and interactions within the data. This algorithm combines the strengths of gradient boosting with tree-based models. The utilization of ensemble learning in XGBoost, wherein multiple weak learners are combined to form a strong learner, serves to augment the predictive accuracy and robustness of the model. Furthermore, the regularization techniques employed by XGBoost serve to mitigate the issue of overfitting and enhance the overall generalization performance. The inclusion of this feature guarantees the model's ability to generalize to unfamiliar data, which is a critical factor in predicting stock market trends where data distributions may vary over time. In addition, the scalability and efficiency of XGBoost render it well-suited for the management of extensive datasets and the execution of real-time prediction tasks. Due to its distributed computing capabilities, parallel processing is facilitated, leading to enhanced efficiency in both training and inference processes. The AO-XGBoost model is highly suitable for real-time applications that require prompt predictions. Furthermore, the accessibility of the model is improved by the clarity of XGBoost's decision-making process.

The AO-XGBoost model has demonstrated promising accuracy in predicting short-term stock prices. These forecasts can assist investors in maximizing profits and mitigating risks in the stock market. It may be advantageous to incorporate high-performing XGBoost models such as BBO-XGBoost or ABC-XGBoost into algorithmic trading systems. These systems can autonomously execute buy or sell orders on the predictions made by the model, thereby enabling traders to take advantage of short-term market fluctuations. It may be worthwhile to investigate risk management solutions for financial institutions that utilize XGBoost models to forecast market trends. Accurate forecasts of stock prices and market volatility have the potential to assist institutions in evaluating and mitigating risks associated with their investment portfolios. It is advisable to develop investor portfolio management tools that utilize XGBoost models to enhance asset allocation and diversification. XGBoost stock price forecasts can be utilized by investors to construct portfolios that optimize returns and mitigate market risk. An alternative approach entails the utilization of XGBoost models in financial advisory services to generate tailored investment recommendations. Financial advisors can customize guidance based on individual client's investment objectives and risk tolerance through the examination of past stock market data and the utilization of sophisticated machine learning techniques. It may be intriguing to develop trading signal platforms that offer real-time buy or sell signals using XGBoost models. Traders can subscribe to these platforms to receive timely notifications and alerts, enabling them to make well-informed trading decisions in rapidly changing markets. Enhancing the accuracy, robustness, and scalability of XGBoost stock market forecasting models through research and development could yield advantageous outcomes. Enhancing machine learning methods enables researchers to enhance predictive models for intricate market dynamics and patterns. These illustrations indicate that XGBoost models have the potential to be employed in various domains such as stock market prediction, investment tactics, algorithmic trading, risk mitigation, and financial advisory consulting. The predictive capabilities of XGBoost models can assist stakeholders in making more informed decisions and effectively navigating the stock market.

V. CONCLUSION

The trading system may provide profitable buy, sell, and hold recommendations. Traders and investors may benefit from the forecasts. The successful model from this study is intended to be applied to the dynamic trading system. Either an automated stock market system or a dynamic trading system may include this method. Thus, it is possible to forecast stock prices in the future with little inaccuracy. In this study, a novel model XGBoost network optimized using the AO is proposed. The AO-XGBoost is the suggested model. The suggested model seeks to forecast stock market values. Stock market prices are predicted using data from Google stock prices. The dataset covers the months of 2015 to mid-2023. The data is separated into 80% train data and 20% test data after normalization. With parameters, a new XGBoost network is constructed. The AO algorithm variables are linked to the parameters. The design of the XGBoost adapts to changes in the variables. To get the best outcomes, AO selects the optimal design. The newly suggested model is compared to the XGBoost, ABC-XGBoost, and BBO-XGBoost networks in order to assess its quality. The results unequivocally demonstrate that, out of all the models, the AO-XGBoost model is the best. It also provides very accurate forecasts.

Although the proposed AO-XGBoost model has demonstrated impressive performance in predicting stock prices, it is important to take into account several limitations. The utilization of historical data derived from Google stock prices within a designated time period may not comprehensively capture the complexities and fluctuations inherent in the wider stock market. Furthermore, although optimization techniques such as BBO, ABC, and AO strive to improve the model's ability to adjust, there is a possibility of overfitting the training data, which could undermine its ability to apply the model to new market conditions. Furthermore, the model's susceptibility to parameter selection and the computational burden involved in determining optimal values present practical obstacles. Moreover, the model's forecasting accuracy is limited by the inherent unpredictability of market dynamics, which is influenced by various factors including geopolitical events and investor sentiments. The implementation and interpretation of the AO-XGBOOST model in real-world trading environments may be impeded by its complexity, thereby restricting its applicability for practitioners with diverse levels of technical proficiency. Therefore, it is imperative to recognize and tackle the limitations of the AO-XGBoost model in order to ensure its successful implementation in dynamic financial markets, despite the promising insights it provides.

References

[1] S. Asadi, E. Hadavandi, F. Mehmanpazir, and M. M. Nakhostin, "Hybridization of evolutionary Levenberg–Marquardt neural networks and data pre-processing for stock market prediction," Knowl Based Syst, vol. 35, pp. 245–258, 2012, doi: https://doi.org/10.1016/j.knosys.2012.05.003.

[2] S. Gupta, V. Tyagi, M. Aadil, and N. Kumari, "Forecasting Stock Market's Performance Based on Grasshopper Optimized Hybrid Neural Network Method," International Journal of Intelligent Systems and Applications in Engineering, vol. 11, no. 8s, pp. 164–170, 2023.

[3] S. Gupta, V. Tyagi, M. Aadil, and N. Kumari, "Forecasting Stock Market's Performance Based on Grasshopper Optimized Hybrid Neural Network Method," International Journal of Intelligent Systems and Applications in Engineering, vol. 11, no. 8s, pp. 164–170, 2023.

[4] Y. Xu, J. Liu, F. Ma, and J. Chu, "Liquidity and realized volatility prediction in Chinese stock market: A time-varying transitional dynamic perspective," International Review of Economics and Finance, vol. 89, no. PA, pp. 543–560, 2024, doi: 10.1016/j.iref.2023.07.083.

[5] S. Akhter and M. A. Misir, "Capital markets efficiency: evidence from the emerging capital market with particular reference to Dhaka stock exchange," South Asian Journal of Management, vol. 12, no. 3, p. 35, 2005.

[6] K. Miao, F. Chen, and Z. G. Zhao, "Stock price forecast based on bacterial colony RBF neural network," Journal of Qingdao University (Natural Science Edition), vol. 2, no. 11, 2007.

[7] J. Lehoczky and M. Schervish, "Overview and History of Statistics for Equity Markets," Annu Rev Stat Appl, vol. 5, pp. 265–288, 2018, doi: 10.1146/annurev-statistics-031017-100518.

[8] A. Aali-Bujari, F. Venegas-Martínez, and G. Pérez-Lechuga, "Impact of the stock market capitalization and the banking spread in growth and development in Latin American: A panel data estimation with System GMM," Contaduría y administración, vol. 62, no. 5, pp. 1427–1441, 2017.

[9] M. P. Naeini, H. Taremian, and H. B. Hashemi, "Stock market value prediction using neural networks," in 2010 international conference on computer information systems and industrial management applications (CISIM), IEEE, 2010, pp. 132–136.

[10] B. Qian and K. Rasheed, "Stock market prediction with multiple classifiers," Applied Intelligence, vol. 26, no. 1, pp. 25–33, 2007, doi: 10.1007/s10489-006-0001-7.

[11] E. S. Olivas, J. D. M. Guerrero, M. Martinez-Sober, J. R. Magdalena-Benedito, and L. Serrano, Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques: Algorithms, methods, and techniques. IGI global, 2009.

[12] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, "Machine learning techniques and data for stock market forecasting: A literature review," Expert Syst Appl, vol. 197, no. February, 2022, doi: 10.1016/j.eswa.2022.116659.

[13] J. Grudniewicz and R. Ślepaczuk, "Application of machine learning in algorithmic investment strategies on global stock markets," Res Int Bus Finance, vol. 66, 2023, doi: 10.1016/j.ribaf.2023.102052.

[14] H. N. Bhandari, B. Rimal, N. R. Pokhrel, R. Rimal, K. R. Dahal, and R. K. C. Khatri, "Predicting stock market index using LSTM," Machine Learning with Applications, vol. 9, no. February, p. 100320, 2022, doi: 10.1016/j.mlwa.2022.100320.

[15] H. N. Bhandari, B. Rimal, N. R. Pokhrel, R. Rimal, K. R. Dahal, and R. K. C. Khatri, "Predicting stock market index using LSTM," Machine Learning with Applications, vol. 9, no. May, p. 100320, 2022, doi: 10.1016/j.mlwa.2022.100320.

[16] B. Mahesh, "Machine learning algorithms-a review," International Journal of Science and Research (IJSR).[Internet], vol. 9, no. 1, pp. 381–386, 2020.

[17] T. Chen et al., "Xgboost: extreme gradient boosting," R package version 0.4-2, vol. 1, no. 4, pp. 1–4, 2015.

[18] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Ann Stat, pp. 1189–1232, 2001.

[19] D. Simon, "Biogeography-based optimization," IEEE Transactions on Evolutionary Computation, vol. 12, no. 6, pp. 702–713, 2008, doi: 10.1109/TEVC.2008.919004.

[20] A. K. Bansal, V. S. Sangtani, P. Dadheech, N. Aneja, and U. Yahya, "Biogeography-based Optimization of Artificial Neural Network (BBO-ANN) for Solar Radiation Forecasting," Applied Artificial Intelligence, vol. 37, no. 1, 2023, doi: 10.1080/08839514.2023.2166705.

[21] V. Garg, K. Deep, K. A. Alnowibet, H. M. Zawbaa, and A. W. Mohamed, "Biogeography Based optimization with Salp Swarm optimizer inspired operator for solving non-linear continuous optimization problems," Alexandria Engineering Journal, vol. 73, pp. 321–341, 2023, doi: https://doi.org/10.1016/j.aej.2023.04.054.

[22] F. Y. Partovi and M. Anandarajan, "Classifying inventory using an artificial neural network approach," Comput Ind Eng, vol. 41, no. 4, pp. 389–404, 2002, doi: https://doi.org/10.1016/S0360-8352(01)00064-X.

[23] D. Karaboga, "Artificial bee colony algorithm," scholarpedia, vol. 5, no. 3, p. 6915, 2010.

[24] L. Abualigah, D. Yousri, M. Abd Elaziz, A. A. Ewees, M. A. A. Al-qaness, and A. H. Gandomi, "Aquila Optimizer: A novel meta-heuristic optimization algorithm," Comput Ind Eng, vol. 157, p. 107250, 2021, doi: https://doi.org/10.1016/j.cie.2021.107250.

[25] 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.

[26] S. Hong and J. Han, "Stock Price Prediction by Using BLSTM (Bidirectional Long Short Term Memory)," Journal of Computational and Theoretical Nanoscience ; volume 18, issue 5, page 1614-1617 ; ISSN 1546-1955, 2021, doi: 10.1166/jctn.2021.9603.

[27] Y. Wen, P. Lin, and X. Nie, "Research of Stock Price Prediction Based on PCA-LSTM Model," IOP Conference Series: Materials Science and Engineering ; volume 790, issue 1, page 012109 ; ISSN 1757-8981 1757-899X, 2020, doi: 10.1088/1757-899x/790/1/012109.

[28] S. Simon and A. Raoo, "Accuracy Driven Artificial Neural Networks in Stock Market Prediction," Jan. 2012, [Online]. Available: https://zenodo.org/record/7389464.

[29] B. Krollner, B. J. Vanstone, and G. Finnie, "Risk management in the Australian stock market using Artificial Neural Networks," Krollner , B , Vanstone , B J & Finnie , G 2012 , ' Risk management in the Australian stock market using Artificial Neural Networks ' , Australian Journal of Intelligent Information Processing Systems , vol. 13 , no. 2 , pp. 1-12 ., 2012, [Online]. Available: https://research.bond.edu.au/en/publications/cedf77f8-58cc-41d4-979b-d9baad2ad936.

[30] 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.

[31] R. Zhu, G.-Y. Zhong, and J.-C. Li, "Forecasting price in a new hybrid neural network model with machine learning," Expert Syst Appl, vol. 249, p. 123697, 2024, doi: https://doi.org/10.1016/j.eswa. 2024.123697.

[32] A. Q. Md et al., "Novel optimization approach for stock price forecasting using multi-layered sequential LSTM," Appl Soft Comput, vol. 134, p. 109830, 2023, doi: https://doi.org/10.1016/j.asoc. 2022.109830.