Presenting an Optimized Hybrid Model for Stock Price Prediction

Liangchao LIU
College of Economics and Management, Jiaozuo University, Henan Jiaozuo, 454000, China

***Abstract*—In the finance sector, stock price forecasting is deemed crucial for traders and investors. In this study, a detailed comparison and analysis of various machine learning models for stock price forecasting were undertaken. Historical stock data and an array of technical indicators were utilized in these models. The enhancement of the Histogram-Based Gradient Boosting (HGBR) method for predicting the Nasdaq stock index was the focus. Optimization techniques such as genetic algorithm optimization, biologically-based optimization, and the grasshopper optimization algorithm were applied. Among these, the most promising results were shown by the grasshopper optimization method. The optimized HGBR models, namely GA-HGBR, BBO-HGBR, and GOA-HGBR, were found to have achieved significant improvements, with coefficient of determination values of 0.96, 0.98, and 0.99, respectively. These figures underscore the substantial advancement of these models as compared to the baseline HGBR model. Metrics such as Mean Absolute Error, Root Mean Square Error, Mean Absolute Percentage Error, and the Coefficient of Determination were employed to assess the performance of the models.**

Keywords*—Stock prediction; machine learning approaches; ensemble learning; grasshopper optimization; histogram-based gradient boosting

I. Introduction

The task of predicting stock prices is undeniably challenging, primarily due to the inherent long-term uncertainties involved [1]. The traditional market hypothesis suggests that stock prices are unpredictable and random, but current technical analysis has revealed that previous records hold valuable information that can assist in predicting future stock values [2]. Furthermore, factors such as political developments, general economic conditions, commodity prices, investor expectations, and other stock market movements can also have a significant impact on the stock market [3]. High market capitalization is utilized to calculate stock group values, and a variety of technical factors can be used to generate statistical data on stock prices [4]. Therefore, it is essential to consider all of these factors when attempting to predict stock prices accurately. Inherent challenges arise when attempting to anticipate stock values because many traditional techniques for doing so rely on stagnant trends.

Furthermore, because there are so many variables at play, forecasting stock values is inherently difficult. The market operates like a voting machine in the near term but more like a weighing machine in the long term, indicating the possibility of forecasting longer-term market changes [5]. Machine learning (ML) is a potent technology that includes a variety of algorithms and has been shown to improve performance in particular case studies considerably. Many people think that ML has the ability to find important information and recognize patterns in datasets [6]. Ensemble models are a machine learning strategy where common algorithms are used to handle a particular problem, in contrast to standard ML approaches, and they have consistently shown higher performance when it comes to time series prediction [7][8][9].

In the field of forecasting, utilizing ensemble approaches has been found to yield more accurate results compared to single models [10]. The reason behind this is that ensembles are able to combine the predictions of multiple models, enabling them to account for potential errors and uncertainties. One of the major challenges in machine learning is overfitting, which occurs when a model performs exceedingly well on training data but fails to generalize to new data. However, ensembles are less prone to overfitting due to their reliance on multiple base models, such as bagging and boosting, which help to mitigate the risk of overfitting [11]. These techniques work by creating multiple models and combining their predictions, thereby reducing the likelihood of a single model overfitting to the training data. Ultimately, the use of ensemble approaches in forecasting can lead to more reliable and accurate predictions [12]. To conduct a comparative analysis of cutting-edge machine learning methods for forecasting stock market returns, ten years of daily historical data pertaining to the top ten equities on the Casablanca Stock Exchange were utilized. When Bilal et al. [13] used an ensemble learning approach was utilized to train six classifiers (ridge regression, LASSO regression, support-vector machine, k-nearest neighbors, random forest, and adaptive boosting) to forecast price directions one day, one week, and one month in advance. In contrast to other models, support vector machines, random forests, and adaptive boosting exhibited superior performance in short-term predictions. Ensemble learning enhanced performance metrics across all prediction horizons by a substantial margin. Sonkavde et al. [14] investigated a range of algorithms to address challenges related to stock price prediction and classification. These algorithms comprised ensemble algorithms, deep learning, supervised and unsupervised machine learning, and time series analysis.

The model presented for forecasting the Nasdaq stock market in this work is a Histogram gradient boosting regressor (HGBR). Nasdaq is one of the major stock exchanges in the United States, particularly associated with technology and internet-based businesses, and renowned for its electronic trading platform. The HGBR is a machine-learning approach that addresses regression-related issues by combining the principles of gradient boosting with histogram-based feature

splitting. It is an adaptation of the popular Gradient Boosting Machine (GBM) technique [15]. Regression and classification are the two primary subtypes of gradient boosting, a machine-learning approach for prediction. This paradigm is intended to manage complicated and substantial difficulties as opposed to simple and small ones, in contrast to previous techniques. The gradient-boosting technique known as HGBR was created expressly to overcome regression issues. This technique is renowned for its quickness and capacity to hasten decision-tree learning. By discretizing the input variables, HGBR does this by splitting extra trees into several values [16].

Providing precise forecasts is the primary goal of prediction models. To achieve this, optimizing these models can significantly improve their accuracy, especially in sectors where even a minor increase in accuracy can have a substantial impact, such as healthcare, banking, and manufacturing [17]. Different methods and models are provided to optimize the HGBR. Some of them, like Moth flame optimization [18], Biogeography-based optimization [19] and gray wolf optimization [20], are inspired by nature. The optimization methods used to optimize for the model of this paper are genetic algorithm, biogeography-based optimization and grey wolf optimization.

The genetic algorithm is a computational optimization technique that draws inspiration from natural selection and evolution. This powerful tool is widely employed to solve or estimate a wide variety of optimization and search problems, ranging from engineering and finance to biology and physics [21]. By mimicking the process of natural selection, genetic algorithms are able to efficiently navigate complex search spaces and identify optimal solutions for a wide range of problems. In essence, this approach is based on the idea that the fittest solutions are more likely to survive and reproduce, leading to a gradual improvement in the overall quality of the solution over time [22]. Overall, the genetic algorithm is a versatile and powerful tool that has revolutionized the field of optimization and has enabled researchers and practitioners to tackle some of the most challenging problems of our time [23]. Another optimization method in this paper is biogeography-based optimization is a method of optimization that takes its cues from nature. Biogeography is the study of how organisms spread and adapt through time in various habitats [19]. BBO is used to solve numerous optimization problems in a variety of disciplines, including engineering, biology, economics, and data science. Biogeography is the scientific study of the geographical distribution of living organisms. The 1960s saw the discovery and development of the fundamental mathematical equations regulating the spread of organisms [24]. The GWO algorithm [20] is an innovative solution that finds its inspiration in the social hierarchy and hunting habits of grey wolves in the wild. This nature-inspired optimization technique has gained popularity in the fields of computational intelligence and machine learning due to its effectiveness in solving complex search and optimization problems [25]. By mimicking the social behavior of grey wolves, the GWO algorithm proves to be an efficient and effective way to tackle real-world optimization challenges. Its unique approach provides a fresh perspective on the problem-solving process, allowing for a more comprehensive and dynamic method of finding solutions [26]. This paper makes a substantial contribution to the current research on predicting stock prices by thoroughly examining and analyzing several machine-learning algorithms. The application of optimization techniques, like genetic algorithm optimization, biologically-based optimization, and the grasshopper optimization algorithm, adds a layer of depth to the inquiry. The focus on improving the Histogram-Based Gradient Boosting technique for forecasting the Nasdaq stock index is remarkable. This study underscores the pragmatic significance for investors, emphasizing the cruciality of utilizing historical data and sophisticated algorithms to guide investment choices. The proposal to utilize ensemble techniques or hybrid models is in line with the progressive nature of stock prediction, recognizing the intricate and dynamic character of the market. The recognition of the grasshopper optimization algorithm as the most efficient optimizer contradicts current beliefs and offers a nuanced viewpoint on optimization methods in predicting stock prices. To summarize, this study enhances previous research by improving and perfecting the HGBR method, demonstrating the efficacy of particular optimization techniques, and providing practical guidance for investors in the ever-changing field of stock prediction.

According to the reviewed literatures, the main research gaps and novelties of the paper can be stated as follows.

A. Research Gaps

The field of stock price prediction, especially for the Nasdaq stock index, has long faced difficulties because of the complex interplay between contributing factors and the stock's intrinsic unpredictability. Due to their dependence on stagnant trends and inadequate analysis of the numerous factors influencing the stock market, traditional tactics frequently fail. The intricacy of this problem is exacerbated by the tendency of many machine learning models to overfit, which causes them to perform incredibly well on training data but poorly on new, untested data. Moreover, ensemble approaches are often underutilized in current models, despite the fact that they have been demonstrated to provide improved time series prediction accuracy by combining several predictions and reducing errors and uncertainties. Furthermore, although a number of optimization strategies, including Moth flame, Biogeography-based, and Gray Wolf optimization, have been studied in the literature, there is a dearth of thorough evaluation and comparison of these strategies, especially when it comes to using them to optimize the Histogram-Based Gradient Boosting (HGBR) method for stock price forecast-making.

B. Novelties of the Work

Using the powerful Histogram-Based Gradient Boosting Regressor (HGBR) in conjunction with cutting-edge optimization methods like genetic algorithms, biologically-based optimization, and most notably, the grasshopper optimization algorithm, this study presents an optimized hybrid model for the prediction of Nasdaq stock prices. The creation of the GA-HGBR, BBO-HGBR, and GOA-HGBR models is the result of a thorough comparison and empirical examination of various optimization techniques, which is where the innovation lies. These models have amazing coefficients of determination values and show significant improvements over

the baseline HGBR model. The exceptional efficacy of the grasshopper optimization method is revealed in this study, which is innovative in its application to stock price prediction. Furthermore, the huge dataset that was obtained from Yahoo Finance and the Nasdaq Stock Exchange and included a wide range of factors over a long period of time offers a distinctive and reliable basis for the predictive analysis. By lowering the chance of overfitting, this study not only fills in the holes in the current predictive models but also offers fresh perspectives on how well ensemble and nature-inspired optimization strategies might improve stock price predictions.

Lastly, the paper's structure is broken down into multiple sections, each of which focuses on a different aspect of the in-depth investigation that was done:

In Section II, the research methodology is the main topic of discussion in this section. It includes the explanation of the data that was utilized, the specifics of the model that was used, the optimization strategies that were used, and the evaluation criteria. The purpose of Section III is to present the study's result and discussion. Finally, Section IV concludes the paper.

II. Methodology

A. Data Description

This data was acquired from the Yahoo Finance Website to compile a complete historical dataset of publicly listed firms. A broad spectrum of valuable data, encompassing daily stock prices and trading volumes, was made available by this source. Five important variables—Open, High, Low, Close prices, and Trading Volume—were the main focus of the analysis in this work in order to train and test our model. To comprehend the dynamics of the stock market, these factors are essential. The opening price of a stock or other financial instrument is the price at which it is traded at the start of the trading day. It establishes the foundation for every trading day. The stock price may fluctuate and reach its highest point, referred to as the High price, during the trading day. This represents the day's peak demand or valuation. On the other hand, the price can potentially fall to what is known as the Low price—its lowest point of the day. This represents the lowest demand or valuation. The closing price is the last trading price at the conclusion of the day. It is frequently used as a benchmark for the day's performance of the stock. Additionally, trading volume is the total number of shares, contracts, or financial instruments that are exchanged in a given trading day. Elevated volumes may suggest heightened attention or involvement in a specific stock. Additionally, this dataset was supplemented with data collected directly from the Nasdaq Stock Exchange, ensuring its comprehensiveness. Access to supplementary trade metrics and market indicators was granted by this esteemed financial data source, enhancing our understanding of market dynamics. The dataset, which spans a significant timeframe from January 2015 to June 2023, includes various market conditions, including periods of stability and volatility, owing to its wide temporal range. Given its diverse array of data points that can be harnessed to construct a comprehensive industry portrait, it can be claimed that this dataset was ideally suited for robust model training and evaluation.

Nasdaq, a preeminent global stock exchange established in 1971, Nasdaq has come to represent innovation, technology, and sophisticated financial markets. Its inception aimed to modernize and streamline the stock trading process, introducing revolutionary changes to the conventional open-outcry system [27].

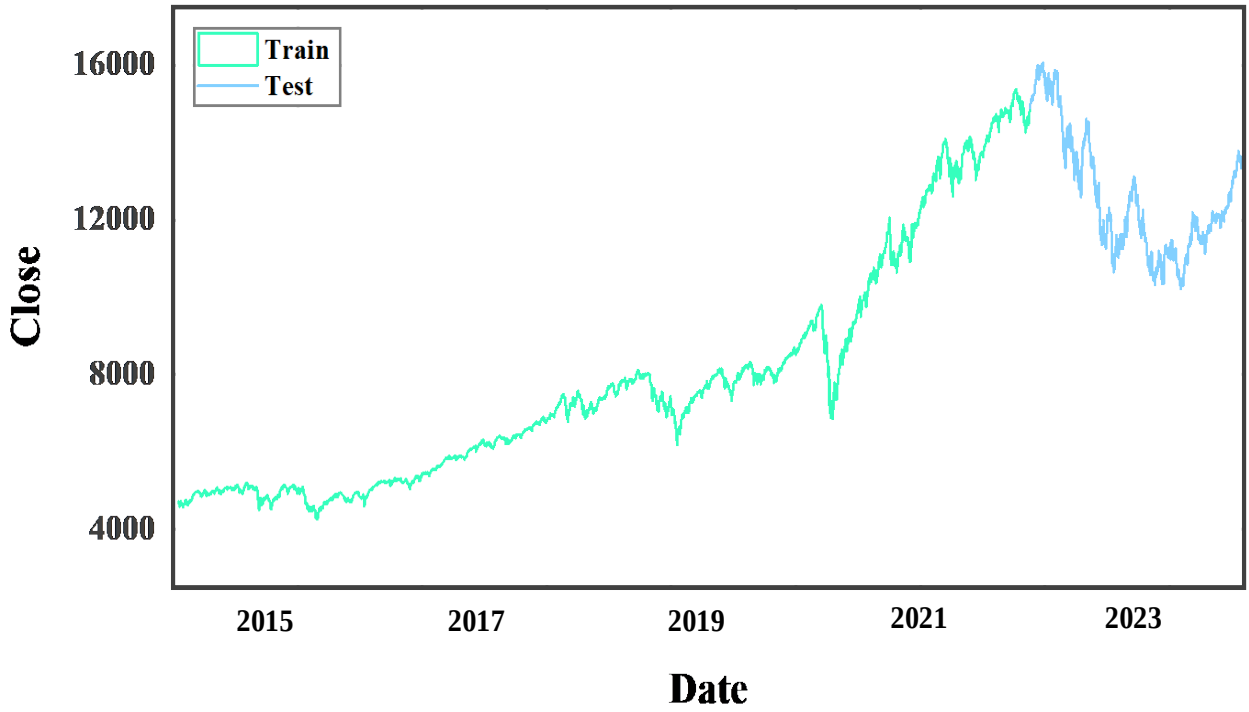

Fig. 1. Data division into training and testing.

The training and testing portions of the produced dataset are separated, as illustrated in Fig. 1. Data analysis and machine learning both start with the division of data into training and test sets. You may evaluate the results of the model and generalization skills using this technique.

B. Description of the Applied Model

1) Histogram-based gradient boosting: HGBR represents a subtype of Gradient Boosting Regressor that accelerates the computation of the gradients and Hessians of the loss function by using histograms [28]. The algorithm starts by fitting a regressor to the training data, and then it fits other regressors to the residual errors of the first ones [15]. Weak learners are weighted together to form the final algorithm. The algorithm's main goal is to reduce the loss function:

$$\mathrm{L} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad (1)$$

The approach fits a weak learner $h_t(x)$ to the residual errors of the prior regressors at each iteration. The decision tree used by the weak learner divides the data into bins according to the values of the input characteristics. The approach then directly determines, rather than estimating, the gradients and Hessians of the loss function using the histogram data. Then, using precise gradients and Hessians, the weight of the learner is determined. Understanding categorical characteristics and values that are missing organically by generating new bins for each category or missing value is one benefit of histogram gradient boosting. The ultimate model is a weighted average of each weak learner separately:

$$\hat{y}(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \quad (2)$$

where, α_t is the learner's weight for the t-th weak learner.

C. Obtained Optimization Method

When it comes to developing a reliable machine learning model, one of the most critical factors is optimizing the hyperparameters correctly. This can significantly impact the accuracy, precision, and recall of the model, which in turn can have a significant impact on the quality of the decisions and forecasts that it produces [29]. By taking the time to fine-tune the hyperparameters, developers can ensure that their model performs optimally and is better equipped to handle real-world scenarios [30].

1) Genetic algorithm: GA is an algorithm used for solving optimization and search issues that replicates the process of natural selection. Its basic principle is to repeatedly apply genetic operators like selection, crossover (recombination), and mutation on a population of candidate solutions or individuals to produce new individuals. The fitness function, which gauges the caliber of the solution, is then used to assess the new people. Until a workable solution is identified, this procedure is repeated over several generations [22].

a) GA consists of three key elements [31]: Each person is represented by a chromosome, which is a string of numbers or letters. The exact issue being handled determines the encoding. Evaluation of the fitness function is used to gauge each person's contribution to the quality of the solution. The fitness feature was created with the current issue in mind. Using evolutionary operators, new individuals can be produced from existing ones. Selection, crossover, and mutation are the three most often utilized operators. To choose the most fertile people, selection is utilized. Chromosomes from two people can be combined through a process called crossover to create a third person. The mutation is utilized to induce minor, random alterations in an individual's chromosomes. It's essential to remember that GA is a heuristic optimization technique; it cannot be relied upon to discover the best overall solution, but it can offer a good one at a reasonable computing cost. However, for large-scale issues, it could be computationally demanding and time-consuming, particularly if the dataset is sizable and the training procedure is drawn out [32].

2) Biological-based optimization: BBO, a natural-inspired optimization approach, is based on the concepts of biogeography, a scientific field that investigates how species are dispersed across time in varied ecosystems. BBO is used to handle optimization difficulties in various fields, including engineering, biology, economics, and data science. Biogeography is the study of how biological organisms are distributed geographically. The discovery and development of mathematical equations that control how organisms disperse occurred in the 1960s [24]. The concept of Biogeography-Based Optimization has caught the attention of an engineer who believes that nature can teach us valuable lessons. This algorithmic approach was developed based on the principles of biogeography, which include the birth of new species, species migration between islands, and the extinction of species. Back in 2008, Dan Simon introduced this flexible and metaheuristic strategy. It uses a mathematical framework to explain how animals move across habitats, seeking refuge from unfavorable conditions and gravitating towards more hospitable ones. The Habitat Appropriateness Index is a helpful tool for evaluating and recording the suitability of different habitats. It relies solely on the objective function of the optimization problem. One of the most esteemed evolutionary algorithms is biogeography-based optimization. This algorithm systematically enhances the best solutions by optimizing a function based on a specific quality or fitness function [33].

3) Grasshopper optimization algorithm: The Grasshopper Optimization Algorithm, a popular metaheuristic algorithm, draws inspiration from nature. Finding the finest solutions that produce the biggest potential outcome is the key objective, and randomization is used to prevent being caught in local optima. The method has shown to be very effective and efficient in optimization thanks to its speedy convergence and impressive exploration abilities. GOA has performed better in test problems than a variety of other approaches, proving its excellence and promise in practical applications. GOA is also adaptable, balancing exploitation and exploration to ensure the optimal result is reached. This unique characteristic makes GOA an excellent choice for research applications. The overall cycle of the GOA optimizer is shown in Fig. 2.

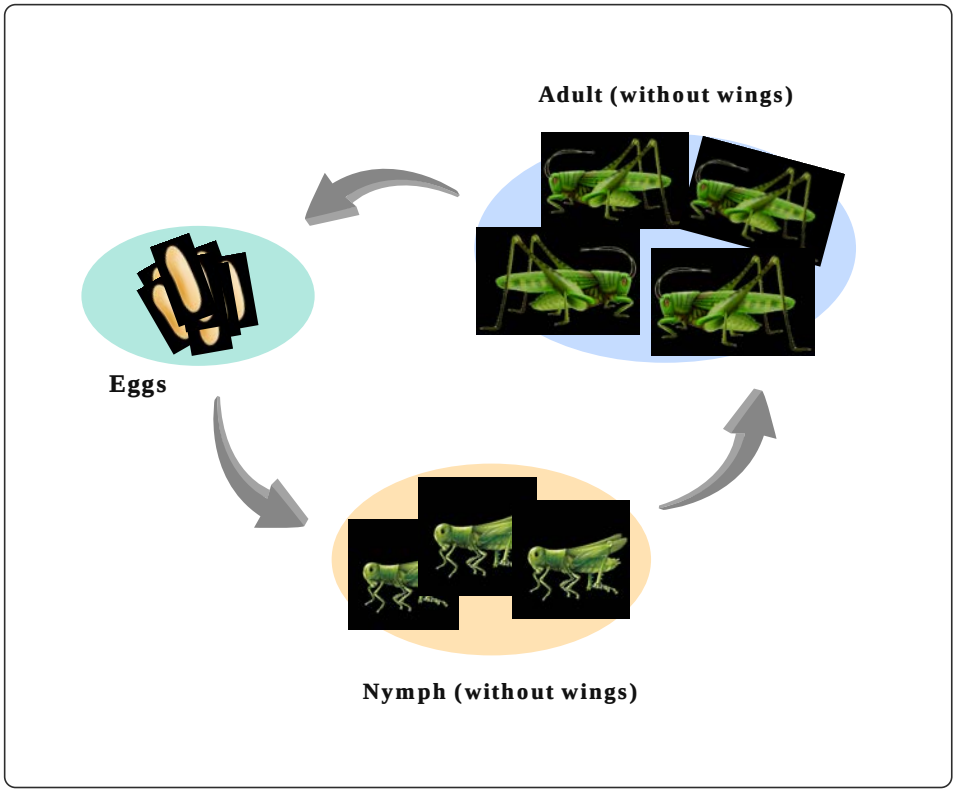

Fig. 2. Comprehensive cycle of GOA.

Suggested GOA, a Swarm Intelligence algorithm. [34] proposed GOA. Each grasshopper's position in the swarm, which is patterned after the behavior of grasshoppers, which regularly form swarms, represents a potential solution. The position of the ith grasshopper is indicated by the following equation:

$$X_i = S_i + G_i + A_i \tag{3}$$

where, S_i is for social interaction, G_i stands for gravity and A_i stands for wind advection.

The following equation, with the gravity element removed and the direction of the wind considered to be toward the target, states the equation adjusted for N grasshopper optimization:

$$X_i^d = c\left(\sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{ub_d - lb_d}{2} s\left(\left|x_j^d - x_i^d\right|\right)\frac{x_j - x_i}{d_{ij}}\right) + \widehat{T_d} \tag{4}$$

The symbol d_{ij} represents the separation among the ith and jth grasshoppers, while the function s represents the strength of the social forces, where l stands for the attractiveness scale and f for the level of attraction, all of which are calculated using the equations below:

$$\begin{aligned} d_{ij} &= \left|d_j - d_i\right| \\ s(r) &= f e^{\frac{-r}{T}} - e^{-r} \end{aligned} \tag{5}$$

The coefficient c, which decreases the comfort zone proportionately to iterations, is found using the equation.

$$c = c_{max} - l\frac{c_{max} - c_{min}}{L} \tag{6}$$

where, l is the current iteration, C_{max} denotes the maximum value, C_{min} denotes the minimum value, and L denotes the maximum number of iterations. How the GOA optimizer works from the beginning to the end of the process is shown in Fig. 3.

D. Evaluation Criteria

In statistics and machine learning, evaluation metrics are the chosen quantitative measurements for assessing the efficacy of prediction models. They assist in determining a model's ability to produce precise predictions on as-yet-unobserved data. The kind of prediction problem and the specific analytic goals determine the optimum assessment metric. Performance metrics, including $RMSE$, MAE, $MAPE$, and R^2 were employed in this study's predictive measures to assess the constructed forecasting models' predictive accuracy. A collection of mathematical formulas for these measurements is provided below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{7}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{8}$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{10}$$

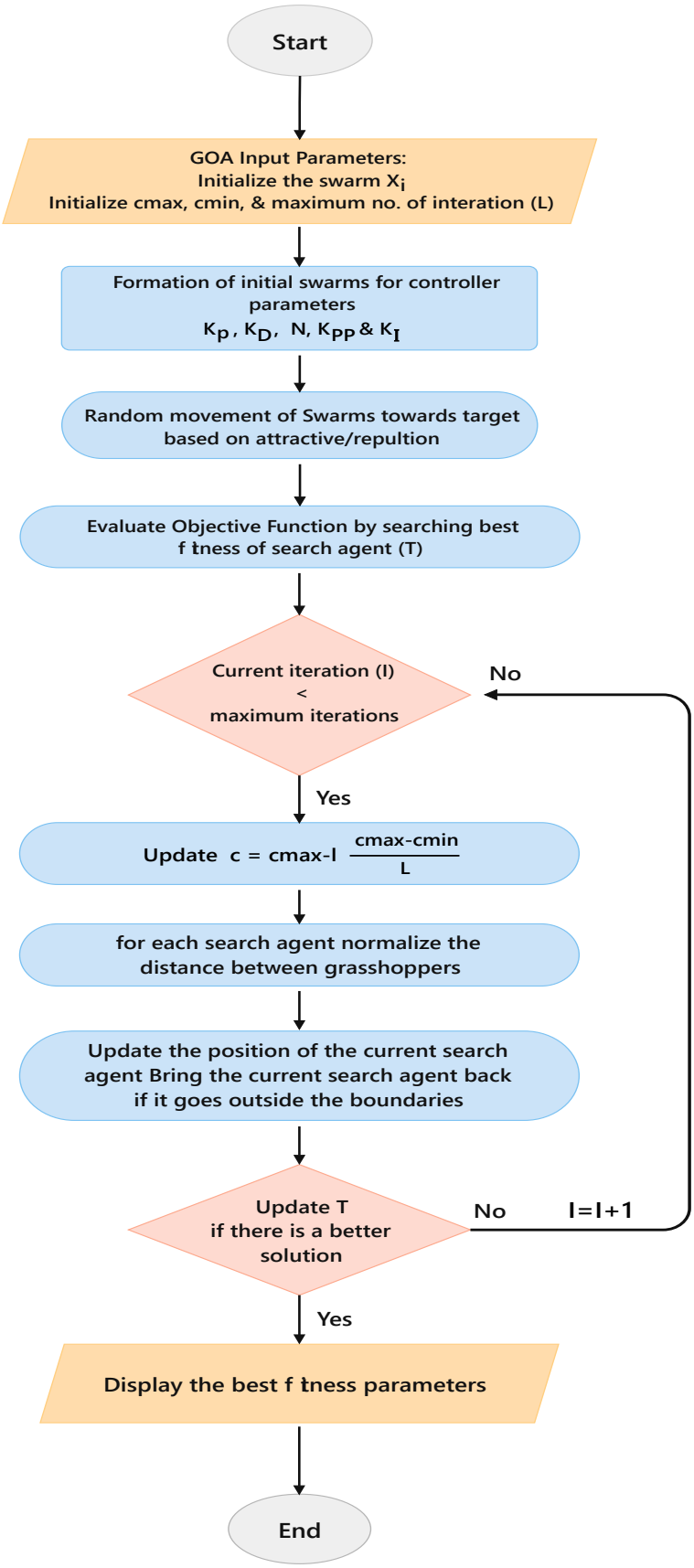

Fig. 3. Flowchart of the main optimization method.

III. Result and Discussion

A. Data Statistical Results

Table I presents the statistical results of these data points, offering insights into their distribution and variability. The table indicates the count, mean, standard deviation, minimum, 25th percentile, 75th percentile, and maximum values for each variable. The count for all variables stands uniformly at 2,137, ensuring a consistent dataset for analysis. The mean values provide an average level of each variable, with the mean Close price at 8,745.821, suggesting an overall higher closing trend in the dataset. The standard deviation, particularly high in the case of High and Low prices (3,362.163 and 3,298.311 respectively), indicates significant variability and potential volatility in the market. The minimum and maximum values highlight the range of the dataset, with a notable range in the High price (from 4,293.22 to 16,212.23). The 25th and 75th percentiles reveal the distribution's skewness, where a noticeable difference is seen in the volume, and indicating periods of both low and high trading activity.

B. Comparative Analysis

The efficacy of the models given was evaluated using a variety of standard metrics including MAE, $MAPE$, R^2, and $RMSE$. These metrics provide a thorough analysis of the forecast accuracy of the models. The performance indicators for four models, HGBR, GA-HGBR, BBO-HGBR, and GOA-HGBR, are summarized in Table II. Utilizing historical stock price information for a Nasdaq stock market index, covering from January 2015 to June 2023, these models were created and assessed.

Based on the results shown in Table II, it is clear that the GOA-HGBR model performs better than the other models in terms of predicting accuracy. The model's ability to accurately represent the complex temporal patterns and correlations contained in stock price data is demonstrated by its impressively low values for MAE, $MAPE$, and $RMSE$. These findings imply that the GOA-HGBR model may be a trustworthy resource for identifying potential market trends and making wise investment choices.

TABLE I. Data Statistical Results

count	2137	2137	2137	2137	2137
mean	8744.356	8805.287	8677.574	3143.8	8745.821
Std.	3332.744	3362.163	3298.311	1551.37	3332.058
Min	4218.81	4293.22	4209.76	706.88	4266.84
25%	5776.33	5821.95	5769.39	1908.94	5793.83
75%	11573.14	11699.63	11476.66	4416.84	11590.78
max	16120.92	16212.23	16017.23	11621.19	16057.44

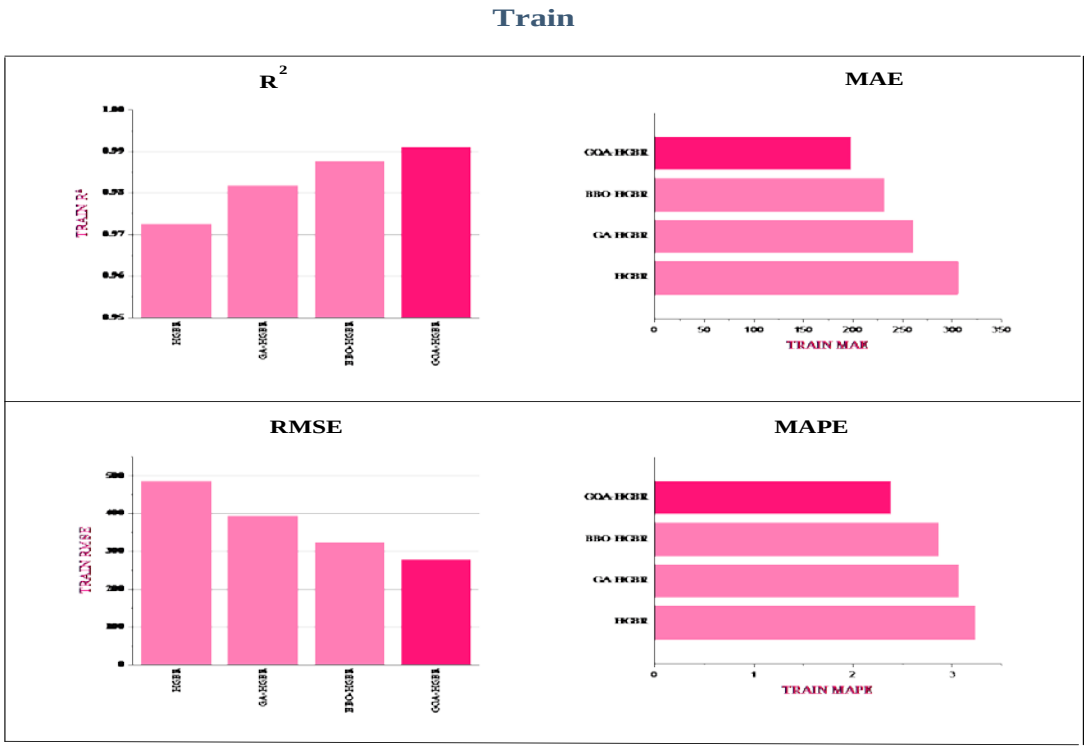

Fig. 4. The results of the evaluation criteria of the developed models during training.

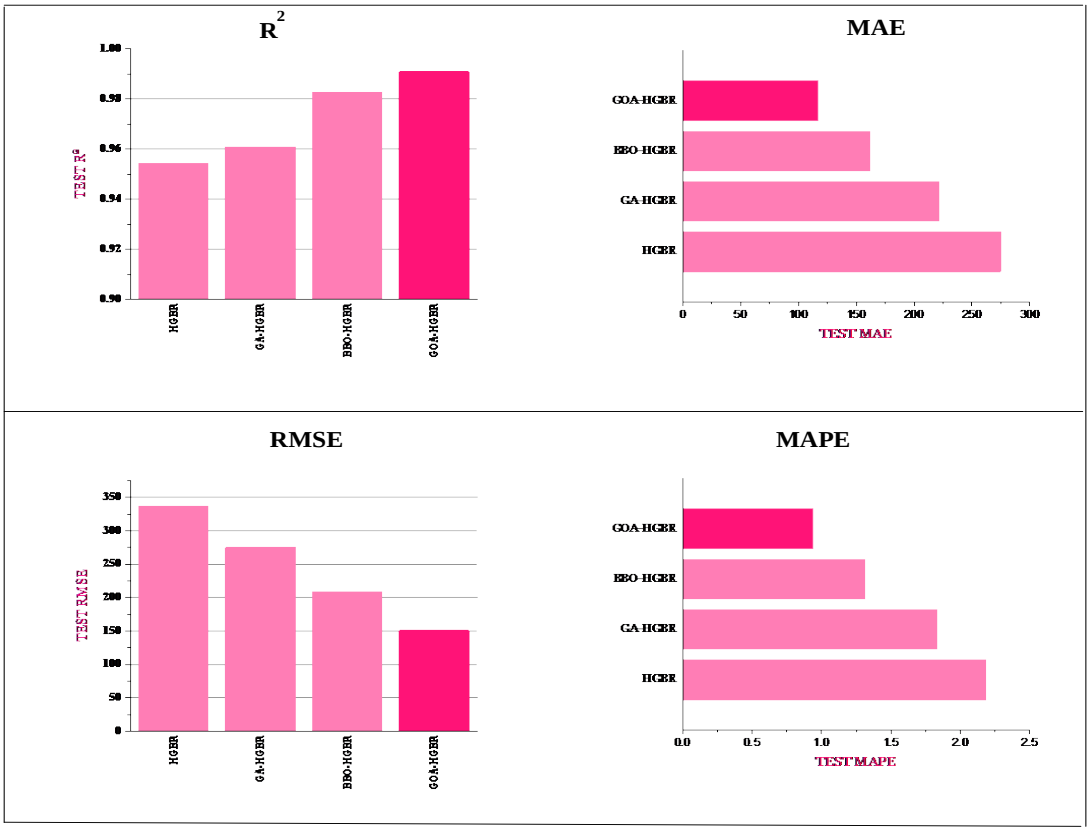

Fig. 5. The results of the evaluation criteria of the developed models during testing.

TABLE II. THE RESULTS OF MODEL PERFORMANCE CRITERIA FOR THE NASDAQ INDEX

MODEL / Metrics	TRAIN SET				TEST SET			
	RMSE	*MAPE*	*MAE*	R^2	*RMSE*	*MAPE*	*MAE*	R^2
HGBR	485.22	3.23	305.86	0.9726	337.05	2.18	274.83	0.9543
GA-HGBR	394.17	3.06	259.86	0.9819	275.18	1.83	221.78	0.9609
BBO-HGBR	324.99	2.86	231.27	0.9877	208.27	1.31	162.18	0.9825
GOA-HGBR	278.33	2.38	197.58	0.9910	150.97	0.94	117.44	0.9908

When the performance of the four models in Table II is compared, it can be observed that the GOA technique, followed by BBO and GA, produced the best results for optimizing the hyperparameters of the model that is being presented. The baseline HGBR model, although demonstrating robust prediction ability, acts as the standard for assessing the effectiveness of hybridization. The test set demonstrates an RMSE of 337.05 and an R^2 of 0.9543, providing a solid basis for evaluating the hybrid models. The incorporation of the genetic algorithm in the hybrid model results in significant enhancements. The GA-HGBR model demonstrates a decrease in the RMSE to 275.18, the MAPE to 1.83, and the MAE to 221.78 in the test set. The increase in R^2 (0.9609) indicates a higher level of accuracy in fitting the data, implying that the genetic algorithm successfully optimizes the hyperparameters to improve prediction accuracy. The integration of BBO into the hybrid model showcases additional enhancement. Significantly, BBO-HGBR demonstrates superior performance in the test set, as evidenced by its lower RMSE (208.27), MAPE (1.31), and MAE (162.18), indicating an enhanced capacity to accurately capture stock price patterns. The R^2 value of 0.9825 confirms the effectiveness of BBO in optimizing the model for enhanced accuracy in forecasting. The outcomes of the GA-HGBR, BBO-HGBR, and GOA-HGBR findings as 0.96, 0.98, and 0.99, respectively, demonstrate the improvement in the model outcomes. The evaluation results of the developed models are shown in Fig. 4 and Fig. 5, and as it is evident in the figures, it can be seen that GOA-HGRB has the best results for all evaluation criteria. The outcomes demonstrate that the prediction result has been enhanced by the optimized model. For the R^2 evaluation criterion, the GOA-HGBR model, which is optimized using the GOA technique, has a result of 0.99. This result demonstrates that optimization has a beneficial impact on predicting when compared to the HGBR model, which is not optimized. Without using the optimization approach, the HGBR was 0.95.

The comparison of the developed models is illustrated in Fig. 6 and Fig. 7. The principal objective of this research endeavor was to assess the efficacy of the GOA-HGBR model in predicting NIKKEI 225 closing prices between 2013 and 2022. The comprehensive findings of this investigation are detailed in Table III, which offers an abundance of information regarding the accuracy and effectiveness of the model across various indices. With a R^2 value of 0.9870 for the NIIKEI 225 data set, the GOA-HGBR model exhibited superior performance compared to other models that were comparable to the NASDQE index data set. The results suggest that the GOA-HGBR model could potentially be a valuable instrument for forecasting the forthcoming values of the aforementioned indices. Financial analysts and investors may utilize this information to aid in the formation of well-informed investment decisions. In its entirety, this research offers substantial contributions to the existing body of knowledge regarding stock price forecasting and underscores the potential of the GOA-HGBR model in predicting forthcoming financial market trends.

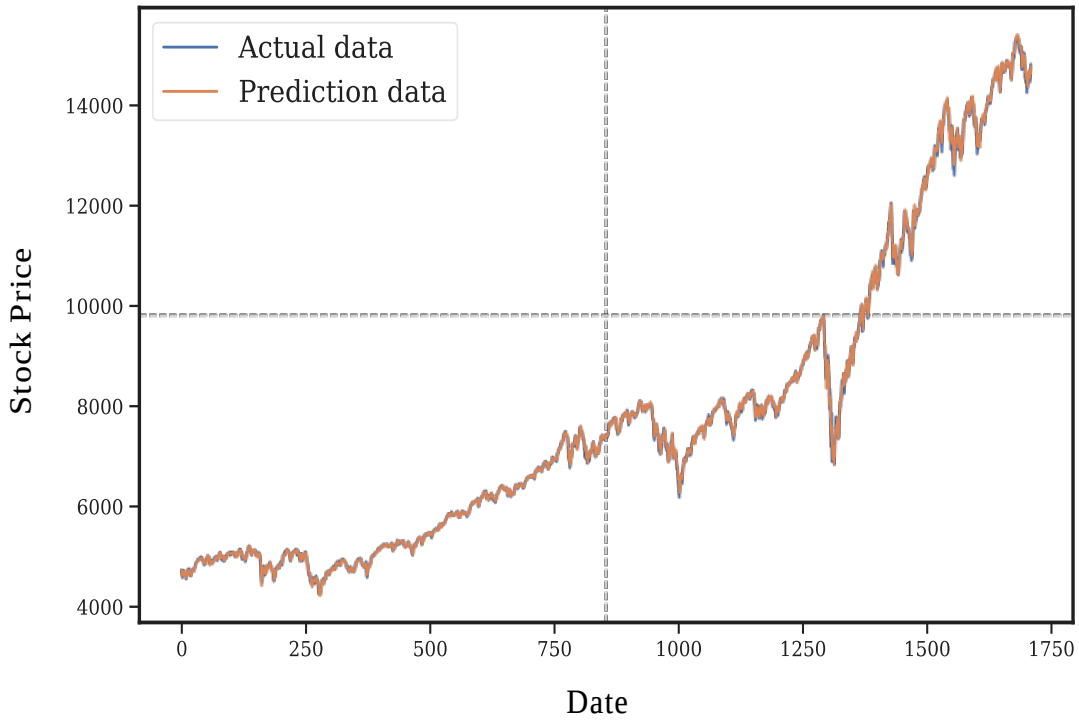

Fig. 6. Fit diagram of GOA-HGBR with other developed models during training.

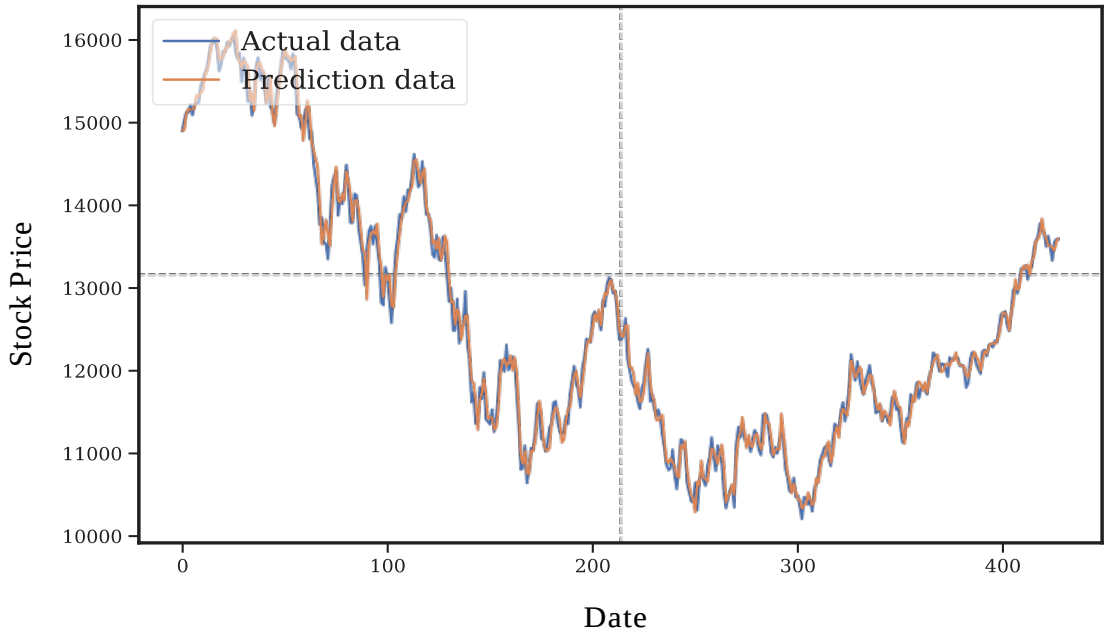

Fig. 7. Fit diagram of GOA-HGBR with other developed models during testing.

TABLE III. THE RESULTS OF MODEL PERFORMANCE CRITERIA FOR THE NIKKEI 225 INDEX

MODEL / Metrics	TRAIN SET				TEST SET			
	RMSE	MAPE	MAE	R^2	RMSE	MAPE	MAE	R^2
HGBR	359.5165	2.7144	288.0514	0.9783	185.5216	1.4299	145.5727	0.9688
GA-HGBR	335.9287	2.4682	260.1081	0.9821	175.7187	1.4188	135.3647	0.9712
BBO-HGBR	270.3312	2.1711	206.9693	0.9842	143.3115	1.4113	121.7996	0.9737
GOA-HGBR	247.8360	2.0554	186.8910	0.9873	121.9168	1.3252	98.1423	0.9870

In conclusion, the study and its findings warrant the following observations regarding future research and limitations:

- One of the model's limitations is its significant dependence on historical stock data. Particularly in the volatile and unpredictable stock market, past performance is not always indicative of future results. Consequently, this may present a constraint.
- The computational demands of sophisticated algorithms such as GA-HGBR, BBO-HGBR, and GOA-HGBR may restrict their practicality in real-time trading situations that require prompt decision-making.
- Without substantial recalibration and testing, these models may not generalize well to other stock indices or markets, despite their impressive performance for the Nasdaq stock index.
- Genetic algorithm, biologically-based optimization, and grasshopper optimization algorithm comprise the bulk of the study's attention. Alternative optimization techniques might potentially produce outcomes that are superior in quality or efficiency.

Further investigations may be warranted to examine the extent to which these models can be applied to diverse financial instruments and stock markets, thereby evaluating their adaptability and resilience.

Future Insights:

- A substantial progression would be the development of a framework for real-time data analysis and prediction, which would enable investors and traders to formulate decisions in accordance with the most up-to-date market conditions.
- Further examination of hybrid models, which amalgamate the merits of distinct algorithms, may result in the development of forecasting tools that are more precise and dependable.
- Incorporating deep learning methodologies, which have demonstrated potential in numerous predictive modeling contexts, into stock price forecasting experiments may yield novel insights and enhancements.
- Subsequent research endeavors may center on enhancing the models' capacity to navigate the frequent abrupt occurrences and market volatility that characterize the financial industry.
- By integrating these sophisticated models into user-friendly applications or platforms, they could be rendered more accessible to a wider spectrum of investors and speculators.

IV. CONCLUSION

The best course of action for investors to take, whether to buy, sell, or hold onto stocks, can be determined by using historical data and advanced algorithms. This approach is essential for investors who are committed to making intelligent investment decisions since it lowers risks and increases the likelihood of achieving profitable results. The complex and dynamic world of stock prediction was examined in this study using a variety of predictive algorithms and data sources. These findings suggest that an ensemble technique or a hybrid model may be able to anticipate more correctly. Last but not least, the creation and evaluation of the prediction model illustrated the need for data-driven insights in order to provide trustworthy conclusions. This shows the benefits of a data-centric approach in the modern, quickly changing business environment, as well as the possible applications of predictive analytics across a wide variety of sectors. In order for interested traders and investors to utilize these algorithms to buy on the correct day and at the appropriate price, this study set out to create models that could more accurately predict stock prices.

- The study's findings both support and question previous research. Utilizing a range of metrics, including Mean Absolute Error, Root Mean Square Error, Mean Absolute Percentage Error, and the Coefficient of Determination, allows for a thorough evaluation of the model's performance. The optimized hybrid genetic algorithm-based regression models, specifically GA-HGBR, BBO-HGBR, and GOA-HGBR, demonstrate substantial enhancements, achieving a coefficient of determination value of 0.9908. This not only confirms the significance of machine learning models in predicting stock prices but also undermines conventional approaches by showcasing their superior prediction powers.
- Deciding on the best model, examining the outcomes, and then modifying its hyperparameters to enhance the performance of the previously provided model.
- To further validate the efficacy of the GOA-HGBR, these algorithms were applied to and contrasted with the NIKKEI 225 index data sets.

- By contrasting the outcomes of several optimizers, the most effective optimization has been determined as the main optimizer of the model. The GOA technique yields the best results when compared to GA, BBO, and GOA, whose R^2 assessment criterion scores are 0.96, 0.98, and 0.99, respectively.

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. 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.

[3] 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.

[4] 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.

[5] D. Shah, H. Isah, and F. Zulkernine, "Stock market analysis: A review and taxonomy of prediction techniques," International Journal of Financial Studies, vol. 7, no. 2, 2019, doi: 10.3390/ijfs7020026.

[6] 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.

[7] M. Ballings, D. Van den Poel, N. Hespeels, and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," Expert Syst Appl, vol. 42, no. 20, pp. 7046–7056, 2015, doi: https://doi.org/10.1016/j.eswa.2015.05.013.

[8] M. M. Aldin, H. D. Dehnavi, and S. Entezari, "Evaluating the employment of technical indicators in predicting stock price index variations using artificial neural networks (case study: Tehran Stock Exchange)," International Journal of Business and Management, vol. 7, no. 15, p. 25, 2012.

[9] C.-F. Tsai, Y.-C. Lin, D. C. Yen, and Y.-M. Chen, "Predicting stock returns by classifier ensembles," Appl Soft Comput, vol. 11, no. 2, pp. 2452–2459, 2011, doi: https://doi.org/10.1016/j.asoc.2010.10.001.

[10] M. Zounemat-Kermani, O. Batelaan, M. Fadaee, and R. Hinkelmann, "Ensemble machine learning paradigms in hydrology: A review," J Hydrol (Amst), vol. 598, p. 126266, 2021.

[11] O. Sagi and L. Rokach, "Ensemble learning: A survey," Wiley Interdiscip Rev Data Min Knowl Discov, vol. 8, no. 4, p. e1249, 2018.

[12] S. Ardabili, A. Mosavi, and A. R. Várkonyi-Kóczy, "Advances in machine learning modeling reviewing hybrid and ensemble methods," in International conference on global research and education, Springer, 2019, pp. 215–227.

[13] A. E. L. Bilali, A. Taleb, M. A. Bahlaoui, and Y. Brouziyne, "An integrated approach based on Gaussian noises-based data augmentation method and AdaBoost model to predict faecal coliforms in rivers with small dataset," J Hydrol (Amst), vol. 599, p. 126510, 2021.

[14] G. Sonkavde, D. S. Dharrao, A. M. Bongale, S. T. Deokate, D. Doreswamy, and S. K. Bhat, "Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications," International Journal of Financial Studies, Vol 11, Iss 94, p 94 (2023), Jan. 2023, doi: 10.3390/ijfs11030094.

[15] A. Guryanov, "Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees," in Analysis of Images, Social Networks and Texts: 8th International Conference, AIST 2019, Kazan, Russia, July 17–19, 2019, Revised Selected Papers 8, Springer, 2019, pp. 39–50.

[16] G. Ke et al., "Lightgbm: A highly efficient gradient boosting decision tree," Adv Neural Inf Process Syst, vol. 30, 2017.

[17] S. Sun, Z. Cao, H. Zhu, and J. Zhao, "A survey of optimization methods from a machine learning perspective," IEEE Trans Cybern, vol. 50, no. 8, pp. 3668–3681, 2019.

[18] S. Mirjalili, "Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm," Knowl Based Syst, vol. 89, pp. 228–249, 2015, doi: https://doi.org/10.1016/j.knosys.2015.07.006.

[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] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," Advances in Engineering Software, vol. 69, pp. 46–61, 2014, doi: https://doi.org/10.1016/j.advengsoft.2013.12.007.

[21] B. Mohan and J. Badra, "A novel automated SuperLearner using a genetic algorithm-based hyperparameter optimization," Advances in Engineering Software, vol. 175, no. September 2022, p. 103358, 2023, doi: 10.1016/j.advengsoft.2022.103358.

[22] S. Mirjalili, "Genetic Algorithm," in Evolutionary Algorithms and Neural Networks: Theory and Applications, Cham: Springer International Publishing, 2019, pp. 43–55. doi: 10.1007/978-3-319-93025-1_4.

[23] A. Lambora, K. Gupta, and K. Chopra, "Genetic algorithm-A literature review," in 2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon), IEEE, 2019, pp. 380–384.

[24] 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.

[25] H. Faris, I. Aljarah, M. A. Al-Betar, and S. Mirjalili, "Grey wolf optimizer: a review of recent variants and applications," Neural Comput Appl, vol. 30, pp. 413–435, 2018.

[26] H. Rezaei, O. Bozorg-Haddad, and X. Chu, "Grey wolf optimization (GWO) algorithm," Advanced optimization by nature-inspired algorithms, pp. 81–91, 2018.

[27] A. Abraham, B. Nath, and P. K. Mahanti, "Hybrid intelligent systems for stock market analysis," in Computational Science-ICCS 2001: International Conference San Francisco, CA, USA, May 28—30, 2001 Proceedings, Part II 1, Springer, 2001, pp. 337–345.

[28] S. Md. M. Hossain and K. Deb, "Plant Leaf Disease Recognition Using Histogram Based Gradient Boosting Classifier," in Intelligent Computing and Optimization, P. Vasant, I. Zelinka, and G.-W. Weber, Eds., Cham: Springer International Publishing, 2021, pp. 530–545.

[29] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," Neurocomputing, vol. 415, pp. 295–316, 2020.

[30] B. Bischl et al., "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," Wiley Interdiscip Rev Data Min Knowl Discov, vol. 13, no. 2, p. e1484, 2023.

[31] E. Alkafaween, A. B. A. Hassanat, and S. Tarawneh, "Improving initial population for genetic algorithm using the multi linear regression based technique (MLRBT)," Communications-Scientific letters of the University of Zilina, vol. 23, no. 1, pp. E1–E10, 2021.

[32] D. M. Rocke and Z. Michalewicz, "Genetic algorithms+ data structures= evolution programs," J Am Stat Assoc, vol. 95, no. 449, p. 347, 2000.

[33] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[34] S. Saremi, S. Mirjalili, and A. Lewis, "Grasshopper Optimisation Algorithm: Theory and application," Advances in Engineering Software, vol. 105, pp. 30–47, 2017, doi: https://doi.org/10.1016/j.advengsoft.2017.01.004.