Stock Market Volatility Estimation: A Case Study of the Hang Seng Index

Shengwen Wu[1], Qiqi Lin[2]*, Xuefeng liu[3]
School of Economics and Management, Harbin University, Harbin 150086, Heilongjiang, China[1, 3]
School of Economics & Management, Guangdong Technology College, Zhaoqing 526100, Guangdong, China[2]

***Abstract*—Among the influential elements in the national economy is the stock market. The stock market is a multifaceted system that combines economics, investor psychology, and other market mechanics. The objective of the financial market investment is to maximize profits; but, due to the market's complexity and the multitude of factors that might impact it, it is challenging to predict its future behavior. The challenging process of stock price prediction requires the analysis of a wide range of social, political, and economic factors. These variables include market trends, financial statements, earnings reports, and other data. The goal of this project is to develop an accurate hybrid stock price forecasting model using Random Forest which is combined with the optimization. Random Forest is one type of machine learning that is often used in time series analysis. This study provides stock price forecasting using the Hang Seng index market, which consists of the largest and most liquid corporations that are publicly traded on the Hong Kong Stock Exchange, data from 2015 to 2023. The Dow Jones and KOSPI were evaluated as two additional indices. This study demonstrates some optimization approaches including genetic algorithm, grey wolf optimization, and biogeography-based optimization, which drew inspiration from the phenomenon of species migrating between islands in search of a suitable habitat. Biogeography-based optimization has shown the best result among these optimizations. The proposed hybrid model obtained the values 0.992, 0.997, and 0.9937 for the coefficient of determination for HSI, Dow Jones, and KOSPI markets, respectively. These results indicate the ability of the model in order to predict the stock market with a high degree of accuracy.**

***Keywords*—Hang Seng index; financial market; stock price prediction; Random Forest; biological bases optimization**

I. Introduction

A. Knowledge Background

The stock market is a network that facilitates nearly all significant global economic transactions at a dynamic rate determined by market equilibrium and known as the stock value. Predicting the stock market is a highly challenging endeavor since many factors might affect the market price, such as economic, political, and investor mood [1], [2], [3]. This resulted in random fluctuation and was caused by changes in stock market prices. Inherently volatile and loud is the stock market [4]. It is necessary to have in-depth stock knowledge to anticipate the stock exchange. Purchasing stocks that will appreciate over time is preferred by investors over stocks whose price will fall. But, to optimize investor profit and reduce loss, it is critical to create a potent stock market algorithm that has the ability to accurately predict stock behavior. Furthermore, a variety of variables that affect the stock market's volatility in the exchange market might affect it. Forecasting the future price of stocks can also be difficult when stock market data is incomplete. To forecast the movement of the stock price, investors use a variety of technical indicators. While the stock is evaluated using these indications, it is difficult to predict market developments. The behavior of stock movements is influenced by both non-economic and economic factors [5]. To comprehend the basic elements impacting stock prices, several models have been created and put to the test. To construct a model or algorithm that would permit investors to anticipate modifications more precisely than they did in the past, research is still ongoing. One of the most well-liked and often-used techniques is the creation of prediction models using machine learning algorithms [6]. Artificial intelligence (AI) and machine learning have the ability to greatly improve stock market precision forecasting in general. This would provide investors with useful information about market movements and assist them in making wise financial choices [7]. The practicality of artificial neural networks as machine learning models is beginning to pose a threat to traditional regression and statistical methods [8].

The Random Forest (RF) is an ensemble of regressors or decision-tree classifiers in which the distribution of each tree is identical across the forest and is dependent on an independent random sample [9]. Both the training data and the input variables are chosen at random during the generation of each decision-tree classifier or decision-tree regressor in RF [10]. Thus, every decision tree within an RF attains an adequate level of robustness to accommodate thousands of variables without experiencing overfitting [11]. In addition, it is possible to decrease the variance and correlation of the trees.

In recent times, meta-heuristic-based algorithms have garnered considerable interest in the optimization of objective functions across diverse domains owing to their straightforwardness, adaptability, resilience, and capacity to circumvent local maxima. To optimize the performance of the RF model, the present investigation employed three meta-heuristic optimization algorithms: the genetic algorithm (GA), grey wolf optimization (GWO), and biological bases optimization (BBO). Mirjalili et al. [12] proposed the GWO algorithm, which draws inspiration from the social hierarchy and hunting strategies exhibited by grey wolf packs and operates on a population-based model. Four levels comprise the hierarchy: omega, alpha, beta, and delta [12]. One of the evolutionary algorithms utilized to resolve optimization problems is the GA [13]. The algorithm in question is a direct

*Corresponding Author.

replication of Darwin's survival of the fittest and the process of natural evolution [13]. An initial population of randomly generated candidate solutions encoded as chromosomes is utilized by the algorithm. By applying the principle of survival of the fittest to generate ever-improving approximations, the solution to the inverse problem may be to gradually identify the elite individual attained through progress [13]. The BBO algorithm, introduced by Dan Simon in 2008 [14], is a novel population-based meta-heuristic algorithm that drew inspiration from the migration of species across various islands in search of a suitable habitat [14]. This algorithm utilizes the habitat suitability index (HSI) to quantify the quality of the homeland (solution); a solution with a high HSI is deemed to be good, whereas one with a low HSI is deemed to be poor [14]. This optimizer has been widely utilized in different tasks [13], [15], [16], [17], [18], [19].

B. Literature Review

In recent decades, substantial potential has existed for the implementation of machine learning algorithms in the context of forecasting future stock market prices. Bhalke et al. [20] examined the arduous and unpredictable characteristics of forecasting stock market prices, with a particular focus on the prevalence of recurring patterns in price curves. They investigated the feasibility of utilizing Long Short-Term Memory (LSTM), which is renowned for its efficacy with sequential data, to predict forthcoming stock prices through the analysis of daily closing prices [20]. The objective of their research was to automate prediction processes and minimize human labor in stock market analysis through the utilization of LSTM, a machine learning technique [20]. In their study, Yuan et al. [21] propose an alternative methodology to the conventional linear multi-factor stock selection model that incorporates the stock market's dynamic and chaotic attributes. They implemented a diverse range of feature selection algorithms to carry out an exhaustive feature selection procedure [21]. Time-sliding window cross-validation is employed to further refine the parameters of stock price trend prediction models that are based on machine learning [21]. The researchers employed an extensive dataset spanning eight years, which pertained to the Chinese A-share market, with the aim of determining the most effective integrated models for predicting trends in stock prices [21]. Through the analysis and evaluation of multiple integrated models, their study established that the Random Forest algorithm demonstrates remarkable effectiveness in both feature selection and stock price trend prediction [21]. Vijh et al. [22] employed Random Forest and Artificial Neural Networks (ANN) in their research to forecast the closing prices of five diversely operating companies across multiple sectors. They utilized financial data encompassing stock opening, closing, high, and low prices to produce original variables that serve as inputs for the predictive models [22]. With the utilization of ANN and Random Forest methodologies, their objective was to forecast the closing prices of stocks on the subsequent business day [22]. Moghar and Hamiche confront the complexities inherent in predicting future asset values in the perpetually volatile and uncertain financial market [23]. Their research is devoted to the development of a predictive model utilizing recurrent neural networks (RNNs), with an emphasis on LSTM models [23]. They aimed to enhance the precision of inventory value predictions through the utilization of RNN capabilities, specifically LSTM [23].

Khan et al. [24] investigated the influence of political occurrences and public opinion on the trajectory of the stock market. Their investigation encompassed not only the performance of individual firms but also the wider market milieu [24]. They aimed to ascertain whether public sentiment and political circumstances of a given day could have an impact on seven-day stock market trends. To achieve this goal, sentiment and political situation features were incorporated into a machine-learning model to assess their influence on prediction accuracy [24]. Their experimental findings revealed that the inclusion of sentiment features marginally improved the accuracy of predictions by a range of 0% to 3%. Nevertheless, the integration of the political situation feature resulted in a significant enhancement of approximately 20% in the precision of forecasts [24]. To enhance the accuracy of trend prediction about stock market volatility, Nabipour et al. [25] initiated an inquiry employing machine learning and deep learning algorithms. An investigation was undertaken to assess the relative efficacy of various prediction models concerning four discrete stock market categories that are publicly traded on the Tehran Stock Exchange: diversified financials, petroleum, non-metallic minerals, and basic metals [25]. The results indicated that when applied to continuous data, the RNN and LSTM performed better than alternative prediction models [25]. Liu and Long [26] proposed a framework for forecasting stock closing prices that takes advantage of the LSTM network's prowess in processing complex financial time series and deep learning capabilities. Their framework employed empirical wavelet transform (EWT) for data preprocessing and an outlier-robust extreme learning machine (ORELM) model for post-processing, as opposed to conventional models [26]. The primary component, a deep learning predictor based on LSTM networks, was optimized via the particle swarm optimization (PSO) algorithm and the dropout technique [26]. The feasibility of employing three machine learning algorithms—Support Vector Machine (SVM), Multilayer Perceptron, and Logistic Regression—to forecast the course of stock prices for the subsequent day was investigated by Parray et al. [27]. The experiments are executed by the researchers utilizing historical stock data spanning the period from December 31, 2018, to January 1, 2013. Approximately fifty stocks were included in the dataset, which was compiled using the NIFTY 50 index of the Indian National Stock Exchange [27]. Their results indicate that the SVM model achieves an average prediction accuracy of 87.35% [27]. Logistic Regression and Multilayer Perceptron follow suit with an accuracy of 86.98% and 75.88%, respectively [27].

Mehtab et al. [28] devised a hybrid modeling approach to forecast stock prices through the integration of machine learning and deep learning methodologies. The data utilized in their analysis is obtained from the National Stock Exchange (NSE) of India's NIFTY 50 index values [28]. The time span encompassed by this dataset is between December 29, 2014, and July 31, 2020. In order to predict the open values of the NIFTY 50 index between July 31, 2020 and December 31, 2018, eight regression models are constructed using training data that covers the period from December 29, 2014 to December 28, 2018 [28]. Ayala et

al. [29] introduced a hybrid approach for generating trading signals in stock market prediction by integrating machine learning methodologies with technical analysis indicators. Future applications of their method [29] involving the integration of machine learning and a technical indicator for the purpose of informing trading decisions may be justified, given its straightforwardness and efficacy [29]. An evaluation of the performance of four machine learning techniques was conducted to ascertain the most suitable one: a random forest, a linear model, and four neural networks. Utilizing daily trading data from prominent indices including the Ibex35, DAX, and Dow Jones Industrial, they assessed their technical trading strategies by employing the Triple Exponential Moving Average and Moving Average Convergence/Divergence [29]. In order to forecast the S&P 500 index's closing price for the subsequent day, Bhandari et al. [30] utilize LSTM, an architecture designed specifically for neural networks. A thorough examination of the behavior of the stock market is accomplished by constructing a meticulously curated collection of nine predictors [30]. This ensemble comprises technical metrics, macroeconomic indicators, and fundamental market data. Subsequently, both single-layer and multilayer LSTM models are constructed utilizing the selected input variables [30].

C. Research Gaps, Motivations, and Main Contributions

The literature review does not incorporate optimization techniques. Additionally, it fails to analyze the effectiveness of these techniques or determine which one produces the most favorable outcomes in the context of stock price forecasting. Although numerous studies examine the creation of hybrid models that combine machine learning and deep learning techniques for predicting stock prices, there is a lack of thorough assessment regarding the effectiveness of these models in comparison to conventional machine learning models. The majority of studies discussed in the literature review concentrate on predicting stock prices for individual companies or indices. However, there is a dearth of research that specifically applies these models to the Hang Seng Index (HSI) market, Dow Jones, and KOSPI indexes. By developing an innovative hybrid stock price forecasting model using these datasets, this research fills in the existing gaps in knowledge. Several optimization techniques, including GA, GWO, and BBO, are incorporated with the widely used machine learning algorithm RF. To identify the most accurate forecasting method for the HSI market, this research entails a comparative evaluation of these optimization techniques. In addition, a comparative analysis of the performance of our hybrid model with pre-existing hybrid models has been conducted. Incorporating optimization techniques with RF for stock price forecasting in the HSI market yielded outcomes that establish the efficacy of this methodology in both accuracy and predictive capability.

In order to address the complexities of the stock market landscape, the motivation is to create hybrid models that combine machine learning and optimization techniques. The primary objective of the model is to optimize its applicability and precision by devoting attention to the distinctive dynamics of the HSI market. To enhance the predictive performance of the forecasting model, a comparative analysis of various optimization strategies was conducted. With the ultimate goal of enabling well-informed decision-making in the financial sector, the research endeavors to furnish investors with dependable insights, thereby promoting economic stability and growth. These are the primary contributions of the study:

- The research paper makes a scholarly contribution to the domain of financial forecasting through the proposition of an optimized hybrid model designed to forecast stock prices. The research enhances the methodology for forecasting and modeling stock market trends by integrating RF with some optimization techniques, including GA, GWO, and BBO.
- By conducting an analysis of HSI market data spanning the years 2015 to 2023, this study offers significant insights into the intricacies of stock market behavior. Additionally, KOSPI and the Dow Jones were evaluated as supplementary indices. Through the identification of critical determinants that impact stock prices and the construction of a model that can effectively capture these intricacies, this study enhances the collective comprehension of investor conduct and market trends.
- This research investigates the utilization of optimization methods to refine the performance of machine learning models. Through the assessment of various optimization techniques, including GA, GWO, and BBO, this study provides valuable insights regarding the enhancement of machine learning algorithms specifically designed for the purpose of financial forecasting.

The remaining portion of this article is structured in the following manner: Section II outlines the materials and methods used, as well as provides details about the dataset and assessment metrics. The experimental findings are presented in Section III. Additionally, the analyses and discussions are outlined in Section IV. Finally, the study's conclusions are presented in Section V.

II. Methodology

A. Random Forest

An ensemble learning algorithm [31] typically includes a Random Forest [32]. The algorithm's core idea is to create and generate a decision tree using the subset of sampled original data, combine several decision trees into a random forest, and then carry out a replacement sampling process using the original data set using the Bootstrap sampling method. The forecast generated by a Random Forest regression model is constructed by aggregating the results generated by numerous decision trees. The mean of the predictions made by each individual decision tree in the Random Forest constitutes the final output. Consequently, the collective forecast generated by the ensemble of decision trees constitutes the mean value or consensus of the overall forecast produced by the random forest. In the random forest algorithm paradigm, every decision tree Fig. 1 has a succession of decision nodes that resemble a tree, which comprises the individual phases of the algorithm. The tree is split into several branches till it reaches the leaf at the tip of the tree based on this sequence. Each decision tree's output prediction is routed through leaf nodes, and the aggregated outputs of several decision trees are then used to make predictions. Among its benefits are its quick training pace and

ability to prevent overfitting. The selection and configuration of the hyperparameters for a Random Forest model significantly influence its overall performance and capacity for generalization. The "Maximum Depth" parameter controls the depth of decision trees, which has implications for the model's complexity and vulnerability to overfitting. Specifically, GA favors a depth of 80, GWO tends towards 50, and BBO converges at 60. The setting for feature selection during splitting, "Maximum Features," consistently prioritizes the "auto" option across all optimization techniques. The minimum number of samples necessary for node splitting and "Minimum Samples Leaf" are determined by "Minimum Samples Split" and "Minimum Samples Leaf," respectively. GA selects 2 samples for both, GWO selects 1 and 4, and BBO selects 2 and 3. In order to ensure reproducibility, "Random State" initializes randomness; GA, GWO, and BBO select 42, 64, and 24 elements, respectively. The ensemble size is ultimately determined by the "Number of Estimators," whereby 300, 200, and 500 trees are selected by GA, GWO, and BBO, respectively. The judicious modifications executed by each optimizer underscore their sophisticated approaches in refining hyperparameters with the aim of improving the random forest model's performance on the dataset, thereby guaranteeing an equilibrium between intricacy and applicability. The setting of the hyperparameters of the RF can be observed in Table I.

B. Genetic Algorithm

GA is a computer technique that solves optimization and search issues by simulating the process of natural selection. The basic notion is to create new persons by continually applying genetic operators like selection, crossover (recombination), and mutation to a population of candidate solutions known as individuals. The quality of the solution is then gauged by a fitness function, which is used to assess the new individuals. Until a workable solution is identified, this procedure is repeated over several generations [33], [34]. GA has three essential components [35]:

Encoding: Every person is represented by a chromosome, which is a collection of numbers or letters. The particular issue being handled determines which encoding is used.

Evaluation: The standard of the response that each person represents is assessed using the fitness function. The current issue guides the design of the fitness function.

Operators that are used to create new individuals from preexisting ones are called evolutionary operators. The operators that are most often utilized are mutation, crossover, and selection. To choose the most qualified people to procreate, selection is utilized. The process of crossover allows the chromosomes of two persons to combine to generate one new person. Individual chromosomes can have minor, random alterations introduced into them through the use of mutations.

GA is a heuristic optimization technique; however, technique cannot guarantee the discovery of the best global solution, but it can yield a respectable result at a manageable computing cost. For large-scale issues, however, it could be computationally demanding and time-consuming, particularly if the dataset is big and the training procedure takes a while [36].

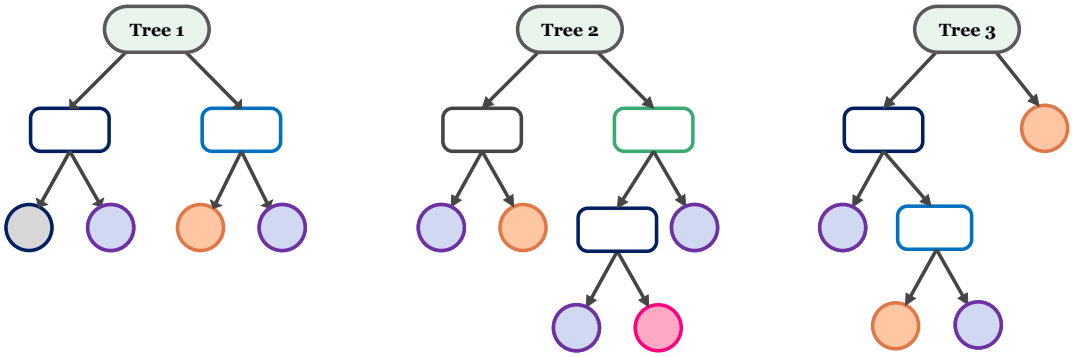

Fig. 1. Illustration of RF.

TABLE I. THE SETTING OF THE HYPERPARAMETERS

Random Forest		GA	GWO	BBO
Max depth	[10, 100, None]	80	50	60
Max features	['auto', 'sqrt']	auto	auto	auto
Minimum samples leaf	[1, 4]	2	1	2
Minimum samples split	[2, 10]	2	4	3
Random state	[4, 24, 42, 64, 88]	42	64	24
Number estimators	[200, 2000]	300	200	500

During training, the number of iterative cycles executed by the optimization algorithm is determined by the epoch parameter, which is configured to 500. With a value of 100, the population size parameter specifies how many potential solutions are assessed during each iteration.

C. Gray Wolf Optimization

The optimization strategy that will be presented in this part is distinct and is modeled after the natural hunting organization of grey wolves. In, Mirjalili et al. [37] introduced the Grey Wolf Optimization (GWO) approach. It is claimed that each wolf in the pack lives in a semi-democratic way, with a specific place in this algorithm. To prepare for the hunt, the wolves first circle their target. As they go closer and loosen the encirclement, they gradually exhaust the victim. When the dominant wolf gives the signal, they attack and capture the prey. The wolf hierarchy is as follows:

The alpha pair (α), the group leader, makes the decisions. Alpha decisions have an effect on the group as a whole. However, one also observes a certain democratic conduct.

Beta wolves (β) help alphas with decision-making and other group activities. The most qualified wolves are the alpha wolves until they are too old or die.

The lowest-ranking wolves in a pack are called omega wolves (ω). These are the wolves that warn of approaching disaster. Wolves have to follow the decisions made by other wolves as they eat their prey. As a result, the Omega Wolves are typically not particularly significant. However, if they are eradicated or disregarded, the group may have major problems, such as civil war.

Wolves that do not fall into the previously stated groups are known as delta wolves (δ). These wolves are superior to omegas even if they follow the alpha and beta hierarchies.

As mentioned before, grey wolves are known for their rapacious hunts for prey. Eq. (1) below simulates grey wolves' hunting habits.

$$\begin{aligned} \vec{D} &= \left|\vec{c} \cdot \overrightarrow{X_p}(t) - \vec{X}(t)\right| \\ \vec{X}(t+1) &= \overrightarrow{X_p}(t) - \vec{A} \cdot \vec{D} \\ \vec{A} &= 2\vec{a}\overrightarrow{r_1} - \vec{a} \\ \vec{C} &= 2\overrightarrow{r_2} \end{aligned} \tag{1}$$

During algorithm iterations, the grey wolf's location vector X, denoted by t in Eq. (1), linearly decreases from a value of 2-0. The coefficients of the prey position vector are represented by vectors A and C. r_1 and r_2 are random vectors in the interval $[0,1]$. The algorithm undergoes 500 rounds of iterative refinement with an epoch value of 500 to improve its predictive capabilities. The algorithm optimizes its search efficiency by concurrently evaluating 100 candidate solutions in each iteration, with a population size of 100.

D. Biogeography-based Optimization

When combined with a more effective exploration method, the BBO algorithm is proven to be effective in exploiting the search space. Because they share qualities, superior solutions tend to draw in inferior ones. The operators listed below are used to process this feature sharing.

Migration Operator: Migration is the process by which, depending on immigration and emigration rates, the poorer solution is replaced with a better habitat. The method by which a species enters a habitat is measured by its emigration rate. Better solutions will see a greater rate of emigration than inferior ones.

The quantity measurement used to determine how a species leaves its environment, however, is the immigration rate. Therefore, in a worse solution than in a better one, the immigration rate will be higher. The simplest form of BBO, the straight lines seen in Fig. 2, have been employed. For the linear functions, it is possessed: Therefore, in a worse solution than in a better one, the immigration rate will be higher. The simplest form of BBO, as seen in the straight lines in Fig. 2, has been employed. For the linear functions, it is assumed that:"

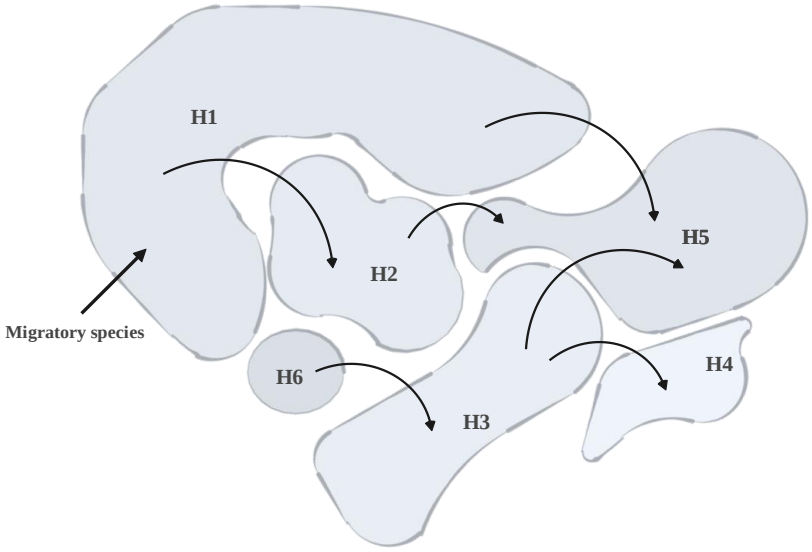

Fig. 2. Visualizing the biogeography-based optimization.

$$\mu_k = \frac{E \times k}{n} \lambda_k = I(1 - \frac{k}{n}) \quad (2)$$

where,

μ_k: Emigration rate of k^{th} habitat.

λ_k: Immigration rate of k^{th} habitat.

I: Maximum immigration rate.

E: Maximum emigration rate.

$n = S_{max}$: Maximum number of species a habitat can support.

k: Number of species count.

As species diversity increases, immigration rates decrease. On the other hand, the emigration rate rises in tandem with the number of species. S_1 and S_2, two possible solutions, exist. While S_1 is a somewhat subpar response, S_2 is a quite good one. On average, immigration rates for S_1 are higher than those for S_2. Compared to S_2 emigration, S_1 emigration will occur at a slower rate.

Mutation: A BBO mutation is comparable to an abrupt shift in living circumstances brought on by other events, such as a tornado, volcanic eruption, or natural disaster. Since the previous environment is no longer adequate for the species to live, the random change in the solution indicates that the animal moves to a new habitat.

The epoch parameter, which is initialized to 500, controls the length of time that the algorithm iteratively processes. This is an essential factor in enhancing predictive models. The population size parameter, when configured to 100, has an effect on the algorithm's capacity for exploration and diversity, thereby influencing its convergent solution generation capability. Fig. 3 provides the overall structure for the Biogeography Based Optimization Algorithm for easier comprehension.

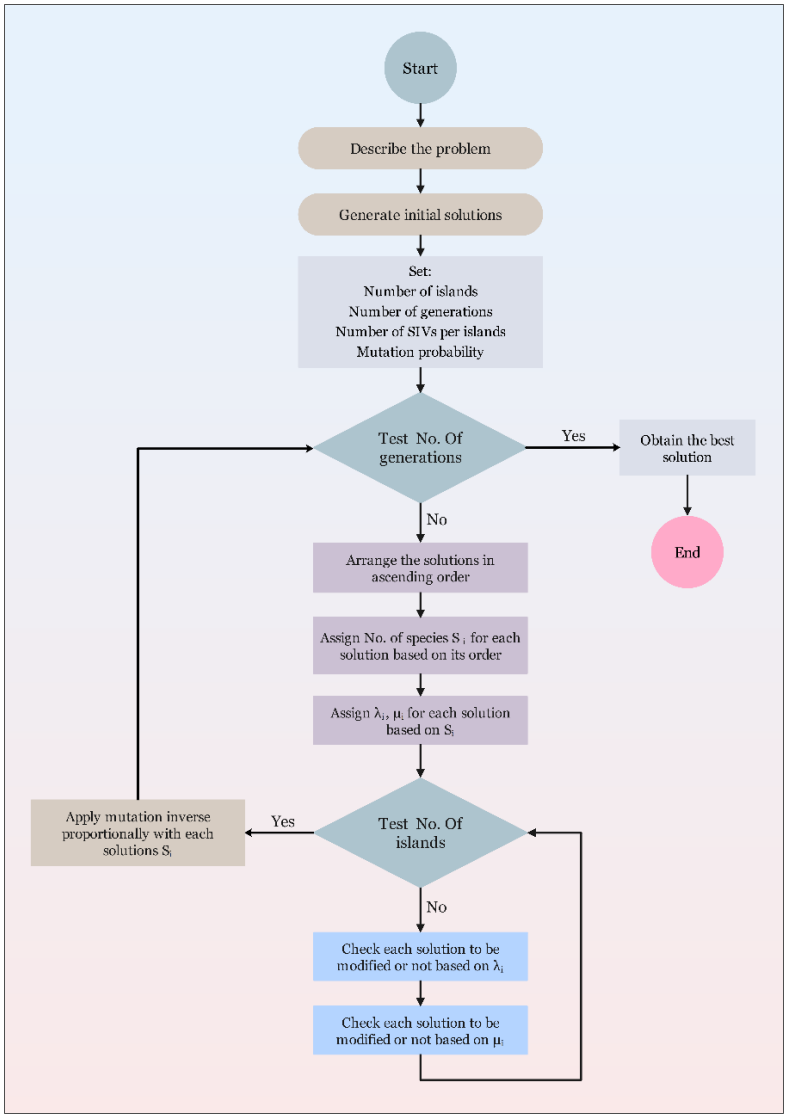

Fig. 3. RF flowchart.

E. Dataset Description

A daily closing price time series shows the observed data for each index in one dimension. The complete dataset was first split up into testing and training groups. The first 80% of the data are part of the training set and are used to train the model parameters. As can be observed from the data shown in Fig. 4, the testing set's last 20% of data is utilized to assess the models' effectiveness.

This article was shown using data from the Hang Seng Index. Several techniques, including normalization, are used to prepare this data, which spans from the start of 2015 to mid-2023. A notable Hong Kong stock market index, the Hang Seng Index monitors the performance of a subset of the largest corporations that are publicly traded on the Hong Kong Stock Exchange [38]. The Hang Seng Index comprises an assortment of corporations that hold leadership positions across multiple sectors of the Hong Kong economy [38]. In addition to manufacturing, these sectors also include finance, real estate, technology, and telecommunications. The Hang Seng Index comprises a number of corporations renowned for their substantial international footprint and profound global impact [38]. This global reach enhances the index's significance as an indicator of market and economic trends outside of Hong Kong. In brief, the Hang Seng Index monitors the performance of major corporations that are publicly traded on the Hong Kong Stock Exchange. It is a significant Hong Kong stock market index. It functions as a benchmark for investors, furnishes valuable insights into the state of the Hong Kong economy, and is indispensable for comprehending market sentiment and trends within the Hong Kong equity market [38]. Fig. 5 illustrates the graphical representation of the daily and time series collection features for the HSI index. It displays the prices of the open, high, low, and close, as well as the volume value for the first five days and last five days. Two additional indices—the Dow Jones and KOSPI—collected from the beginning of 2015 to the middle of 2023 were assessed in order to demonstrate the efficacy of the proposed model. The Dow Jones is an exceptionally followed stock market index on a global scale. It monitors the performance of thirty sizable, publicly traded companies in the United States that are publicly traded on stock exchanges. The KOSPI Index serves as the primary benchmark for the South Korean stock market. The investment vehicle monitors the progress of every common stock that is listed on the Korea Exchange (KRX), the exclusive operator of a securities exchange in South Korea. The KOSPI Index comprises an extensive array of industries, such as consumer goods, finance, technology, and automotive, among others.

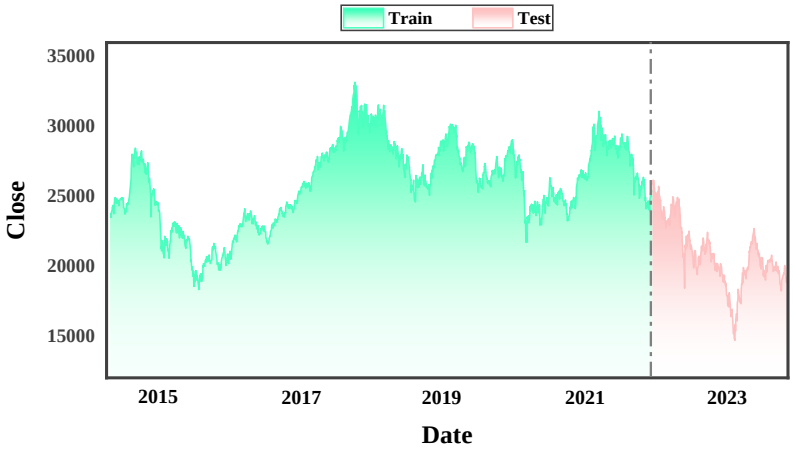

Fig. 4. Dataset example and its division into training and testing.

Date	Open	High	Low	Volume	Close
2015-01-02	23699.199219	23998.900391	23655.500000	1801713100	23721.300781
2015-01-05	23699.189453	23998.869141	23655.519531	2585193100	23721.320312
2015-01-06	23515.130859	23611.000000	23312.500000	2617976900	23485.410156
2015-01-07	23396.699219	23715.710938	23332.029297	2181069500	23681.259766
2015-01-08	23920.349609	23941.640625	23719.050781	2011642900	23835.529297
⋮					
2023-06-23	19135.019531	19138.419922	18800.339844	1689313400	18889.970703
2023-06-26	18845.900391	19001.619141	18767.150391	2066052200	18794.130859
2023-06-27	18851.660156	19226.320312	18842.419922	2059536300	19148.130859
2023-06-28	19099.390625	19222.130859	19019.259766	1675196500	19172.050781
2023-06-29	19180.279297	19180.279297	18837.320312	1690353400	18934.359375

Fig. 5. A visual example of head and tail analysis of the HSI index.

Maintaining the relative connections between the values is the aim while bringing all the qualities to the same scale. This may be crucial for machine learning algorithms that depend on the amount of data that is supplied. The data normalization procedure uses the following formula:

$$\text{XScaled} = \frac{(X - Xmin)}{(Xmax - Xmin)} \quad (3)$$

F. Model Evaluation

A comprehensive array of evaluation metrics was utilized in this investigation of stock prediction in order to assess the performance of the predictive models. The aforementioned metrics consist of the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2).

MAE calculates the mean discrepancy between anticipated and observed values. By calculating and averaging the absolute differences between predicted and observed values, MAE offers a straightforward indication of the predictive accuracy of a model, irrespective of the error direction. When applied to the domain of stock prediction, MAE provides insight into the average discrepancy that occurs between our forecasts and the real prices of stocks [39]. It can be calculated by using the following equation:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \quad (4)$$

The accuracy of predictions is quantified by MAPE in percentage format. The metric calculates the mean percentage discrepancy between the values predicted and those observed. Regardless of the scale of the data, MAPE is particularly useful in financial forecasting, such as stock prediction, because it provides insights into the relative accuracy of predictions. It provides a percentage-based understanding of the degree to which our forecasts differ from actual stock prices [39]. The calculation can be performed utilizing the subsequent equation:

$$MAPE = \left(\frac{1}{n} \sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right|\right) \times 100 \quad (5)$$

Another widely employed metric for assessing the accuracy of predictions is RMSE. The square root of the mean of the squared discrepancies between the predicted and observed values is computed. By assigning greater penalties to larger errors as opposed to smaller ones, RMSE offers a more nuanced evaluation of predictive performance. The RMSE metric is utilized in stock prediction to assess the overall adequacy of our models by taking into account the error's magnitude and direction [39]. The calculation can be performed utilizing the subsequent equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (6)$$

R^2 assesses the extent to which the independent variables (predictors) in the model account for the variability observed in the dependent variable (stock prices). It is bounded between 0 and 1, where higher values signify a more optimal correspondence between the model and the data. Determining the extent to which our predictive models can account for the variability observed in stock prices requires R^2 as a critical metric. The evaluation of the model's ability to account for the observed variations in stock prices is facilitated by this [39]. The following equation could be employed to compute it:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (7)$$

where, y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ represents the mean value [39].

III. Experimental Results

A. Statistical Values

As part of the study report, Table II provides a thorough analysis of the dataset. Information on OHLC price and volume is presented statistically in a comprehensive manner in the table. A more thorough comprehension of the facts is made possible by this. Several statistical measures are displayed in the table, such as the count, 50%, kurtosis, skewness (skew), mean, standard deviation (Std.), minimum (min), and maximum (max) values. An exact and comprehensive data analysis is provided by these measures. From the central tendency to the variability to the dispersion of the data, each of these metrics provides insightful information about a range of aspects of the data.

B. Algorithms Outcomes

An exact and thorough data analysis is provided by these measures. Among the several features of the data that each of these measures offers valuable insights into are the central tendency, variability, and dispersion of the analysis. This work's main goal is to identify and assess the best hybrid algorithm for stock price prediction. To do this, the study created forecasting models and examined intricate factors impacting stock market movements. The objective is to provide analytical data that helps investors and analysts make well-informed investing decisions. The efficacy and performance ratings of each model are fully analyzed in Table III and Table IV and Fig. 6 and Fig. 7.

TABLE II. Statistical Findings from the Dataset

	count	mean	Std.	min	50%	max	skew	kurtosis
Open	2090	24877.8	3492.279	14830.69	25002.49	33335.48	-0.19992	-0.65433
High	2090	25026.72	3486.289	15113.15	25118.69	33484.08	-0.18469	-0.6701
Low	2090	24689.52	3484.234	14597.31	24755.93	32897.04	-0.21056	-0.64255
Volume	2090	4013.656	1462.996	0	3679.685	12025.52	1.660448	4.339923
Close	2090	24862.03	3486.437	14687.02	24973	33154.12	-0.20035	-0.64908

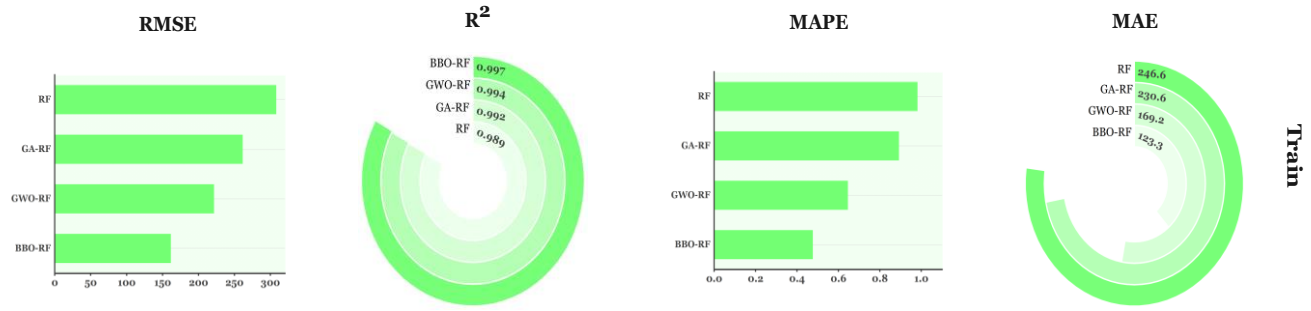

Fig. 6. Result of the Evaluation metrics for the presented models during train.

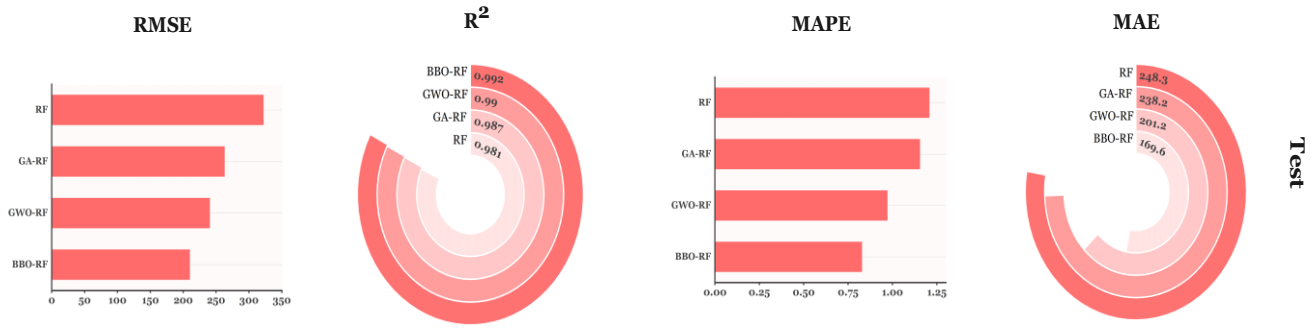

Fig. 7. Result of the Evaluation metrics for the presented models during the test.

The primary objective of each model was to forecast the HSI, and the same dataset was used in each model. Additionally, two other indexes were evaluated to prove the efficiency of the proposed model, which these indexes are the Dow Jones and KOSPI collected from the start of 2015 to mid-2023.This article presents a comprehensive and informative study by carefully comparing and evaluating each model's performance. It is crucial to clarify the performance measures used to evaluate the models to provide a fair and adequate comparison. Assessing the models using a range of important metrics, as explained in the methodology section. It is possible to evaluate every model's performance using a range of indicators and then choose the model that best fits the needs. Table III and Table IV offer an in-depth analysis of all the subtle aspects of each model's operation, along with the outcomes.

TABLE III. ANALYZING DATA WITH METRICS FOR TRAING SET

TRAIN SET	MODEL/Metrics	RF	GA-RF	GWO-RF	BBO-RF
HSI	R^2	0.9891	0.9922	0.9944	0.997
	RMSE	307.81	261.17	221.27	161.23
	MAPE	0.98	0.89	0.65	0.48
	MAE	246.64	230.62	169.22	123.3
DOW JINES	R^2	0.9883	0.9906	0.9930	0.9982
	RMSE	562.04	503.87	434.14	217.25
	MAPE	2.23	2.02	1.38	0.67
	MAE	533.98	481.98	327.09	159.03
KOSPI	R^2	0.9864	0.9897	0.9922	0.9958
	RMSE	6.35	5.53	4.81	3.53
	MAPE	1.64	1.60	1.36	0.96
	MAE	5.01	4.67	3.93	2.84

TABLE IV. ANALYZING DATA WITH METRICS FOR TESTING SET

TEST SET	MODEL/Metrics	RF	GA-RF	GWO-RF	BBO-RF
HSI	R^2	0.981	0.9875	0.9895	0.992
	RMSE	322.02	262.85	240.4	209.87
	MAPE	1.21	1.15	0.97	0.83
	MAE	248.31	238.22	201.2	169.6
DOW JONES	R^2	0.9864	0.9899	0.9921	0.9970
	RMSE	213.54	201.48	186.04	164.18
	MAPE	0.66	0.54	0.42	0.37
	MAE	189.60	166.74	148.92	130.50
KOSPI	R^2	0.9846	0.9880	0.9911	0.9937
	RMSE	3.81	3.36	2.91	2.44
	MAPE	1.01	0.79	0.69	0.56
	MAE	3.39	2.63	2.33	1.90

IV. DISCUSSION

Initially, based on the acquired result, the RF model was chosen. The higher performance of the RF model led to its formulation following a comprehensive analysis of the data. About the HSI market data from the start of 2015 to the end of 2023, appropriate data were selected and normalized. Furthermore, to substantiate the efficacy of the suggested model, two additional indices—the KOSPI and the Dow Jones—gathered from the beginning of 2015 to the middle of 2023—were assessed. This rigorous strategy will extract pertinent facts to aid in decision-making. R^2, RMSE, MAE, and MAPE were used to analyze the data analysis in detail. These indicators have a solid reputation for offering an accurate assessment of the analysis's overall dependability, efficacy, and correctness. The R^2, RMSE, MAPE, and MAE criteria were used to assess the effectiveness of the RF model both with and without an optimizer. This evaluation improved the ability to comprehend the model's performance and make judgments based on the results. Table III and Table IV shows that the evaluation result for the RF alone in testing is 0.9810 from R^2, which has increased due to the advances of the optimizers. The R^2 criteria values for GA-RF, GWO-RF, and BBO-RF are 0.9875, 0.9895, and 0.9920, respectively, indicating that selecting the optimal course of action is possible. The RMSE values were 307.81 and 322.02 for RF during training and testing, while the MAE values for training and testing were 246.64 and 248.31, respectively. The MAPE values were 0.98 and 1.21 for RF during train and test. The testing results are a major factor in determining the optimal approach for predicting stock values in the HSI, Dow Jones, and KOSPI markets. When examining the metrics of the testing set, the BBO applied to the RF model, denoted BBO-RF, is the focal point. BBO-RF demonstrates exceptional performance on both the DOW JONES and KOSPI indices, highlighting its capability to enhance predictive accuracy and reduce errors. The impressive R^2 score of 0.9970 achieved by BBO-RF for Dow Jones indicates a high degree of accuracy in predicting market movements. It is worth mentioning that it attains the lowest RMSE, MAPE, and MAE values, which emphasizes its ability to produce accurate predictions with minimal discrepancy from the actual values. In the same way, BBO-RF exhibits its superior performance on the KOSPI index by producing significantly reduced error metrics in comparison to RF and other optimization techniques. The findings unequivocally illustrate the efficacy of BBO in enhancing the precision and dependability of RF models, which is especially conspicuous in the context of financial forecasting where exactness is critical. The BBO-RF model that has been proposed yields productive results. The graphs showing the findings may be seen in Fig. 8 and Fig. 9. The training and testing data sets have therefore shown the BBO-RF model to have remarkably high accuracy. For predicting stock prices with remarkable accuracy, the BBO-RF model is an excellent resource.

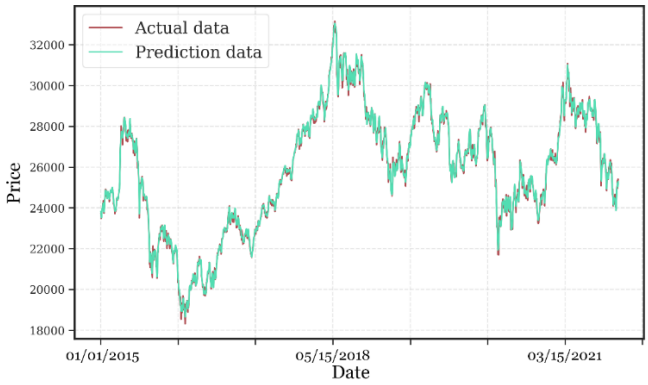

Fig. 8. Evaluation of the performance of the proposed model in comparison to real data during training.

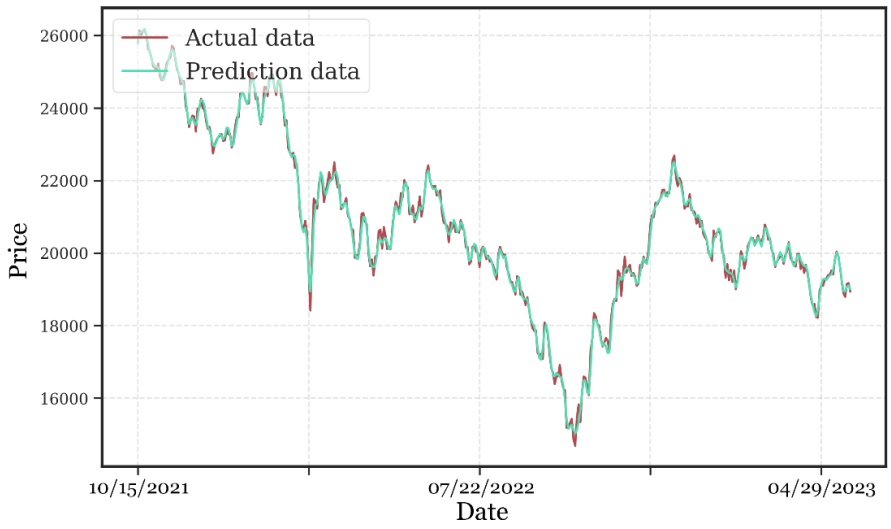

Fig. 9. Evaluation of the performance of the proposed model in comparison to real data during testing.

TABLE V. A COMPARISON BETWEEN THE EVALUATION AND PRIOR RESEARCH

References	Models	R^2
[40]	RNN	0.9784
	LSTM	0.9782
	Bi LSTM	0.9785
[41]	CNN-LSTM	0.9787
[42]	CNN-Bi LSTM	0.9787
	CNN-Bi LSTM-AM	0.9787
[43]	SDTP	0.9788
Current study	BRO-RF	0.992

R^2 values for several predictive models are displayed in comparison Table V. It is essential, when assessing the efficacy of our proposed BBO-RF method for predicting the stock market, to compare its performance to that of previously documented models. The findings of various models, including Bidirectional LSTM (Bi LSTM), Long Short-Term Memory (LSTM), and Recurrent Neural Networks (RNN), are detailed in [40]. The accuracy of these models varies, with RNN attaining a R^2 value of 0.9784, LSTM 0.9782, and Bi LSTM 0.9785. In a similar vein, Convolutional Neural Network (CNN)-based models such as CNN-LSTM, CNN-Bi LSTM, and CNN-Bi LSTM-AM are presented in [41], [42]. Each of these models has a R^2 value of 0.9787. The Stacked Denoising Transfer Learning Process (SDTP) is examined in [42], which provides a R^2 value of 0.9788. The BBO-RF model, which is presented in this study, attains a significantly elevated R^2 value of 0.992. The exceptional performance observed highlights the efficacy of our hybrid methodology in forecasting the stock market. The convergence of BBO and RF enables us to generate more precise forecasts by capitalizing on the combined advantages of optimization and machine learning methodology.

V. CONCLUSION

A robust market may boost confidence among consumers and companies, spurring additional economic growth. As such, the stock market can be used as an indication of the state of the economy overall. The analysis and discussion above make it evident that the study's findings offer insightful information about the prediction model's performance and accuracy. Essential markers of the model's efficacy are the statistical measures of RMSE, MAPE, MAE, and R^2. The Random Forest model has consistently shown remarkable predictive power. This article suggested a new, enhanced model for the technique, which was based on the original Random Forest approach. Machine learning algorithms heavily rely on optimization techniques to help identify the optimal solution to a given problem. To improve accuracy, machine learning models' parameters can be adjusted with the use of optimization techniques. Better judgment and more precise forecasts may result from this. To increase the efficiency of the model used in this research, three optimization methods were used, among which BBO obtained the best results.

This research uses HSI market data from 2015 to 2023 to forecast stock prices along with two Dow Jones and KOSPI indexes. This paper utilized a few optimization techniques, such

as biogeography-based optimization, genetic algorithms, and Grey Wolf Optimization. Out of all of these adjustments, biogeography-based optimization has produced the best results.

FUNDING

This work was supported by Heilongjiang Higher Education Teaching Reform Project (SJGY20220492).

Research on Ideological and political Function Construction and Education Paradigm of Finance Course under the background of "New Liberal Arts" construction (Shengwen Wu).

REFERENCES

[1] J. L. Ticknor, "A Bayesian regularized artificial neural network for stock market forecasting," Expert Syst Appl, vol. 40, no. 14, pp. 5501 – 5506, 2013, doi: 10.1016/j.eswa.2013.04.013.

[2] A. Arévalo, J. Niño, G. Hernández, and J. Sandoval, "High-frequency trading strategy based on deep neural networks," in International conference on intelligent computing, Springer, 2016, pp. 424–436.

[3] D. P. Gandhmal and K. Kumar, "Systematic analysis and review of stock market prediction techniques," Comput Sci Rev, vol. 34, p. 100190, 2019, doi: https://doi.org/10.1016/j.cosrev.2019.08.001.

[4] K. Chourmouziadis and P. D. Chatzoglou, "An intelligent short term stock trading fuzzy system for assisting investors in portfolio management," Expert Syst Appl, vol. 43, pp. 298–311, 2016.

[5] A. Emin, "Forecasting daily and sessional returns of the ISE-100 index with neural network models," Doğuş Üniversitesi Dergisi, vol. 8, no. 2, pp. 128–142, 2011.

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

[7] Y. Chen and Y. Hao, "A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction," Expert Syst Appl, vol. 80, pp. 340–355, 2017.

[8] V. S. Dave and K. Dutta, "Neural network based models for software effort estimation: a review," Artif Intell Rev, vol. 42, no. 2, pp. 295–307, 2014.

[9] A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.," 2022.

[10] H. J. Park, Y. Kim, and H. Y. Kim, "Stock market forecasting using a multi-task approach integrating long short-term memory and the random forest framework," Appl Soft Comput, vol. 114, p. 108106, 2022, doi: https://doi.org/10.1016/j.asoc.2021.108106.

[11] M. C. and L. C. and S. P. Cootes Tim F. and Ionita, "Robust and Accurate Shape Model Fitting Using Random Forest Regression Voting," in Computer Vision – ECCV 2012, S. and P. P. and S. Y. and S. C. Fitzgibbon Andrew and Lazebnik, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 278–291.

[12] 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.

[13] Z. Zhang, Y. Gao, Y. Liu, and W. Zuo, "A hybrid biogeography-based optimization algorithm to solve high-dimensional optimization problems and real-world engineering problems," Appl Soft Comput, vol. 144, p. 110514, 2023, doi: 10.1016/j.asoc.2023.110514.

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

[15] A. I. Hammouri, "A modified biogeography-based optimization algorithm with guided bed selection mechanism for patient admission scheduling problems," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 3, pp. 871–879, 2022, doi: 10.1016/j.jksuci.2020.01.013.

[16] A. Reihanian, M. R. Feizi-Derakhshi, and H. S. Aghdasi, "An enhanced multi-objective biogeography-based optimization for overlapping community detection in social networks with node attributes," Inf Sci (N Y), vol. 622, pp. 903–929, 2023, doi: 10.1016/j.ins.2022.11.125.

[17] F. Liu, B. Gu, S. Qin, K. Zhang, L. Cui, and G. Xie, "Power grid partition with improved biogeography-based optimization algorithm," Sustainable Energy Technologies and Assessments, vol. 46, no. April, p. 101267, 2021, doi: 10.1016/j.seta.2021.101267.

[18] Z. Cao, J. Li, Y. Fu, Z. Wang, H. Jia, and F. Tian, "An adaptive biogeography-based optimization with cumulative covariance matrix for rule-based network intrusion detection," Swarm Evol Comput, vol. 75, no. December 2021, p. 101199, 2022, doi: 10.1016/j.swevo.2022.101199.

[19] 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: 10.1016/j.aej.2023.04.054.

[20] D. G. Bhalke, D. Bhingarde, S. Deshmukh, and D. Dhere, "Stock Price Prediction Using Long Short Term Memory," SAMRIDDHI - A JOURNAL OF PHYSICAL SCIENCES, ENGINEERING & TECHNOLOGY; Vol 14 No Spl-2 issu (2022): A Journal of Physical Sciences, Engineering and Technology (2022);; 271-273 ; 2454-5767 ; 2229-7111, May 2022, [Online]. Available: https://myresearchjournals.com/index.php/SAMRIDDHI/article/view/11072

[21] X. Yuan, J. Yuan, T. Jiang, and Q. U. Ain, "Integrated Long-Term Stock Selection Models Based on Feature Selection and Machine Learning Algorithms for China Stock Market," IEEE Access, vol. 8, pp. 22672–22685, 2020, doi: 10.1109/ACCESS.2020.2969293.

[22] M. Vijh, D. Chandola, V. A. Tikkiwal, and A. Kumar, "Stock Closing Price Prediction using Machine Learning Techniques," Procedia Comput Sci, vol. 167, pp. 599–606, 2020, doi: https://doi.org/10.1016/j.procs.2020.03.326.

[23] A. Moghar and M. Hamiche, "Stock Market Prediction Using LSTM Recurrent Neural Network," Procedia Comput Sci, vol. 170, pp. 1168–1173, 2020, doi: https://doi.org/10.1016/j.procs.2020.03.049.

[24] W. Khan, U. Malik, M. A. Ghazanfar, M. A. Azam, K. H. Alyoubi, and A. S. Alfakeeh, "Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis," Soft comput, vol. 24, no. 15, pp. 11019–11043, 2020, doi: 10.1007/s00500-019-04347-y.

[25] M. Nabipour, P. Nayyeri, H. Jabani, S. Shahab, and A. Mosavi, "Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis," Ieee Access, vol. 8, pp. 150199–150212, 2020.

[26] H. Liu and Z. Long, "An improved deep learning model for predicting stock market price time series," Digit Signal Process, vol. 102, p. 102741, 2020, doi: https://doi.org/10.1016/j.dsp.2020.102741.

[27] I. R. Parray, S. S. Khurana, M. Kumar, and A. A. Altalbe, "Time series data analysis of stock price movement using machine learning techniques," Soft comput, vol. 24, no. 21, pp. 16509–16517, 2020, doi: 10.1007/s00500-020-04957-x.

[28] J. and D. A. Mehtab Sidra and Sen, "Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models," in Machine Learning and Metaheuristics Algorithms, and Applications, S. and L. K.-C. and B. S. and W. M. and S. D. Thampi Sabu M. and Piramuthu, Ed., Singapore: Springer Singapore, 2021, pp. 88–106.

[29] J. Ayala, M. García-Torres, J. L. V. Noguera, F. Gómez-Vela, and F. Divina, "Technical analysis strategy optimization using a machine learning approach in stock market indices," Knowl Based Syst, vol. 225, p. 107119, 2021, doi: https://doi.org/10.1016/j.knosys.2021.107119.

[30] 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, p. 100320, 2022, doi: https://doi.org/10.1016/j.mlwa.2022.100320.

[31] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," 2019.

[32] J. Choi, B. Gu, S. Chin, and J.-S. Lee, "Machine learning predictive model based on national data for fatal accidents of construction workers," Autom Constr, vol. 110, p. 102974, 2020.

[33] B. Gülmez, "Optimizing and comparison of market chain product distribution problem with different genetic algorithm versions," 2023.

[34] B. Gülmez and E. Korhan, "COVID-19 vaccine distribution time optimization with Genetic Algorithm," 2022.

[35] 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.

[36] B. Gülmez, "A novel deep neural network model based Xception and genetic algorithm for detection of COVID-19 from X-ray images," Ann Oper Res, vol. 328, no. 1, pp. 617–641, 2023.

[37] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer Adv Eng Softw 69: 46–61." ed, 2014.

[38] H. W. Kot, H. K. M. Leung, and G. Y. N. Tang, "The long-term performance of index additions and deletions: Evidence from the Hang Seng Index," International Review of Financial Analysis, vol. 42, pp. 407–420, 2015.

[39] L. N. Mintarya, J. N. M. Halim, C. Angie, S. Achmad, and A. Kurniawan, "Machine learning approaches in stock market prediction: A systematic literature review," Procedia Comput Sci, vol. 216, pp. 96–102, 2023, doi: 10.1016/j.procs.2022.12.115.

[40] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of LSTM and BiLSTM in forecasting time series," in 2019 IEEE International conference on big data (Big Data), IEEE, 2019, pp. 3285–3292.

[41] W. Lu, J. Li, Y. Li, A. Sun, and J. Wang, "A CNN-LSTM-based model to forecast stock prices," Complexity, vol. 2020, pp. 1–10, 2020.

[42] W. Lu, J. Li, J. Wang, and L. Qin, "A CNN-BiLSTM-AM method for stock price prediction," Neural Comput Appl, vol. 33, no. 10, pp. 4741–4753, 2021.

[43] Z. Tao, W. Wu, and J. Wang, "Series decomposition Transformer with period-correlation for stock market index prediction," Expert Syst Appl, vol. 237, p. 121424, 2024.