Analysis of the Financial Market via an Optimized Machine Learning Algorithm: A Case Study of the Nasdaq Index

Lei Wang, Mingzhu Xie*
School of Accounting and Finance, Anhui Xinhua University, Hefei Anhui, 230088, China

***Abstract*—The complex interaction among economic variables, market forces, and investor psychology presents a formidable obstacle to making accurate forecasts in the realm of finance. Moreover, the nonstationary, non-linear, and highly volatile nature of stock price time series data further compounds the difficulty of accurately predicting stock prices within the securities market. Traditional methods have the potential to enhance the precision of forecasting, although they concurrently introduce computational complexities that may lead to an increase in prediction mistakes. This paper presents a unique model that effectively handles several challenges by integrating the Moth Flame optimization technique with the random forest method. The hybrid model demonstrated superior efficacy and performance compared to other models in the present investigation. The model that was suggested exhibited a high level of efficacy, with little error and optimal performance. The study evaluated the efficacy of a suggested predictive model for forecasting stock prices by analyzing data from the Nasdaq index for the period spanning from January 1, 2015, to June 29, 2023. The results indicate that the proposed model is a reliable and effective approach for analyzing and forecasting the time series of stock prices. The experimental findings indicate that the proposed model exhibits superior performance in terms of predicting accuracy compared to other contemporary methodologies.**

Keywords*—Stock market prediction; Nasdaq index; random forest; moth-flame optimization; MFO-RF

I. Introduction

Retail and institutional investors can purchase and profitably sell shares of publicly traded corporations on the stock market. The stock market is a vital gauge of a nation's overall economic health since it represents company success and the business climate. Exchanges, both physical and virtual, are places where buyers and sellers come together to exchange assets [1, 2]. Traders and investors use a variety of strategies to evaluate equities and identify profitable opportunities. Many models have been created and put to the test to comprehend the underlying elements that influence stock prices. Both investors and scholars have long been interested in the study of stock price behavior. Fama conducted a study on this topic in 1965, which is now a major area of study in finance and has had a big impact on the way how currently understand stock price behavior [3], analyzing the stock market is challenging due to its volatile and ever-changing character. Analysts find it difficult to effectively analyze and anticipate price changes due to the market's noisy, chaotic, dynamic, non-linear, non-stationary, and nonparametric characteristics [3]. These qualities raise the possibility that conventional statistical techniques won't be enough for efficient stock market analysis.

To get around the challenges of making accurate stock market forecasts, academics have created several machine learning and artificial intelligence algorithms. These novel techniques aim to handle complex, noisy, chaotic, and non-linear stock market data more effectively than traditional time series methods, which often ignore the dynamic nature of the financial markets. Machine learning approaches employ sophisticated algorithms to analyze vast quantities of financial data and identify intricate patterns that may elude human observation. These methodologies have the capacity to provide more precise predictions through the analysis of an extensive array of data sources, including news articles, social media postings, and financial information. Furthermore, machine learning algorithms have the capability to consistently acquire knowledge and adapt to novel data, hence improving their predictive abilities as time progresses. In general, the utilization of artificial intelligence and machine learning has the potential to significantly enhance the precision of stock market prediction, providing investors with valuable information on market fluctuations and facilitating prudent investment choices [4].

Decision trees (DT) are a widely utilized machine learning approach that finds frequent application in both classification and regression tasks. This basic and comprehensible method is employed to construct a tree-like model that represents various choices and their corresponding probable outcomes [5]. The method partitions the data into subsets based on the values of the input features iteratively until a specified stopping condition is satisfied. Every partition is a node in the tree, and every node contains a decision rule that determines which feature will be separated first. DTs provide several benefits in comparison to other machine learning approaches. First of all, because they can handle both continuous and categorical data, they are suitable for a wide range of applications. Secondly, they are easy to use and need little feature engineering or data preparation.

Moreover, they are easily interpreted, which simplifies the process of understanding how the model arrived at a certain prediction. Despite its advantages, DTs have some disadvantages. For example, the tree may be susceptible to overfitting if it is deep and complex. Additionally, owing to their sensitivity to even minute changes in data, they could

*Corresponding Author. Email: dm35261478@126.com generate different trees for different training sets. To get around these problems, many DT variants have been developed; one of the best is random forest (RF).

RF employs an ensemble learning methodology to combine many decision trees, resulting in a robust and accurate model [6]. RF is a versatile machine-learning algorithm that can be applied to both classification and regression tasks. It has exceptional performance in handling large-scale datasets with high-dimensional features. The technique operates by constructing a collection of decision trees, whereby each tree is trained on a randomly selected portion of the characteristics and data. During the training process, every tree within the forest generates a forecast, and the collective projections of all the trees culminate in the ultimate prediction. This methodology enhances the model's robustness against noise and outliers, hence mitigating the risk of overfitting. When juxtaposed with alternative machine learning methodologies, RF offers several advantages. The versatility of this method is shown in its ability to effectively handle both continuous and categorical data, rendering it highly suitable for a wide range of applications.

Furthermore, this approach does not need an extensive feature engineering or data preprocessing, making it very straightforward to implement. In conclusion, the system exhibits a high degree of scalability and possesses the ability to effectively process extensive datasets comprising millions of samples and numerous attributes [6]. In their study, Park et al. [7] developed a comprehensive framework for predicting stock market trends by combining long short-term memory and random forest techniques. To evaluate the effectiveness of their proposed approach, the researchers utilized three prominent global stock indexes and incorporated 43 financial, technical indicators. In their study, Basher et al. [8] employed the RF algorithm to predict Bitcoin prices. Their findings indicate that the RF algorithm outperforms logit models in accurately anticipating trends in both Bitcoin and gold prices Basher et al. [8]. Illa et al. [9] proposed a methodology for estimating pattern-matching expectations by employing artificial intelligence techniques such as RF and support vector machines.

Artificial intelligence techniques are being more widely used in the stock market due to their capacity to process large amounts of data and identify intricate patterns that humans sometimes find difficult. It is important to keep in mind that these strategies' initial parameter configuration has a significant impact on how effective they are. Inaccurate estimates and outcomes might arise from improperly configured beginning settings. As such, it's important to pay close attention to the parameter settings while using artificial intelligence in stock trading. Consequently, there are numerous optimization algorithms that can be used to overcome these restrictions, such as the whale optimization algorithm (WOA) [10], particle swarm optimization (PSO) [11], Aquila optimizer (AO) [12], battel royal optimization (BRO) [13], biogeography-based optimization (BBO) [14], genetic algorithm (GA) [15], grey wolf optimization (GWO) [16], moth–flame optimization (MFO) [17], and others, can be used to get around these limitations. In 2015, S. Mirjalili proposed the MFO algorithm [17]. The MFO algorithm is a stochastic optimization technique inspired by the natural navigational mechanism of moths. Moths may navigate by keeping their angle concerning a far-off light source, like the moon or a flame, constant. The MFO method leverages this idea to optimize complex problems by varying the position and brightness of synthetic moths, which act as potential solutions. The algorithm explores the issue space to find the optimal answer as fast as possible. Justifications for Selecting the Proposed Model:

Addressing Complex Interactions: The model under consideration adeptly manages the complex interplay between investor psychology, market forces, economic variables, and market dynamics, which poses a significant challenge in the realm of financial forecasting. The comprehensive methodology, which merges the Moth Flame optimization and random forest processes, has been purposefully developed to address the intricacies that are intrinsic to the stock market.

Capability to Adapt to Non-Linear and Non-Stationary Conditions: Predicting stock price time series data presents a significant challenge due to their non-stationary, non-linear, and hypervolatile characteristics. The model being proposed is customized to effectively navigate these intricacies, rendering it well-suited for depicting the ever-changing dynamics of the stock market.

Addressing Computational Complexities: While conventional approaches may improve the accuracy of forecasts, they impose significant computational burdens. The complexities are effectively addressed by the proposed hybrid model, which guarantees precise predictions while maintaining computational efficiency.

Outstanding Efficacy and Performance: In comparison to the alternative models examined in the study, the hybrid model that integrated Moth Flame optimization and the random forest method consistently exhibited superior efficacy and performance. This demonstrates its resilience in addressing the unique difficulties associated with forecasting stock prices.

Optimal Performance and Minimal Error: The proposed model demonstrated an exceptional degree of effectiveness, characterized by minimal error. Preciseness is of the utmost importance when it comes to financial forecasting, as it enables one to make well-informed decisions.

The method proposed is thoroughly elucidated, effectively tackling the complex obstacles inherent in financial forecasting. The complex interaction among economic variables, market forces, and investor psychology poses a substantial obstacle to the ability to make precise forecasts in funding. The intricacy of this matter is compounded by the pronounced volatility, nonstationarity, and non-linearity of stock price time series data within the securities market. Acknowledging the constraints of conventional approaches, the article presents an innovative framework that adeptly surmounts these obstacles through the integration of the Moth Flame optimization methodology and the random forest method. Not only does this hybrid model demonstrate exceptional effectiveness, but it also surpasses other modern models examined in the study. The proposed model exhibits exceptional performance, minimal error, and high efficacy, providing a potentially viable resolution to the computational intricacies that are intrinsic to conventional forecasting approaches. The research assesses the predictive model that has been proposed by employing Nasdaq index data that covers the period from January 1, 2015, to June 29, 2023. The conclusive findings validate the efficacy and dependability of the suggested model in the domains of stock price analysis and prediction. The experimental results demonstrate that this methodology exhibits a higher level of predictive accuracy in comparison to other approaches. In brief, the methodology that has been proposed effectively tackles the complex issues associated with financial forecasting. It presents an innovative and successful approach that outperforms current models in terms of effectiveness and performance. The reliability of the model is further reinforced through its exhaustive evaluation and validation using real-world data, thereby establishing it as a significant contribution to the domain of stock price prediction. The research investigated many models, including RF, GA-RF, and PSO-RF, to assess their respective levels of reliability. The inquiry encompassed a comprehensive analysis of the data source and all relevant components in the subsequent section. A variety of analytical tools, including optimizer methods, evaluation metrics, and the RF model, were employed to examine the data. The findings of the study are given and compared with those obtained by alternative methodologies in the third part. The fourth part gives information about discussion of the results. The findings of the investigation are succinctly examined in the concluding section.

II. Methods and Materials

A. Random Forest

A well-liked machine learning technique for situations involving both regression and classification is the Random Forest algorithm, as seen in Fig. 1. It is a subset of the supervised learning algorithms of the support vector machines family. Two other popular tree-based techniques are naive Bayes and Adaboost. Breiman et al. [6] developed and presented the method, which is well known for being simple and efficient. The RF algorithm creates a variety of intricate decision trees, which improves forecast accuracy. The model is constructed decision trees by selecting the optimal feature from a given collection of characteristics in a non-deterministic manner, resulting in a lower level of predictability compared to alternative tree-based techniques. The methodology operates by iteratively training several decision trees through the utilization of bootstrapping, normalization, and bagging techniques. The grouping strategy involves the simultaneous construction of several decision trees, each using different subsets of characteristics and training data chunks. By employing bootstrapping to ensure the distinctiveness of each decision tree, the variance of the RF is reduced. The RF technique exhibits a high level of promise for generalizability due to the use of many tree-based models for evaluation. By employing this approach, the RF classifier can successfully mitigate challenges associated with unbalanced datasets and overfitting, hence surpassing the performance of current methodologies in accurately recognizing data.

Moreover, the methodology was specifically developed to address the analysis of datasets characterized by a large number of dimensions and strong interdependencies among variables. The reliability of the results increases proportionally with the number of trees included in the ensemble. The high level of precision exhibited by the RF approach can be attributed to the amalgamation of outputs derived from several decision trees. The utilization of ensemble methods mitigates the issue of overfitting and improves the predictive capabilities of the algorithm. Machine learning practitioners highly favor the RF method due to its ability to handle missing data and noisy inputs effectively. The above equation may be utilized to calculate the mean square error for an RF:

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

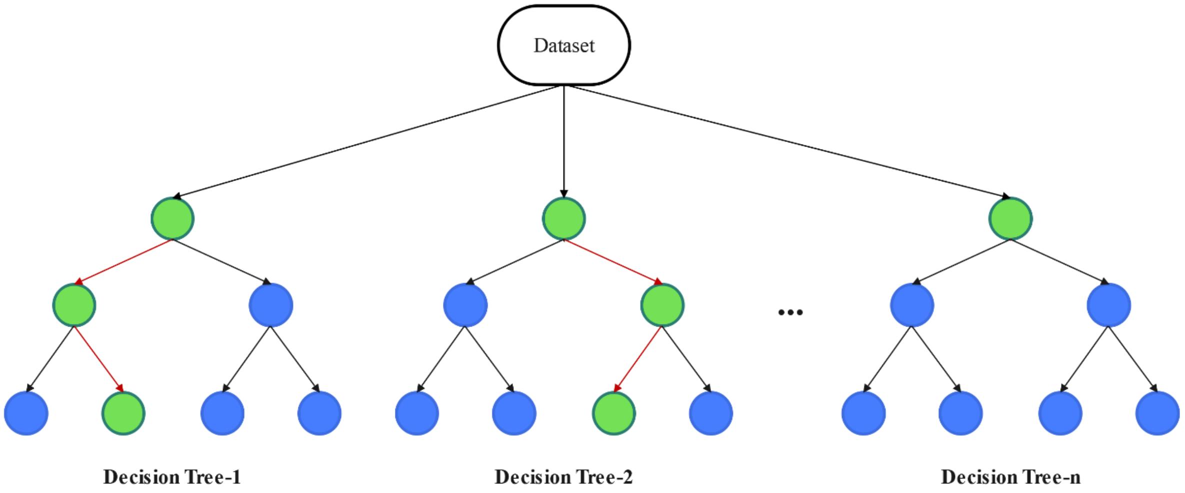

Fig. 1. The structure of Random forest.

B. Particle Swarm Optimization

PSO is an approach that, in order to get optimal results, imitates the cooperative habits of a flock of birds or a school of fish. Even if the exact location of the food supply is unclear at the outset, the swarm will nevertheless follow a set of rules to get there. By working together, the swarm is able to locate the source of nourishment. Swarms of fish or birds will eventually arrive at the near-optimal solution at the same moment. By following these three rules—separation, alignment, and cohesiveness—a bird swarm may efficiently navigate the search space and arrive at the correct solution [18]. Particles undergo separation by moving apart from each other to avoid overcrowding. The particles tend to align with their neighboring particles, resulting in a positional update influenced by the cohesion with said neighbors. Kennedy and Eberhart devised the PSO approach as a means to address optimization challenges. This method draws inspiration from the collective behavior observed in a swarm of particles [11]. PSO technique tends to converge rapidly and necessitates a limited number of parameters, hence reducing computational overhead.

Moreover, the likelihood of encountering a suboptimal local solution is reduced because of the extensive exploration conducted by several particles in quest for an optimal solution. In addition, the algorithm possesses an efficient global search mechanism and does not rely on derivatives. In the PSO, each particle searches a large search space to get the best possible answer. The search process begins with the random generation of candidate solutions, also known as particles, in the search space. Particle velocities and fitness scores are typically computed using a weighted mean of classification accuracy and the number of features in the feature subset. This computation aids in updating the velocity and heading of their trajectories after the initial iteration, and the method is continued until the stopping criterion is reached. The PSO algorithm's particles accelerate and decelerate per the following formula:

$$\boldsymbol{v_{id}^{t+1} = v_{id}^{t} + C_1 r_1^t \left(Pbest_{id}^{t} - x_{id}^{t}\right) + C_2 r_2^t \left(Gbest_{id}^{t} - x_{id}^{t}\right)} \quad (2)$$

The velocity of the ith particle at a given time iteration is denoted as $\boldsymbol{v_{id}^{k}}$ in a search space with d dimensions. The variables $\boldsymbol{Pbest_{id}^{t}}$ and $\boldsymbol{Gbest_{id}^{t}}$ represent the optimal particle and position for each individual and iteration of the ith function. The parameters $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$ are utilized to modify the velocity of particles, whereas $\boldsymbol{r_1^t}$ and $\boldsymbol{r_2^t}$ represent random values within the range of 0 to 1. Furthermore, the particles in the PSO algorithm have the ability to alter their locations by utilizing the equation shown below:

$$\boldsymbol{x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1}} \quad (3)$$

The variable $\boldsymbol{x_{id}^{t}}$ represents the spatial coordinates of the ith particle at iteration $\boldsymbol{t}$ inside a search space characterized by $\boldsymbol{d}$ dimensions.

C. Genetic Algorithm

The Genetic algorithm is a computational approach that emulates the mechanism of natural selection in order to address optimization and search problems [15]. Using this approach, a set of potential solutions referred to as individuals is generated. In order to generate novel individuals, these individuals are subsequently exposed to genetic mechanisms such as mutation, recombination, and selection. The assessment process employed in this study is iterative in nature and is repeated through several generations until a viable solution is identified. Consequently, the utilization of GA is prevalent throughout several areas, including but not limited to engineering, finance, and science [19]. GA is comprised of three fundamental components [20]. A chromosome refers to a sequence of numerical or textual symbols that are assigned to each individual by the encoding entity. The selection of an appropriate encoding technique is contingent upon the specific issue that has to be addressed.

Furthermore, the fitness function is utilized to evaluate the degree to which each individual's representation of the answer is accurate. The fitness function has been particularly tailored to address the current problem. To generate novel individuals from existing ones, the evolutionary operators employ the mechanisms of selection, crossover, and mutation. A crossover is a genetic process that combines the chromosomes of two individuals to generate a novel offspring. Mutation, on the other hand, introduces random alterations to an individual's chromosomes. Selection is employed to identify the most reproductively successful individuals.

D. Mouth Flame Optimization

The Moth Flame Optimizer is a computational model that draws inspiration from natural phenomena and is specifically influenced by the nighttime behavior of butterflies, which is displayed in Fig. 3. [17]. Butterflies have a consistent behavior of fluttering towards the moon when they are attracted by a light source. The Moth Flame Optimizer utilizes and formalizes this approach into an optimization algorithm, which is illustrated in Fig. 2. The optimizer exhibits versatility in its applicability to many optimization issues across several domains, including power and energy systems, economic dispatch, engineering design, image processing, and medical applications. Additionally, the optimizer derives inspiration from the behavior of butterflies as they navigate light sources. Researchers utilize the transverse orientation as a method to investigate the phenomenon of moths maintaining a straight flight trajectory toward the moon [21]. This examination explores the possible use of moths as solutions, which possess the ability to navigate in several dimensions, including $\boldsymbol{1D, 2D, 3D}$, and hyperdimensional space, by manipulating their position vectors. The focus of this study is on examining the spatial distributions of these moths, as they represent the aspects under consideration. The provided methodology guarantees convergence, and the Multi-Objective Firefly Algorithm (MFO) is both reliable and computationally efficient. MFO is commonly utilized as:

$$M = \begin{bmatrix} CO_{1,1} & CO_{1,2} & \cdots & \cdots & CO_{1,h} \\ CO_{2,1} & CO_{2,2} & \cdots & \cdots & CO_{2,h} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ CO_{a,1} & CO_{a,2} & \cdots & \cdots & CO_{n,h} \end{bmatrix} \quad (4)$$

In this context, h represents the number of dimensions, whereas a represents the number of moths.

$$S = \begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & \cdots & S_{1,h} \\ S_{2,1} & S_{2,2} & \cdots & \cdots & S_{2,h} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{a,1} & S_{a,2} & \cdots & \cdots & S_{2,h} \end{bmatrix} \quad (5)$$

The process of global optimization is conducted by the implementation of the three-step MFO approach.

$$MFO = (I, F, T) \quad (6)$$

Function $\boldsymbol{I}$ denotes a specific mathematical function, while $\boldsymbol{F}$ symbolizes the flight pattern of a moth as it navigates its environment in search of suitable space. Additionally, the symbol $\boldsymbol{T}$ is used to indicate the criteria that determine when the moth's flight comes to a halt.

$$X_i = t(C_i, S_j) \quad (7)$$

The formula employed in this context involves the twisting function denoted as $\boldsymbol{t}$, the number of the i-th moths represented by $\boldsymbol{C_i}$, and the number of the $\boldsymbol{j}$-th flames denoted as $\boldsymbol{S_j}$:

$$S(C_i, S_j) = Z_i \cdot e^{bt} \cdot cos(2\pi t) + S_j \, (8)$$

The variable $\boldsymbol{Z_i}$ represents the spatial separation between the moth and the flame. The constant $\boldsymbol{b}$ is a parameter in the context of this study. Additionally, the variable t is a random number selected from the interval [-1, 1].

$$\boldsymbol{Zi} = |\boldsymbol{S_j} - \boldsymbol{X_i}| \quad (9)$$

E. Data Collection and Preparing

To conduct a comprehensive analysis, it is important to include the trade volume as well as the open, high, low, and closing (OHLC) prices within a certain temporal interval. The data collection period was from January 2, 2015, to June 29, 2023, during which data was obtained from the Nasdaq index on the Yahoo Finance website. The precise details are encompassed inside the dataset that was employed for the investigation. A thorough data-cleaning procedure was conducted to ensure the accuracy and consistency of the forecasting models following the collection of the dataset. The implementation of a multi-step method was devised to safeguard the integrity of the dataset and prevent the inclusion of erroneous or incomplete data that might potentially lead to complications. The data were subjected to a thorough analysis in order to discover any anomalies, extreme numbers, or contradictions that could potentially compromise the validity of the results. This was one of the key aims. Several processes were utilized in order to clean and prepare the data in order to guarantee that it could be utilized. Several procedures, including scaling and normalization, were performed on the data in order to reduce the likelihood of gradient mistakes and unpredictable training results. Before beginning training, the data were normalized by employing the MinMaxScaler method. This was done in order to construct a stable model and reduce the likelihood of extremely high weight values occurring. This normalization process was achieved by employing the equation [22].

$$Xscaled = \frac{(\boldsymbol{X-Xmin})}{(\boldsymbol{Xmax-Xmin})} \quad (10)$$

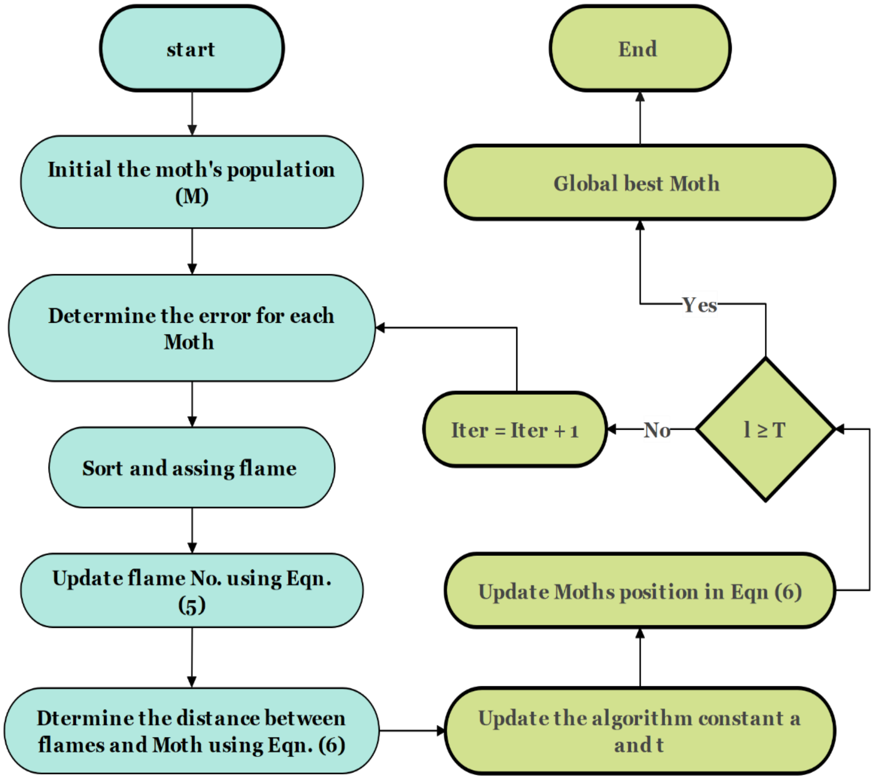
start
Initial the moth's population (M)
Determine the error for each Moth
Sort and assing flame
Update flame No. using Eqn. (5)
Dtermine the distance between flames and Moth using Eqn. (6)
Update the algorithm constant a and t
Update Moths position in Eqn (6)
I ≥ T
No
Iter = Iter + 1
Yes
Global best Moth
End

Fig. 2. The framework of MFO.

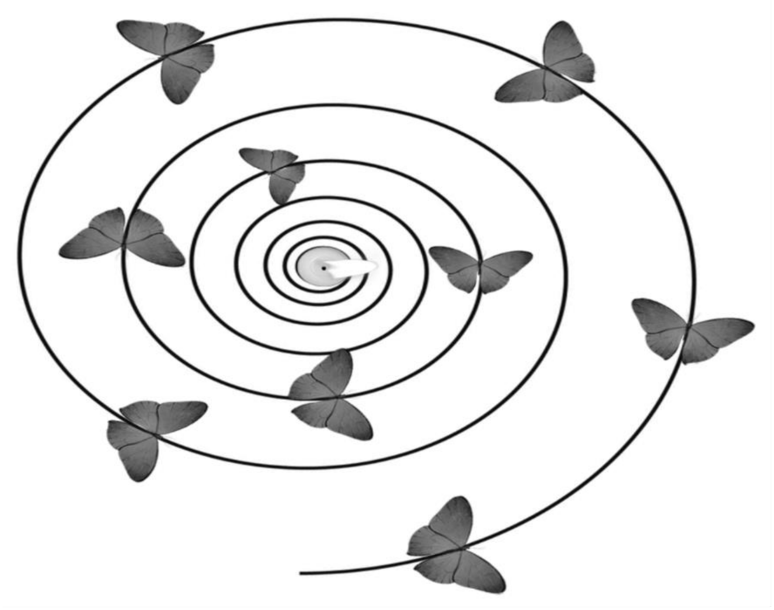

Fig. 3. Movement of the propeller towards the light source.

The training data provided to the model consisted of prices and volume for OHLC. The model was evaluated by including all features except for the close price data. The data were divided into two sets: 80% for training and 20% for testing, as seen in Fig. 4.

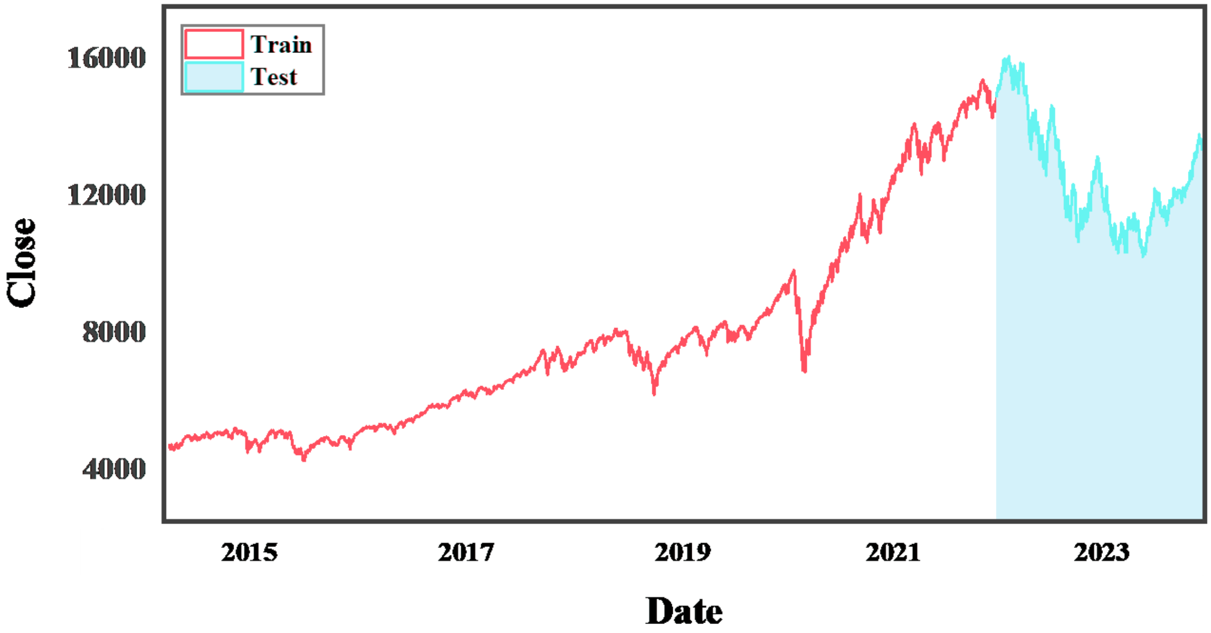

Fig. 4. Dividing the data into training and test.

F. Evaluation Metrics

The accuracy of the projected future was evaluated using a range of performance measures. Carefully selected, these measures provide a comprehensive assessment of the forecasts' validity and accuracy. The assessment method took into account a number of criteria. The Root Mean Square Error (RMSE) determines the root mean square of the errors between the predicted and actual values, the Mean Absolute Percentage Error (MAPE) computes the average absolute difference between the predicted and actual values, and the Coefficient of Determination ($\boldsymbol{R^2}$) quantifies the percentage of the dependent variable's variance that can be predicted based on the independent variable. These techniques help with and are very helpful for evaluating the forecasting models' accuracy.

$$MAPE = \left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|\right) \times 100 \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{12}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{13}$$

III. Results and Discussion

A. Hyperparameters Setting

Performance is significantly impacted by the parameters that machine learning algorithms employ. Under these circumstances, it is imperative to guarantee the precise specification of the model parameters. This particular segment provides a comprehensive explanation of the processes involved in hyperparameter configuration. The three optimizers are utilized in order to optimize the parameters of the RF model. In the context of problem-solving, the RF model demonstrates remarkable comfort in tasks that require classification and regression. In prior discussions, we have examined the upper and lower limits of the parameters employed in the configuration of the RF model. In addition to the optimal values derived by the primary optimizer, a detailed dissection of these limits is provided in Table I. Ultimately, the utilization of this data will aid in the determination of the most effective parameters for the RF model, thereby augmenting its overall performance.

TABLE I. Setting and Obtaining the Optimal Values of the Hyperparameters

Random Forest	Upper and lower bounds	Best values
Maximum depth	[10 -100 and None]	80
Maximum features	[Auto and squared]	auto
Minimum samples Leaf	[1 and 4]	2
Minimum samples split	[2 and 10]	2
Number of estimators	[200 and 2000]	500

B. Statistical Values

This phase of the inquiry encompasses Table II, which presents comprehensive statistical data about the dataset. The inclusion of OHLC price and volume statistics in the table enhances the clarity of the data. To comprehensively and accurately evaluate the data, statistical measures such as mean, count, minimum, maximum, standard deviation (Std.), and variations can be employed.

TABLE II. Statistics Summary for the Data Set

	Open	High	Low	Volume	Close
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
Max	16120.92	16212.23	16017.23	11621.19	16057.44
Variance	11107186	11304139	10878852	2406747	11102609

C. Outcomes of the Models

The primary objective of this study is to identify and assess the most effective hybrid algorithm for the prediction of stock prices. This research is grounded on the establishment of forecasting models and a comprehensive comprehension of the intricate aspects that impact stock market trends. The primary objective is to provide investors and analysts with valuable information that enables them to make informed and prudent investment choices. The models were processed by using different optimizers and different hyperparameters which were obtained by using those techniques. Table III and Fig. 5, 6 presents a comprehensive examination of the performance of each model. An exhaustive evaluation of the efficacy of each model is also incorporated.

TABLE III. Predicted Assessment Results for Benchmarking Approaches

	TRAIN SET			TEST SET		
	R^2	*RMSE*	*MAPE*	R^2	*RMSE*	*MAPE*
RF	0.979	423.25	4.96	0.974	254.85	1.61
GA-RF	0.983	382.32	4.07	0.978	234.67	1.50
PSO-RF	0.987	333.04	3.98	0.983	206.35	1.31
MFO-RF	0.992	260.90	1.70	0.988	173.45	1.07

IV. Discussion

This study aims to find the best hybrid stock price prediction algorithm. This study relies on predicting models and a deep understanding of stock market dynamics. The goal is to help investors and analysts make smart investment decisions. Fig. 5, 6 and Table III detail in each model's performance. Also included is a thorough model efficacy review. RMSE, MAPE, and R^2 were the three metrics that were utilized in the evaluation of the data analysis, other measures that were utilized included R^2 and RMSE. The ability of the aforementioned metrics to give a comprehensive evaluation of the correctness, dependability, and overall efficacy of the analysis is widely acknowledged and widely accepted. Both with and without the utilization of an optimizer, the performance of the RF model has been evaluated by utilizing the RMSE, MAPE, and R^2 criteria. Through the utilization of this approach, it is possible to acquire a more thorough understanding of the performance of the model and to make educated judgments based on the insights that are obtained. After doing an analysis on both the training and test sets, it was observed that the RF model, while it was not utilizing the optimizer, generated R^2 values of 0.979 and 0.974 for the training set and the testing set, respectively. While the MAPE values were 4.96 and 1.61, the RMSE values for the training and testing sets were 423.25 and 254.85, respectively. This is in contrast to the MAPE values, which were 4.96 and 1.61. There was a significant improvement in the effectiveness of the RF model as a result of the incorporation of optimizers. An increase in the R^2 value to 0.983 for the training dataset and 0.978 for the testing dataset is evidence that the utilization of the GA optimizer has led to considerable improvements. This can be observed by comparing the numbers. The RMSE and the MAPE have both shown a decline in both the training data set and the testing dataset. To be more specific, the RMSE values for the training set were 382.32 and the testing set 234.67, respectively. As an additional point of interest, the MAPE values for the training set are 4.07, whereas the values for the testing set are 1.50. As a result of doing a comparative analysis between the GA-RF model and the PSO-RF model, it has been concluded that the latter model shows superior performance. For the training phase, the R^2 values for the PSO-RF model were found to be 0.987, and for the testing phase, they were found to be 0.983. Despite the fact that the RMSE and MAPE values for training and testing both declined to 333.04 and 3.98, respectively, the values for testing decreased to 206.35 and 1.31, respectively. This is an important observation to make. According to these findings, the PSO-RF model is superior to the GA-RF model in terms of both its degree of effectiveness and its level of efficiency. The remarkable R^2 values of 0.992 and 0.988 for training and testing, respectively, demonstrate the efficacy of the MFO-RF model through its remarkable performance. Despite having the lowest possible testing value of 1.07, it performed exceptionally well. The MFO-RF model performed the best when compared to the other models, with the lowest RMSE values of 260.90 for training and 173.45 for testing. For comparison, the other models performed the worst. Considering that these results demonstrate how highly precise and reliable the MFO-RF model is, it is clear that it is an effective instrument for this particular application.

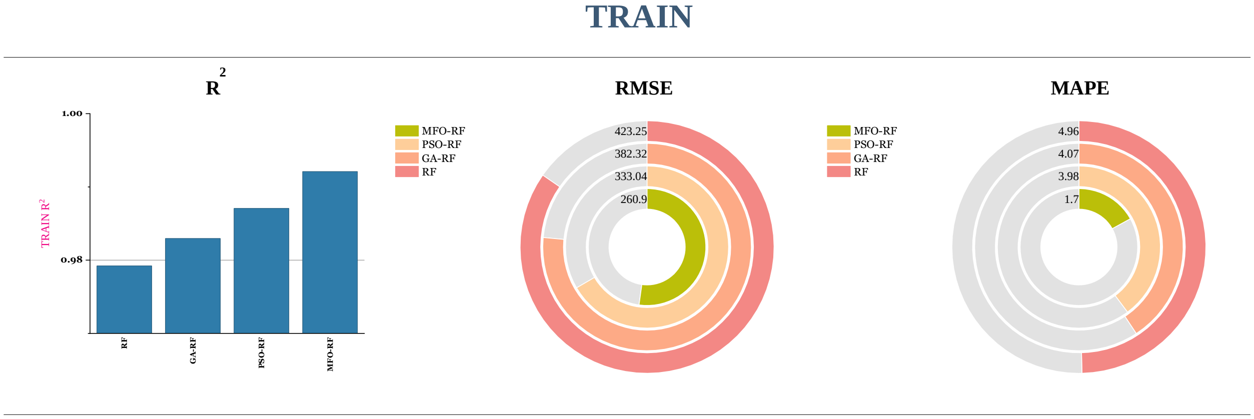

Fig. 5. The results obtained for $\boldsymbol{R^2}$, RMSE, and MAPE by the proposed model during training.

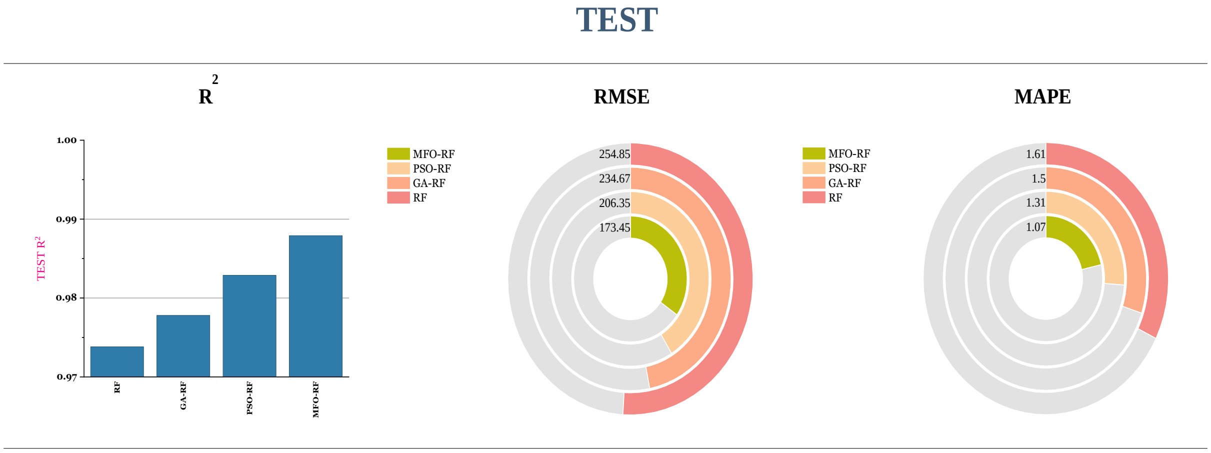

Fig. 6. The results obtained for $\boldsymbol{R^2}$, RMSE, and MAPE by the proposed model during testing.

After thorough research, the MFO-RF model is a reliable instrument for accurately forecasting stock values, thereby establishing its credibility. The efficacy of the model may be assessed by examining the Nasdaq index curves and comparing them to the corresponding curves depicted in Fig. 7 and Fig. 8. The MFO-RF model has superior performance in forecasting stock prices compared to other models, such as RF, GA-RF, and PSO-RF. A comprehensive analysis of the model's efficacy unveiled that the MFO-RF model predicts stock prices through the integration of the Mouth-Flame optimization technique and the random forest algorithm. The utilization of the RF technique not only mitigates fluctuations in stock prices but also enhances the precision of future trend predictions, hence further augmenting the accuracy of the model. One distinguishing characteristic of the MFO-RF model, in comparison to other models, is its ability to acquire knowledge from previous datasets. To accurately predict stock prices, a model must possess the capability to acquire knowledge from historical datasets and adjust its predictions in response to evolving market patterns. In summary, the MFO-RF model's reliability, accuracy, and ability to derive insights from historical datasets render it a highly valuable tool for predicting stock prices. The utilization of the RF algorithm and MFO optimizer, together with its adaptability in addressing dynamic market trends, positions it as the preferred option for those seeking to achieve profitable stock market transactions. These are the limitations of the research:

Temporal Scope: The study encompasses the period that commences on January 1, 2015, and concludes on June 29, 2023. The temporal scope may fail to encompass specific market conditions or events that transpired beyond the specified time period. Subsequent investigations may delve into the durability of the suggested algorithms across prolonged historical epochs.

Generalization of Algorithms: Although the suggested algorithms demonstrate exceptional performance when applied to Nasdaq data, their ability to be applied to diverse market conditions and financial instrument types is still unknown. It is critical to evaluate the adaptability of the algorithms to a wide range of datasets in order to obtain a thorough comprehension of their practicality.

Insufficient Comparative Research Against Non-Machine Learning Approaches: The study predominantly conducts a comparative analysis of various machine learning models, neglecting to delve deeply into their performance in relation to conventional forecasting methods or statistical methodologies. By incorporating these comparisons, a more comprehensive assessment of the proposed algorithms could be achieved.

The research investigates the impact of hyperparameter selection on sensitivity: To optimize hyperparameters, the study employs genetic algorithms, particle swarm optimization, and Moth Flame optimization. Nevertheless, the explicit consideration of the algorithms' sensitivity to various sets of hyperparameters is absent. A more comprehensive sensitivity analysis may yield valuable insights regarding the models' stability.

Market Volatility Attributable to Inherent Market Volatility: Predicting stock prices necessitates addressing vagaries. The accuracy of the proposed algorithms is commendable; however, the unpredictability inherent in financial markets may introduce unforeseen fluctuations that have an adverse effect on the precision of forecasts.

The potential compromise of model interpretability may arise from the complexity of the proposed algorithms, particularly when they are integrated with optimization techniques. Comprehending the fundamental mechanisms that propel predictions may present a formidable task, thereby constraining the model's applicability in specific decision-making contexts.

Alterations in Market Dynamics: Throughout the period under analysis, the study presupposes a stable set of market dynamics. Variations in economic policies, geopolitical occurrences, or worldwide economic transformations might give rise to modifications in market conduct that the suggested algorithms might not sufficiently account for.

Overfitting Risk: It is crucial to recognize the potential for overfitting, particularly when algorithms are being optimized for particular datasets. While the models might exhibit outstanding performance on the training data, they might encounter difficulties when implemented on unseen data, which raises doubts about their efficacy in real-world scenarios.

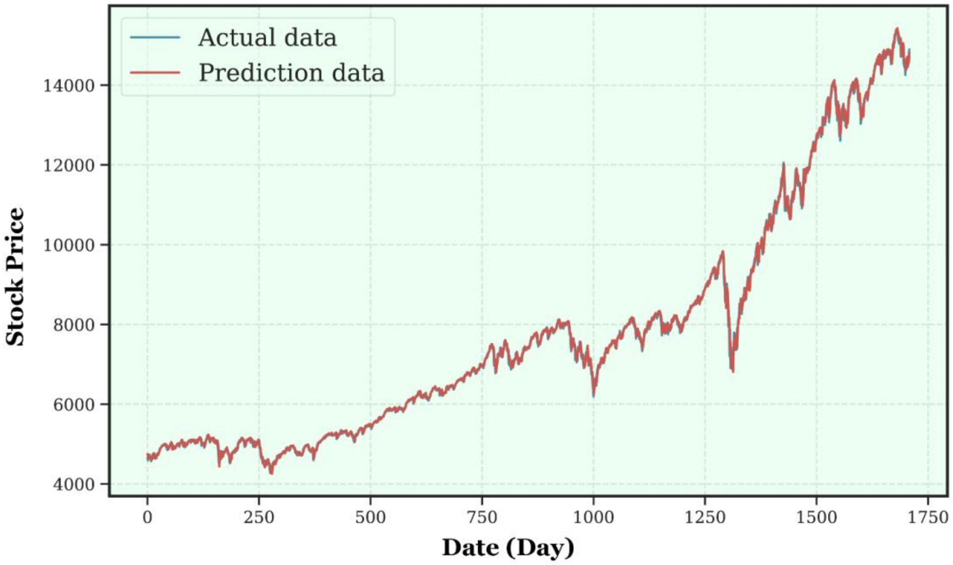

Fig. 7. The prediction curve generated via MFO-RF during training.

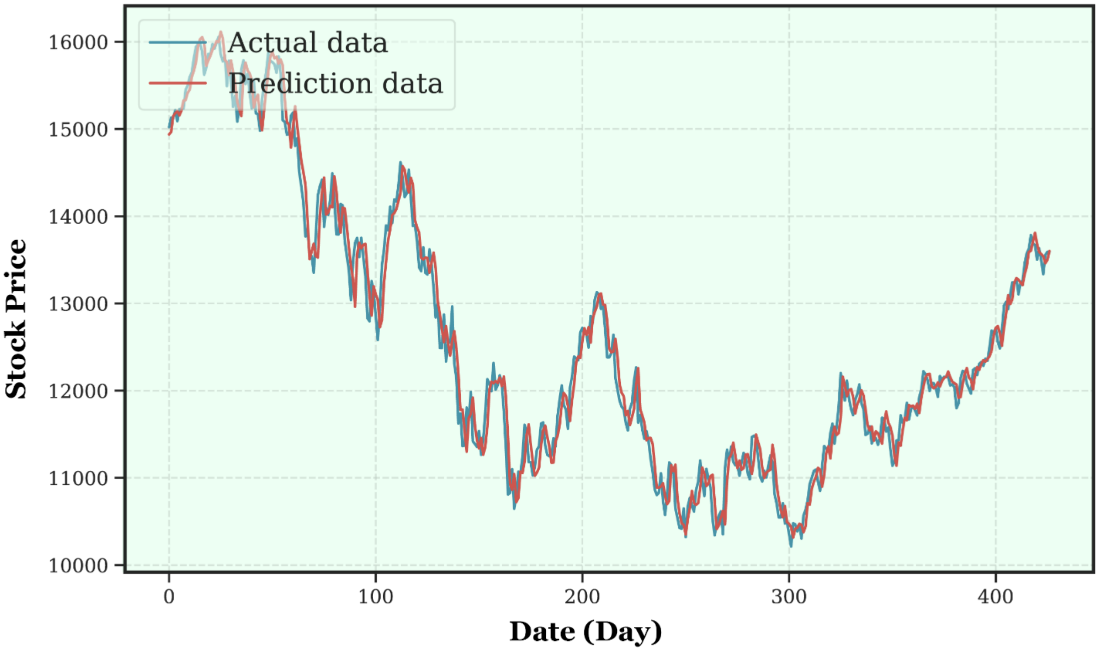

Fig. 8. The prediction curve generated via MFO-RF during testing.

V. Conclusion

The process of projecting stock prices is a complex and involved task that involves several interconnected components. The stock market is subject to several influences, including but not limited to politics, society, and the economy. It is a complex and always-changing system. To make accurate predictions on future stock values, it is important to consider a range of financial statements, earnings reports, market trends, and other relevant elements. Moreover, it is important to note that macroeconomic factors, such as interest rates, inflation, and worldwide market conditions, wield significant influence over the behavior of the stock market. Developing accurate and dependable prediction models can pose significant challenges owing to the intricate nature and multitude of factors inherent in forecasting stock prices. Understanding the unpredictable and non-linear nature of the market is essential to making accurate forecasts. Fortunately, the MFO-RF model offers a practical answer to these problems and has shown to be accurate and trustworthy. The effectiveness of many stock price prediction models, such as RF, GA-RF, and PSO-RF, was assessed in the current study. By employing the GA, PSO, and MFO hyperparameter optimization techniques, the RF's parameters were enhanced.

Nevertheless, when combined with RF, the MFO optimizer method produced the best results. The OHLC prices and volume for the Nasdaq index from January 2, 2015, to June 29, 2023, made up the dataset utilized in the study. The results of the investigation demonstrate how reliable and accurate the MFO-RF model is at forecasting stock prices.

Throughout the study, the accuracy and predictive capability of the MFO-RF model were evaluated by comparing it to many other models. Based on the obtained data, it can be concluded that the MFO-RF model consistently exhibited superior performance compared to the other models. The testing $\boldsymbol{R^2}$ score of 0.988 indicates a good level of accuracy in the predictions made. The RMSE value of the model, which was observed to be 173.45, indicates that the model's predictions exhibited a satisfactory level of accuracy. The model had a low MAPE score of 1.07, suggesting a consistent ability to provide reliable predictions. In terms of accuracy and efficacy, the MFO-RF model demonstrated superior performance compared to the other models that were examined.

The MFO-RF model is a helpful resource for stock price prediction in general and offers insightful data to investors who are attempting to make well-informed investment decisions.

Funding

Anhui Provincial Social Science Innovation Development Project： Aggregation of Emerging Industries in Anhui Province: Efficiency Evaluation and Path Optimization (NO.2021CX025). The first-class undergraduate professional construction fund of Anhui Xinhua University (NO.2020ylzyx05).

References

[1] C. Zhang, J. Ding, J. Zhan, and D. Li, “Incomplete three-way multi-attribute group decision making based on adjustable multigranulation Pythagorean fuzzy probabilistic rough sets,” International Journal of Approximate Reasoning, vol. 147, pp. 40–59, 2022.

[2] Y.-H. Wang, C.-H. Yeh, H.-W. V. Young, K. Hu, and M.-T. Lo, “On the computational complexity of the empirical mode decomposition algorithm,” Physica A: Statistical Mechanics and its Applications, vol. 400, pp. 159–167, 2014, doi: https://doi.org/10.1016/j.physa.2014.01.020.

[3] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, “Machine learning techniques and data for stock market forecasting: A literature review,” Expert Syst Appl, vol. 197, no. December 2021, p. 116659, 2022, doi: 10.1016/j.eswa.2022.116659.

[4] 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.

[5] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, “An introduction to decision tree modeling,” Journal of Chemometrics: A Journal of the Chemometrics Society, vol. 18, no. 6, pp. 275–285, 2004.

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

[7] 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.

[8] S. A. Basher and P. Sadorsky, “Forecasting Bitcoin price direction with random forests: How important are interest rates, inflation, and market volatility?,” Machine Learning with Applications, vol. 9, p. 100355, 2022.

[9] P. K. Illa, B. Parvathala, and A. K. Sharma, “Stock price prediction methodology using random forest algorithm and support vector machine,” Mater Today Proc, vol. 56, pp. 1776–1782, 2022.

[10] S. Mirjalili and A. Lewis, “The whale optimization algorithm,” Advances in engineering software, vol. 95, pp. 51–67, 2016.

[11] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in Proceedings of ICNN’95-international conference on neural networks, IEEE, 1995, pp. 1942–1948.

[12] L. Abualigah, D. Yousri, M. Abd Elaziz, A. A. Ewees, M. A. A. Al-Qaness, and A. H. Gandomi, “Aquila optimizer: a novel meta-heuristic optimization algorithm,” Comput Ind Eng, vol. 157, p. 107250, 2021.

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

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

[15] 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.

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

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

[18] C. W. Reynolds, “Flocks, herds and schools: A distributed behavioral model,” in Proceedings of the 14th annual conference on Computer graphics and interactive techniques, 1987, pp. 25–34.

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

[20] 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.

[21] S. Mohammad, A. Laith, A. Hamzeh, A. Mohammad, and A. M. Khasawneh, “Moth–flame optimization algorithm: variants and applications,” Neural Comput Appl, vol. 32, no. 14, pp. 9859–9884, 2020.

[22] P. J. M. Ali, R. H. Faraj, E. Koya, P. J. M. Ali, and R. H. Faraj, “Data normalization and standardization: a technical report,” Mach Learn Tech Rep, vol. 1, no. 1, pp. 1–6, 2014.