Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2022, 9(5):33-47**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Forecasting Stock Prices by Applying the Whale Optimization Algorithm and Genetic Programming

Chih-Ming Hsu*

*Department of Business Administration, Minghsin University of Science and Technology, Hsinchu, Taiwan

Abstract It is a critical task to accurately forecast stock prices in the future for an investor to make more money in the dynamic stock market. However, there are various factors, such as politics, business trade cycle, government financial policy, variation of an exchange rate, inflation, as well as business operation of a corporation, etc., that can affect the stock prices issued by a corporation, thus making the problems of forecasting the stock prices very complicated and difficult to resolve. Hence, the problems regarding forecasting stock prices always can attract the great interest of researchers and practitioners. In this study, the whale optimization algorithm (WOA) and genetic programming (GP) are utilized to propose a systematic approach to tackle the problems of forecasting stock prices. The WOA intends to divide the stocks into appropriate groups such that the corporations within the same group can have business operation of a similar style, thus making the modelling process easier. The GP is applied to construct the forecasting model for each cluster. Theses forecasting models for all clusters are then combined into an integrated one. The usefulness, effectiveness, as well as efficiency of our proposed forecasting approach is demonstrated by a real case study aiming to forecast the closing prices in the next day for stocks listed in the Taiwan Stock Exchange Corporation (TWSE). The experimental results show that our proposed method can outperform the single GP forecasting model, thus the clustering mechanism by using the WOA can be proven an effective method for clustering the data thus lowering the difficulty of modelling and improving the forecasting accuracy.

Keywords Stock prices, Forecasting, Whale optimization algorithm, Genetic programming, Clustering, TWSE

1. Introduction

In our daily life, there are various kinds of forecasting problems arisen in different fields, e.g. science, engineering, culture, finance, election, gambling, etc. In the domain of finance, forecasting the stock prices in the future accurately is an important task for an investor to make money in the stock market. However, the factors that can affect the trend regarding the stock prices might diverse, such as business trade cycle, financial policy, variation of an exchange rate, inflation, operation of a corporation, etc. Hence, the forecasting of stock prices is always a very complicated and difficult problem, thus attracting the great interest from researchers and practitioners. Because of the enhancement of computers' efficacy, as well as the decreasing of computers' hardware, many heuristic optimization algorithms that are inspired from different natural phenomena had been used and demonstrated their effectiveness and usefulness in tackling such a complex problem of forecasting stock prices. For example, [1] combined the stock time series and stock chart images, which are the features learned from different representing for the same data, to propose a feature fusion long short-term memory-convolutional neural network, shortly named LSTM-CNN, model to predict stock prices. In their study, the LSTM and CNN are used to extract the temporal features and image features. The performance of their proposed approach relative to those using a single model, i.e. CNN or LSTM, is measured by evaluating the SPDR S&P 500 ETF data, and the experimental results reveal that their proposed method can outperform the

single CNN or LSTM model for predicting the stock prices. Furthermore, the authors find that it is the most appropriate to apply a candlestick chart to demonstrate the stock chart image for predicting the stock prices, thus the combination of temporal and image features from the same data can efficiently reduce the prediction error. [2] utilized technical indicators that focus on the interpretability-accuracy trade-off to proposes a neuro-fuzzy system, with the efficiency and the interpretation, for predicting the stock prices. The interpretability of their proposed system can be ensured through two manners. First, the rule base reduction is made by selecting the best rules based on the rule performance criteria to yield an efficient and compact rule base that can be comprehend easily. Second, the constrained learning during the model optimization stage makes the simple constraints can be imposed on the updates of fuzzy set parameters. The daily stock data of Bombay Stock Exchange, CNX Nifty and S&P 500 stock indices are used to evaluate their proposed system. According to the simulation results, their proposed approach can attain a better balance between accuracy and interpretability while comparing to the other two artificial intelligence techniques, as well as two statistical methods that are common tools for predicting the stock prices. [3] proposed a hybrid ARI-MA-LS-SVM model to create basic predictions for the stock market through analyzing the defects of the current approaches for predicting stock markets, as well as the standard support vector machines. A cumulative auto-regressive moving average that combines the least squares support vector machine is then proposed. Next, the predictive indicators are processed by utilizing the cumulative auto-regressive moving average. The least squares support vector machine with a simple indicator system is then applied to predict the fluctuations regarding the stock prices. Based on numerous simulation experiments, their proposed method can provide the general applicability, the market applicability, as well as the feasibility. [4] proposed a model by hybridizing the empirical mode decomposition (EMD), convolutional neural network (CNN) and Long Short-Term Memory (LSTM) to predict stock prices. The EMD is applied to decompose the original stock price series to form a finite number of intrinsic mode functions (IMFs) under different frequencies. Then, the features are extracted through a CNN for each component. Finally, the temporal dependencies of all extracted features are modeled, thus the prediction for the stock prices can be obtained by using a linear transformation. Their proposed approach is evaluated by testing the two prediction steps, one day and one week, for the Shanghai Stock Exchange Composite Index (SSE). Compared with other state-of-the-art models, their proposed method can achieve the better performance based on the experimental results. [5] combined the CNN and LSTM to propose a method to forecast the stock prices. The stock prices are also predicted by utilizing the MLP, CNN, RNN, LSTM, CNN-RNN, and other forecasting models, as well as the forecasting results from these models are analyzed and compared. The experimental data are gathered from the daily stock prices of 7127 trading days during July 1, 1991 and August 31, 2020. Eight features, including opening price, highest price, lowest price, closing price, volume, turnover, ups and downs, and change, are chosen and the CNN is adopted to extract the features from these data efficiently. The LSTM is then applied to predict the stock prices through using the extracted feature data. The experimental results show that their proposed CNN-LSTM model can forecast the stock prices with reliability and the highest prediction accuracy. [6] proposed a systematic LLE-BP approach based on a local linear embedding dimensional reduction algorithm (LLE) and back propagation (BP) neural network to resolve the problems of stock price forecasting. The dimension of the factors that can affect the stock prices is first reduced by the LLE. The BP neural network is then utilized to predict the stock prices where the dimension-reduced data takes as the input variables of BP. Their proposed method is compared with the traditional BP neural network model, PCA-BP model, and ARIMA (3,1,1) model. The experimental results show that the LLE-BP neural network model can obtain the higher accuracy in predicting stock prices, as well as it is a stock price prediction approach with effectiveness and feasibility. [7] applied the extreme learning machine (ELM) optimized by the differential evolution (DE) algorithm based on the secondary decomposition techniques of variational mode decomposition (VMD) and ensemble empirical mode decomposition (EEMD) to propose a hybrid model of predicting stock prices. The original stock index price sequence is first processed through the VMD technology thus obtaining different modal components and the residual item that is further handled by the EEMD method. The prediction results of the DE-ELM model for each modal component and the residual item are then superimposed to obtain the final prediction results. Their proposed model is demonstrated by verifying a series of benchmark models, as well as testing the samples of the S&P 500 index and the HS300 index by one-step, three-step, and five-step forward

forecasting. Based on the empirical results, their proposed hybrid model can achieve the best prediction performance for all prediction scenarios. In addition, the model combined with the decomposition technology can yield the superior prediction performance compared to the single DE-ELM model. [8] combined the Long Short-Term Memory (LSTM) with Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) to develop a mixture model to forecast the stock prices in the Standard & Poor's 500 index (S&P500) and China Securities 300 Index (CSI300) markets. The original data is decomposed by using the CEEMDAN technique thus obtaining several IMFs and one residue. The LSTM forecasting model is then applied to the decomposed data to yield the prediction sequences that are further reconstructed to acquire the final prediction. They also construct some contrast models including Support Vector Machine (SVM), Backward Propagation (BP), Elman network, Wavelet Neural Networks (WAV), as well as their mixture models combined with the CEEMDAN method. These models are tested with the MCS (Model Confidence Set) evaluation criterion. According to the empirical results, the CEEMDAN-LSTM approach can provide the optimal forecasting performance in the developed and emerging stock markets. [9] applied the decision tree model, CRISP-DM (Cross-Industry Standard Process for data mining) for analysis through using the WEKA software based on the data mining techniques thus aiming to predict stock prices in an emerging market. The sample in their study includes ten firms in five different sectors listed on Pakistan Stock Exchange (PSX). The experimental findings indicated that the accuracy ratios can attain a range from 50% to 60%, and the market participants can disclose the higher returns through considering the embedded information in their previous performance of stock prices. Furthermore, the investors can take more prudent buy-sell decisions according to their experimental results. [10] improved the shortcomings of the traditional self-adaptive optimized grey model, including (1) extracting more value through sufficiently exploring the new data without information lapses to generate a dynamic weighted sequence, (2) various samples are adjusted thus augmenting the applicability of the proposed model by using the weighted coefficient and modified initial condition, and (3) the background value is reconstructed by applying Simpson's formula as well as is then integrated with the modified initial condition thus smoothing the data saltation, to propose a new self-adaptive optimized grey model. In their study, four cases regarding the sales and stock of EVs (electric vehicles) are simulated by their proposed approach to compare with six benchmarks to verify the rationality and efficacy of their model. The demonstration results show that their improved model outperforms in the forecasting precision for most cases. [11] proposed a multiple-staged method to predict the "ups and downs" of stock market prices. First, the attributes of a stock dataset are analyzed by using a correlation analysis, thus processing the missing values as well as determining the attributes to form the retained data which is then further divided into the training and testing sets. The retained attributes are subsequently predicted based on the LSTM model, and a new testing set is constructed according to the retention of prediction results. Furthermore, the training for the original training set is performed by the Bo-XG Boost model based on XGBoost. Four evaluation indexes, including the root mean square error (RMSE), average absolute error (MAE), accuracy rate and F1-score, are used to evaluated their proposed LSTM-BO-XGBoost model in forecasting the "ES=F", "YM=F", "CL=F", "<^>TNX", "<^>N225", "NQ=F", "AAPL", "GC=F", "JPY=X" and "SI=F" rates with 10 stocks. According to experiments, the LSTM-BO-XGBoost model can perform better than the LSTM in predicting stock prices. In addition, their proposed model is also compared with the single LSTM network model and RNN network model, as well as the LSTM-BO-XGBoost hybrid model. The empirical results show that the LSTM-BO-XGBoost model provides high performance, stability and feasibility compared to the others. [12] integrated the modified crow search algorithm (CSA) and extreme learning machine (ELM) to enhance the forecasting performance in stock markets. Their proposed Particle Swarm Optimization (PSO)-based Group oriented CSA (PGCSA) can outperform other current algorithms through solving 12 benchmark problems. The PGCSA algorithm is then applied to acquire the relevant weights and biases of ELM thus further improving the effectiveness of the traditional ELM. The performance measures, technical indicators and hypothesis test (paired t-test) are utilized to evaluate the effects of their propose hybrid PGCSA ELM model in predicting the next day closing prices in seven different stock indices. Notably, the seven stock indices are considered by incorporating data during COVID-19 outbreak, and their proposed model is also compared to the existing techniques proposed in published works. Based on the simulation results, the PGCSA ELM model can be considered as an appropriate tool for predicting the closing

prices in the next day. [13] proposed prediction model of stock prices to overcome the shortcomings of the traditional neural network based on an improved time convolution network (TCN) method. In addition, the trading data in the stock market are utilized, as well as the preprocessed data of financial news are fed into the model for training thus improving the prediction performance. The most representative 50 stocks with a sufficient large scale and good liquidity in Shanghai stock market are selected as sample stocks, as well as the information texts picked from the financial web pages takes as samples to demonstrate their proposed method to predict the rise/fall direction for the SSE 50 Index. The model structure is adjusted by altering the hyperparameters for the network structure, and the prediction performance is then compared with other models. Experimental results indicate that their proposed improved TCN model can significantly improve the effect for predicting the rise/fall direction for the SSE 50 index. Furthermore, the improved TCN model can train the model and predict the stock prices more efficiently. [14] applied the CNN-based model to forecast the stock prices for the aim to achieve the better forecasting performance as well as to avoid losing value information that might happen in the transformation process of images. Their proposed approach can efficiently extract the features of images while forecasting. In their study, the CNN forecasting model consisting of three convolutional layers and five full connected layers, and the nonlinear relation between input and output Relu and Elu activation functions can be determined is used. The financial time series of the Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) and Financial Time Stock Exchange for London stock market data (FTSE) that are frequently evaluated are used to demonstrate their proposed framework. The results from some aspects as an error criterion, a regression analyses and a visual demonstration show that the CNN structure can provide outstanding forecasts compared to some other state-of-the-art forecasting methods, e.g. ANN, LSTM, fuzzy-based approaches, and some traditional methods.

The above literature review shows that various kind of heuristic optimization algorithms had been successfully applied to resolve the forecasting problems regarding the stock prices. Furthermore, these algorithms can not only provide the superior effectiveness but also give the high efficiency in dealing with the complex stock price forecasting problems compared to those traditional methods. However, the factors that influence the stock prices are diverse. Among these factors, the business operation of a corporation is fundamental to determine the performance of stocks issued by this firm since it can reflect investors' basic expectation for this corporation in the future. Therefore, the value of a corporation should vary according to the nature regarding its business operation, so as to determine the stock performance in the market. Therefore, this study aims to divide the stocks into appropriate groups based on the fundamental business operations of corporations. The corporations within the same group have business operation with the most similar style, thus having stock prices' performance of the similar manner. Then, the forecasting technique is applied to construct the stock price forecasting model for those stocks within the same group in order to provide the better forecasting performance for these stocks issued by the corporations with business operation of the similar style. Hence, the whale optimization algorithm (WOA) and genetic programming (GP) are sequentially applied to develop a systematic approach to tackle the forecasting problems of stock prices in this study. The remaining sections are organized as follows. Section 2 briefly introduces the analyzing and modelling methods including the WOA and GP. Our proposed forecasting approach is then presented in Section 3. In Section 4, a real case study that aims to forecast the closing prices of stocks listed in the Taiwan Stock Exchange Corporation (TWSE) is provided to demonstrate the usefulness, effectiveness, as well as efficiency of our proposed forecasting approach. Finally, the conclusions are provided in Section 5.

2. Methodologies

Whale Optimization Algorithm

The whale optimization algorithm proposed by [15] is a meta-heuristic optimization algorithm that inspired by the social behavior of humpback whales to simulate the bubble-net hunting strategy. There are three mechanisms including (1) encircling prey, (2) spiral bubble-net feeding maneuver and (3) search for prey, in WOA. According to the experimental results of applying the WOA to 29 mathematical optimization problems and 6 structural design problem, WOA algorithm can be proven to be is very competitive compared to the other

state-of-art meta-heuristic and conventional optimizations methods. In WOA, the behavior of encircling prey can be expressed by Equations (1) to (4).

$$\vec{X}(t+1) = \overrightarrow{X^*}(t) - \vec{A} \cdot \vec{D} \quad (1)$$

$$\vec{A} = 2\vec{a} \cdot \vec{r} - \vec{a} \quad (2)$$

$$\vec{D} = |\vec{C} \cdot \overrightarrow{X^*}(t) - \vec{X}(t)| \quad (3)$$

$$\vec{C} = 2 \cdot \vec{r} \quad (4)$$

where t is the current execution iteration, $\vec{A}$ and $\vec{C}$ are coefficient vectors, X^* represents the position vector of the best solution found so far, $\vec{X}(t)$ is the position vector, $||$ indicates the absolute value, $\cdot$ is an element-by-element multiplication, $\vec{a}$ is number that linearly decreases from 2 to 0 over the course of iterations, and $\vec{r}$ is a random vector in [0,1].

Next, the bubble-net behavior of humpback whales can be mathematically designed as follows:

1. Shrinking encircling mechanism

This behavior is achieved by decreasing the value of $\vec{a}$ in Equation (2) to allow $\vec{A}$ to be a random value lying in the interval [−a, a].

2. Spiral updating position

The helix-shaped movement of humpback whales is mimicked by Equation (5).

$$\vec{X}(t+1) = \overrightarrow{D'} \cdot e^{bl} \cdot cos(2\pi l) + \overrightarrow{X^*}(t) \quad (5)$$

where $\overrightarrow{D'} = |\overrightarrow{X^*}(t) - \vec{X}(t)|$ to represents the distance of the ith whale to the prey, i.e. the best solution found so far, b is a constant to define the shape of the logarithmic spiral, l is a random number in [−1,1], and $\cdot$ is an element-by-element multiplication.

In addition, there is a probability of 50% to choose between either the shrinking encircling mechanism or the spiral model for updating the position of whales in WOA to simulate the humpback whales swim around the prey within a shrinking circle and along a spiral-shaped path simultaneously through Equation (6).

$$\vec{X}(t+1) = \begin{cases} \overrightarrow{X^*}(t) - \vec{A} \cdot \vec{D} & if \quad p < 0.5 \\ \overrightarrow{D'} \cdot e^{bl} \cdot cos(2\pi l) + \overrightarrow{X^*}(t) & if \quad p \geq 0.5 \end{cases} \quad (6)$$

where p is a random number in [0,1].

Finally, the search for prey in the exploration of a whale can be modelled by Equations (7) and (8).

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A} \cdot \vec{D} \quad (7)$$

$$\vec{D} = |\vec{C} \cdot \overrightarrow{X_{rand}} - \vec{X}| \quad (8)$$

Where $\vec{A}$ and $\vec{C}$ are calculated based on Equations (2) and (4), respectively. In addition, $\vec{X}(t)$ is the position vector and $\overrightarrow{X_{rand}}$ is a position vector of a whale randomly selected from the WOA population.

Based on the above information, the execution process of WOA can be diagramed in Figure 1. The WOA has been widely applied to resolve optimization problems in various fields, e.g. [16-18].

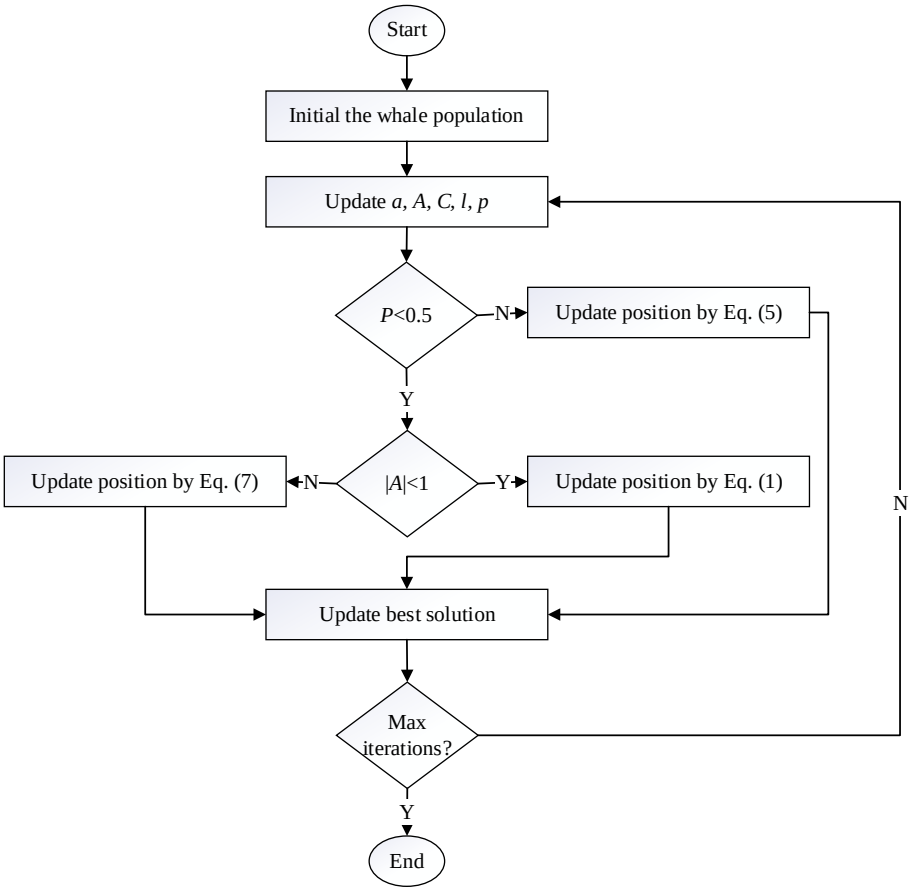

Figure 1: The execution process of WOA

Genetic Programming

[19] presented the genetic algorithm (GA) to resolve an optimization problem by imitating the evolution progress regarding organisms in the natural world, inspired by the famous theory of natural selection and evolution addressed by Darwin. In GA, a chromosome that consists of a series of genes is used to mimic the chromosome of an organism to represent a feasible solution, also called an individual, in an optimization problem. Therefore, all individuals form the solution population for an optimization problem. In addition, how well a feasible solution in the population can tackle an optimization problem is assessed by a well-designed fitness function based on the objective function of the optimization problem. The value of obtained evaluation result is called fitness. Next, a matching pool is constructed through using an appropriate mechanism of natural selection and matching to simulate the marriage process of these individuals. Each pair' individuals, called parents, in the matching pool can use the well-defined crossover functions, that closely correlate to the fitness of a feasible solution in the optimization problem, thus expectedly producing new individuals, called offspring, that have the superior quality. Furthermore, a mutation function is designed to simulate the unusual situation in crossover, i.e. extraordinary genetic changes. Each individual in the offspring is then assessed by the fitness function thus a new population of the next generation can be constructed by using the better individuals among the offspring to replace the worse (weak) individuals in the current generation, i.e. the parents' generation. [20] proposed the genetic programming (GP) to further extend GA into the field of computer programs. In GP, a tree-based structure is utilized to express a feasible solution, i.e. program, as shown in Figure 2, that can be easily decoded into an equation from left to right, as well as from bottom to top, as follows

$$6 + 10x - \frac{15}{\ln(z)} \tag{9}$$

There are two parts in the elements of a GP tree: (1) terminal set and (2) function set. First, the terminal set defines the available elements, that can be an independent variable, zero-argument function, or random constant,

etc. for each terminal branch in a GP tree. For example, the 6, 10, *x*, 15, and *z* in Figure 2 are the elements from the terminal set. The second one is the function set to define a set of primitive functions available for each branch in a GP tree, e.g. addition, square root, multiplication, sine and others. Hence, the +, -, ×, ÷, and ln in Figure 2 are the elements coming from the function set. By feeding the values of variables *x* and *y* into Equation (9), as well as referring the objective function in the optimization problem, the fitness value corresponding to the solution expressed by Equation (1) then can be obtained. Further, the crossover and mutation operators in GA must be re-designed to fit for the tree-based structure of a solution in GP. The original paired solutions illustrated in Figure 3 are

$$6 + 10x - \frac{15}{\ln(y)} \tag{10}$$

and

$$4 - \cos(x) + \log(y)\sqrt{z} \tag{11}$$

The new paired solutions then can be obtained by conducting the crossover operator as follows

$$6 + 10x - \frac{15}{\sqrt{z}} \tag{12}$$

and

$$4 - \cos(x) + \log(y)\ln(z) \tag{13}$$

Similarly, the mutation operator makes the original tree $6 + 10x - \frac{15}{\ln(y)}$ into a new solution $6 + 10x - \frac{15}{\sqrt{y}}$ as shown in Figure 4.

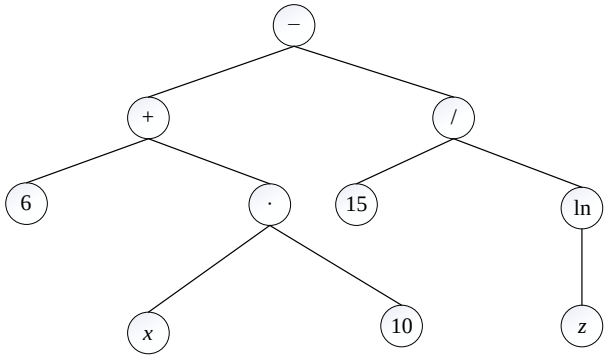

Figure 2: Tree-based structure in GP

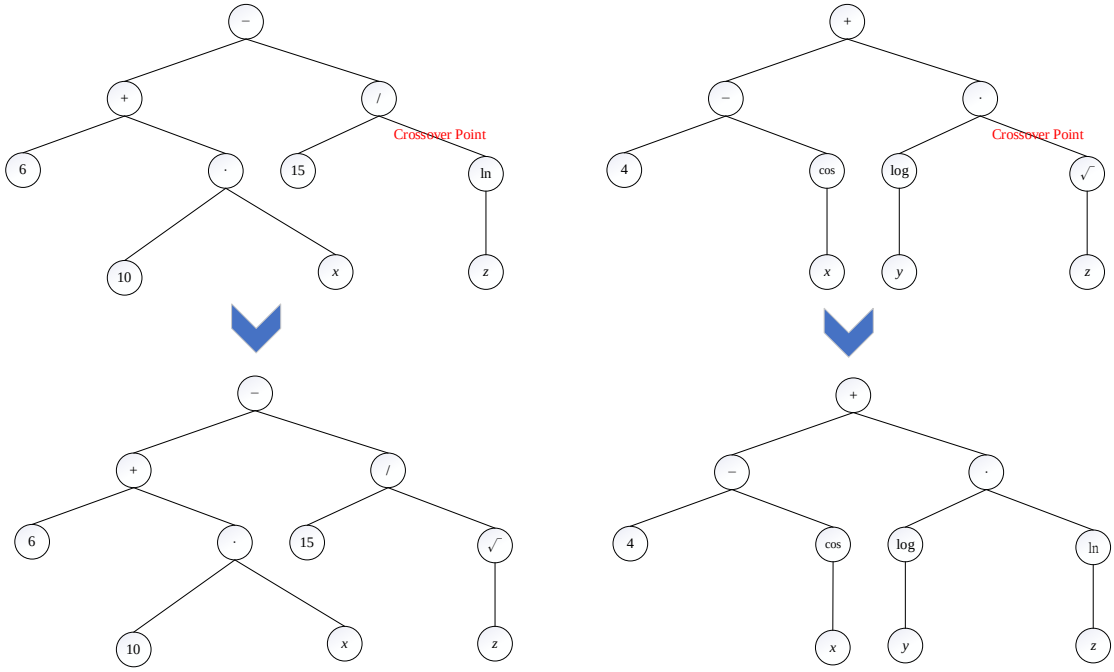

Figure 3: Crossover in GP

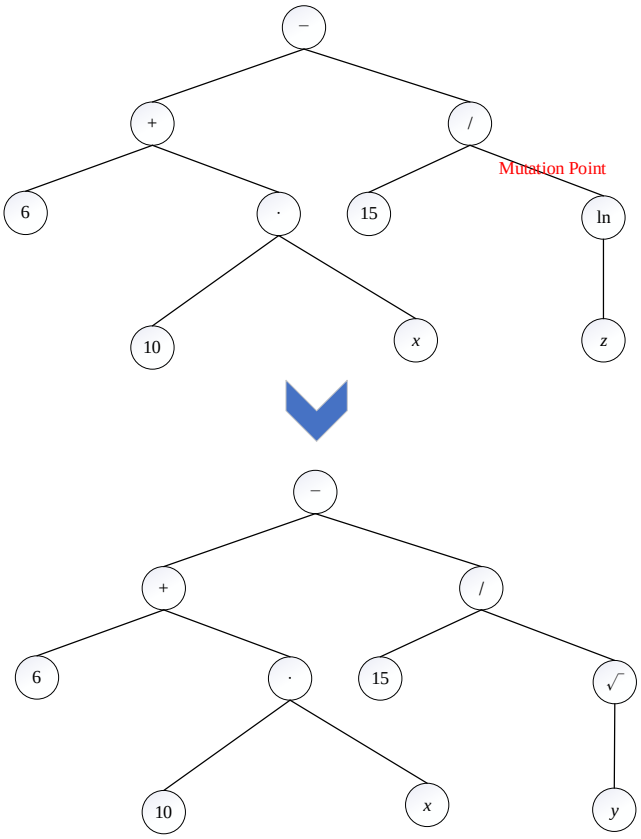

Figure 4: Mutation in GP

Figure 5 diagrams the basic procedure in a simple GP algorithm and is briefly summarized as follows [21-23]:

1. Initialization
   First, the required GP parameters including the population size, maximum size of programs, crossover rate, mutation rate etc., are determined. Subsequently, the initial solutions, i.e. programs or individuals, of a population with the pre-determined size are generated based on a pre-designed or random mechanism. Notably, these initial solutions can have different sizes, as well as different shapes.
2. Evaluation
   In this stage, each program in the population is first executed. Then, a pre-defined fitness function is applied to explicitly or implicitly measure how well the program can resolve the optimization problem. Generally speaking, there are several evaluation methods such as the amount of error between its output and target values, the total cost or time to make the system to be a stable state, or the classification accuracy, to yield the evaluation result, i.e. the fitness.
3. Creation
   Each individual is first selected based on the probability determined by its fitness value to form a matching pool with a pre-determined size. Then, the well-designed genetic operators are put on the selected individuals (programs) to produce the offspring population. These genetic operators include:
   (1) Reproduction: This operator creates new copied individual through duplicating the selected program.
   (2) Crossover: Randomly select two programs, called parents, to be a pair from the matching pool. The chosen paired programs are then recombined through the crossover mechanism with random crossover points thus forming two new programs, called children, in the offspring generation.
   (3) Mutation: This operator applies the mutate mechanism to the randomly selected programs from the children to produce new offspring individuals.

(4) Architectures altering: This stage can generate a new offspring program through altering the architecture of a selected program.

After the applications of the above genetic operators, the programs in the offspring population can replace the original individuals in the current population, i.e. current generation, based on a pre-proposed strategy, such as the elitist strategy. Therefore, the new population in the next generation then can be obtained.

4. Termination

When the terminating criteria can be satisfied, the best program, i.e. the program with the largest fitness value, ever encountered during the searching process of GP, will be designated as the final solution for the optimization problem. Otherwise, the GP algorithm will return to the "evaluation" stage and iteratively execute the subsequent steps until the termination criteria can be fulfilled.

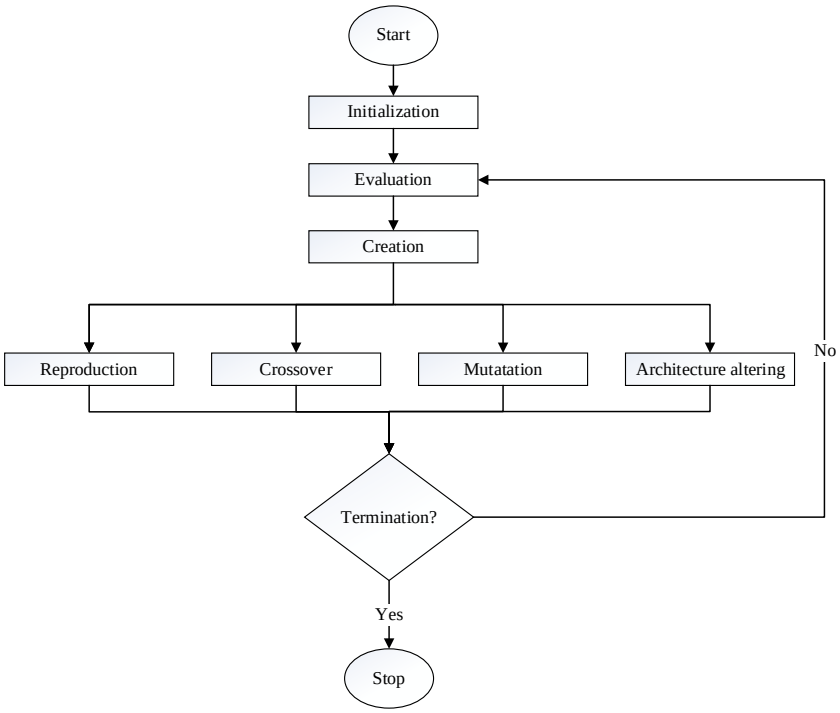

Figure 5: The basic procedure of a simple GP

Davies-Bouldin Index

To evaluate the performance regarding a clustering algorithm, [24] proposed a metric, called Davies-Bouldin index (DBI), to assess how well the clustering algorithm can divide the data into appropriate groups with the sufficient diversity. The DBI is an internal evaluation scheme through validating how well the obtained clustering results can made based on the quantities and features inherent to the data. Notably, a good DBI value does not always imply that it can retrieve the best information from the original data. Given some data points of n dimensions, whose feature vector is represented by X_j. Let C_i be the ith cluster for these data points, the scatter within the cluster i then can be measured by S_i defined as

$$S_i = \left(\frac{1}{T_i}\sum_{j=1}^{T_i}\left|X_j - A_i\right|^p\right)^{1/p} \quad (14)$$

where A_i and T_i are the centroid of C_i and the size of cluster i, respectively, and p is a parameter that determines the distance metrics.

The parameter p is usually set to 2 for measuring the Euclidean distance between the data point and centroid of the cluster. Notably, various distance metrics can be used in the situations where Euclidean distance may not be the optimal measurement to determine the cluster for the data of a higher dimension. Furthermore, the applied distance metric and the metric that is used in the clustering scheme must be identical. Next, the separation between clusters i and j can be measured by

$$M_{i,j} = \left\|A_i - A_j\right\|_p = \left(\sum_{k=1}^{n}\left|a_{k,i} - a_{k,j}\right|^p\right)^{1/p} \quad (15)$$

where $a_{k,i}$ and $a_{k,j}$ are the kth element of A_i and A_j, respectively. Specifically, $M_{i,j}$ is essentially the Euclidean distance measuring the centers of clusters i and j when p is set for 2.

To evaluate the quality of a clustering scheme, Davies and Bouldin defines an index as follows [24]

$$DBI \equiv \frac{1}{N}\sum_{i=1}^{N} D_i \quad (16)$$

where N is the total number of clusters, and D_i is calculated by

$$D_i \equiv \max_{j \neq i} R_{i,j} \quad (17)$$

and

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \quad (18)$$

Notably, the DBI depends on both of the data and algorithm. Next, a smaller DBI represents the clustering result is superior. In addition, a smaller $R_{i,j}$ indicates that the distances for the data points within clusters i and j are relatively smaller than the distance between the centroids between clusters i and j. Therefore, Equation (17) takes the maximum value of D_i to consider the worst case of clustering results for any group to a certain one.

3. Proposed Approach

In this study, the whale optimization algorithm (WOA) and genetic programming (GP) are utilized to propose a systematic approach for resolving the forecasting problems regarding stock prices as briefly depicted in Figure 6 and illustrated as follows:

Step 1: Data Collection

First, a certain number of corporations are randomly selected from a stock market to be the research objects. Some financial indexes, that are generally used to evaluate the performance regarding the business operation of a corporation, are then collected according to the published financial databank. Since theses financial indexes have their own physical meaning with diverse scales, each collected financial index must be identically normalized into [-1, 1] based on its corresponding maximum and minimum values, thus avoiding that the index with the wider measurement range will dominate the effects of other indexes which have relatively the narrower evaluation ambits.

Step 2: Data Clustering

The WOA is used to cluster the normalized financial data obtained in Step 1 into groups with an appropriate total number. Notably, each feasible solution in WOA is formed by all of the clustering centers, and the initial solutions with a pre-determined population size are randomly generated. Which cluster that a corporation belongs to is determined by the minimum value among the distances from the normalized financial indexes of this corporation to all clustering centers. Furthermore, the clustering performance for each feasible solution in WOA is evaluated by the Davies-Bouldin index (DBI) to produce its fitness. Therefore, a smaller DBI value implies a better clustering result, i.e. the distance of data points within the same cluster is relatively small, as well as the distance between different clustering centers is large enough. Hence, the corporations selected in Step 1 can be divided into several groups according to the performance of business operation for a corporation.

Step 3: Technical Indicators Calculation

For the corporations selected in Step 1, the daily data of stock trading, including the opening price, highest price, lowest price, closing price, and trade volume, are first collected through the stock exchange corporation or off-the-shelf financial database. The technical indicators are then calculated based on these collected stock trading data. Notably, it is hard to definitely determine which technical indicators are beneficial to forecast stock prices. Hence, the appropriate technical indicators can be selected according to the previous related researches.

Step 4: GP Models Construction

For each cluster i obtained in Step 2, the technical indicators along with the stock prices in the next trading day for each corporation in this cluster are first merged to form the original modelling data. Similarly, the technical indicators are normalized into a range between -1 and 1 based on their corresponding maximum and minimum values to avoid that the indicator with a larger range will dominate the effects of others with smaller ranges. The normalized data are then divided into two parts: (1) training data and (2) test data, according to a ratio pre-

determined by the users, e.g. 4:1. The GP tool is then applied to construct the forecasting model for stock prices in the next day based on the training data. The GP procedure is implemented several times since each execution result of GP, i.e. the candidate GP model, may differ. The candidate GP model with the best forecasting performance, e.g. the minimum root-mean-square error (RMSE), the maximum R squared (R^2), or the minimum mean absolute percentage error (MAPE) is selected as the final GP forecasting model for this cluster, named GP_i. Hence, the total number of the final GP forecasting models is identical with the total number of groups decided in Step 2. Notably, the GP technique is also implemented to the whole normalized technical indicators' data calculated based on all corporations selected in Step 1 to form a single GP forecasting model, named GP_{All}.

Step 5: Performance Evaluation

The acquired GP forecasting model for each cluster *i*, i.e. GP_i, in Step 4 is combined into an integrated forecasting model for the overall data of all clusters, named GP_{Int}. Therefore, the forecasting performance of applying the GP_{Int} to the normalized training data for all corporations can be obtained. In addition, the all normalized training data are also fed into the single GP forecasting model GP_{All} obtained in Step 4 and the forecasting performance is assessed, too. Next, the forecasting model for each cluster *i*, the integrated forecasting model GP_{Int}, and the single forecasting model GP_{All}, are applied to the test data set up in Step 4 for assessing the forecasting ability of GP models to the data that are never encountered before by using the forecasting measures including the MSE, MAPE and R^2.

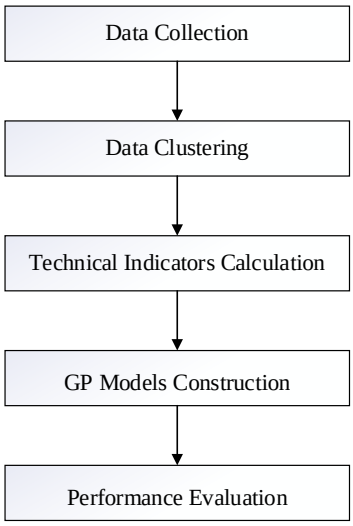

Figure 6: Propose forecasting approach

4. Case Study

In this section, the usefulness, effectiveness as well as efficiency of the proposed forecasting approach are demonstrated by presenting a case study on forecasting the closing price in the next trading day for the stocks listed in the Taiwan Stock Exchange Corporation (TWSE).

Data Collection

In this study, twenty (20) corporations that are randomly selected from the semiconductor sub-section in Taiwan stock market are considered. The research period ranges from 1 Jan. 2021 to 30 Apr. 2022. First, the financial reports announced at the end of the fourth quarter in 2021 were gathered from the TEJ (Taiwan Economic Journal)[1] database for the selected 20 corporations. Among the financial indexes, seven important indicators, including the earnings per share (EPS), return on assets (ROA), return on equity (ROE), gross profit margin (GPM), operating profit margin (OPM), debt ratio (DR), and price-earnings ratio (P/E) are chosen to evaluate

[1] https://www.tej.com.tw/

the performance regarding the business operation of a corporation. In addition, theses financial indexes are identically normalized into [-1, 1] according to its corresponding maximum and minimum values.

Data Clustering

Set the population size and constant for defining the shape of the logarithmic spiral b as 20 and 2, respectively, the WOA is used to cluster the normalized financial data obtained in Step 1. The clustering performance for each feasible solution in WOA is evaluated by the Davies-Bouldin index (DBI) with a parameter determining the distance metrics $p = 2$. The clustering performance of clustering results obtained by setting different total number of clusters as summarized in Table 1. Hence, the optimal total number of clusters thus can be determined as 3 since its minimum DBI value. The clustering results of corporations with a total number of clusters is 3 are shown in Table 2.

Table 1: Clustering results of WOA

Total number of clusters	DBI
2	1.0409
3	0.6698
4	0.7395
5	0.6723
6	3.9583

Table 2: Clustering results of corporation (total number of clusters is 3)

Cluster 1 Stock Code	Cluster 2 Stock Code	Cluster 3 Stock Code
6756	6415	2454
6239		3592
3014		3034
3443		8016
3530		
2481		
2338		
4968		
3583		
3257		
3450		
2369		
2329		
2401		
8028		

Technical Indicators Calculation

The daily data of stock trading, including the opening price, highest price, lowest price, closing price, and trade volume, for the selected corporations in this study are first collected from the TEJ database during 1 Jan. 2021 to 30 Apr. 2022. The technical indicators are then calculated based on these collected stock trading data. Notably, there are 16 technical indicators including (1) the 10-day moving average, (2) 20-day bias, (3) moving average convergence/divergence, (4) 9-day stochastic indicator K, (5) 9-day stochastic indicator D, (6) 9-day Williams overbought/oversold index, (7) 10-day rate of change, (8) 5-day relative strength index, (9) 24-day commodity channel index, (10) 26-day volume ratio, (11) 13-day psychological line, (12) 14-day plus directional indicator, (13) 14-day minus directional indicator, (14) 26-day buying/selling momentum indicator, (15) 26-day buying/selling willingness indicator, and (16) 10-day momentum, are considered in this study according to the previous researches in [25-32].

GP Models Construction

For the first cluster obtained in the previous step, the technical indicators along with the stock prices in the next trading day for each corporation in the first cluster are combined to be the original modelling data. Next, the technical indicators are normalized into [-1, 1] based on their corresponding maximum and minimum values. A ratio of 4:1 is used to separate the normalized data into training data (2,940 rows) and test data (735 rows). The GP tool is then applied to construct the forecasting model for stock prices in the next day based on the training data. The GP algorithm is executed for 5 times as shown in Table 3 where an asterisk indicates the final selected best GP forecasting model for the first cluster, named GP_1 based on the forecasting performance evaluated by maximizing the R squared (R^2). The same procedure is implemented for cluster 2 and 3 and the GP results are also shown in Table 3. In addition, the whole normalized technical indicators' data for all corporations selected in this study are also fed into GP to execute for 5 times to construct a single GP forecasting model, named GP_{All}, that is the run indicated by an asterisk in Table 3.

Table 3: GP implementation results

Cluster 1		Cluster 2		Cluster 3		All	
Run No.	**R^2**	**Run No.**	**R^2**	**Run No.**	**R^2**	**Run No.**	**R^2**
1	0.9972	1	0.9515	1	0.99670*	1	0.9971
2	0.9970	2	0.9597*	2	0.99648	2	0.9974*
3	0.9970	3	0.9581	3	0.99631	3	0.9973
4	0.9971	4	0.9546	4	0.99660	4	0.9972
5	0.9972*	5	0.9590	5	0.99619	5	0.9972

Performance Evaluation

The acquired final GP forecasting models in Table 3, i.e. GP_1, GP_2, and GP_3, for three clusters are merged into an integrated forecasting model for the overall data of all clusters, named GP_{Int}. Next, the single GP forecasting model GP_{All} obtained based on Table 3 is executed on the all normalized training data. The forecasting performance is assessed and summarized in Table 4. Furthermore, the forecasting models GP_{Int} and GP_{All} are then applied to the test data set up previously to assessing their forecasting ability to the data that are never seen before. The evaluation results by using the forecasting measures MSE, MAPE and R^2 are illustrated in Table 4. Based on Table 4, the integrated GP model GP_{Int} obtained by combining the GP forecasting models GP_1, GP_2, and GP_3 that are constructed individually based on the data in each cluster can outperform the single GP forecasting model GP_{All} that is the established for the whole normalized technical indicators' data for all corporations selected in this study. Therefore, we can conclude that it is an effective way to first construct a GP forecasting model for the stock trading data in each cluster individually, and combined these individual GP models for all clusters into an integrated one to forecast the stock closing prices in the future. The forecasting performance is superior than that acquired by the one single model directly constructed for the whole stock trading data of all selected corporations. In other words, the clustering mechanism can be proven an effective method for dividing the corporations into appropriate groups such that the corporations within the same group can have business operation of a similar style, thus reducing the difficulty of modelling, as well as improving the forecasting accuracy.

Table 4: Comparison of forecasting performance

Model	Training Data			Test Data		
	R^2	**MAPE**	**RMSE**	**R^2**	**MAPE**	**RMSE**
GP_{Int}	0.9987	0.0277	30.5880	0.9987	0.0295	30.9839
GP_{All}	0.9974	0.0303	43.2582	0.9981	0.0297	38.0627

5. Conclusions

There are various kinds of forecasting problems in our living world, e.g. science, engineering, culture, finance, politics, economics, etc. Forecasting the stock prices accurately is a traditional important work for an investor in the domain of finance. Sine the stock prices can be influenced by various factors, e.g. business trade cycle in the world, financial policy of a nation, variation of an exchange rate and inflation, as well as the business operation

regarding a corporation, etc., thus making the forecasting of stock prices very complex and attract much attentions from both researchers and practitioners. This study applies the whale optimization algorithm (WOA) and genetic programming (GP) **to** propose an approach to deal with the forecasting problems of stock prices. The WOA aims to group the stocks data into appropriate groups according to the fundamental business operations of corporations such that the business operations have the similar style for these corporations clustered in the same group. By doing so, the forecasting model can be constructed for each cluster individually thus increase the efficiency and effectiveness. The modelling procedure of forecasting is done by using the GP technique. A real case study of forecasting the closing prices of stocks listed in the Taiwan Stock Exchange Corporation (TWSE) is used to illustrate the usefulness, effectiveness, and efficiency of our proposed forecasting model. According to the experimental results and comparison, the integrated GP model by merging the individual GP forecasting models can outperform the single GP forecasting model established for the whole ungrouped data. It can prove that the clustering mechanism indeed can be an effective method to group the corporations into appropriate clusters such that the business operations of corporations within the same group are similar, thus decreasing the difficulty in constructing forecasting models, and enhancing the forecasting efficiency and accuracy.

Acknowledgments

The author would like to thank the partial support of Minghsin University of Science and Technology, Taiwan, R.O.C. under Contract No. MUST-111BA-01.

References

[1]. Kim, T., & Kim, H. Y. (2019). Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data. *PLOS ONE*, *14*(2), Article ID e0212320.

[2]. Rajab, S. & Sharma, V. (2019). An interpretable neuro-fuzzy approach to stock price forecasting. *Soft Computing*, *23*(3), 921-936.

[3]. Xiao, C. L., Xia, W. L., & Jiang, J. J. (2020). Stock price forecast based on combined model of ARI-MA-LS-SVM. *Neural Computing & Applications*, *32*(10), 5379-5388.

[4]. Shu, W. W., & Gao, Q. (2020). Forecasting Stock Price Based on Frequency Components by EMD and Neural Networks. *IEEE Access*, *8*, 206388-206395.

[5]. Lu, W. J., Li, J. Z., Li, Y. F., Sun, A. J., & Wang, J. Y. (2020). A CNN-LSTM-Based Model to Forecast Stock Prices. *Complexity*, *2020*, Article ID 6622927.

[6]. Yu, Z. X., Qin, L., Chen, Y. J., & Parmar, M. D. (2020). Stock price forecasting based on LLE-BP neural network model. *Physica A-Statistical Mechanics and Its Applications*, *553*, Article ID 124197.

[7]. Tang, Z. P., Zhang, T. T., Wu, J. C., Du, X. X., & Chen, K. J. (2020). Multistep-Ahead Stock Price Forecasting Based on Secondary Decomposition Technique and Extreme Learning Machine Optimized by the Differential Evolution Algorithm. *Mathematical Problems in Engineering*, *2020*, Article ID 5604915.

[8]. Lin, Y., Yan, Y., Xu, J. L., Liao, Y., & Ma, F. (2021). Forecasting stock index price using the CEEMDAN-LSTM model. *North American Journal of Economics and Finance*, *57*, Article ID 101421.

[9]. Farid, S., Tashfeen, R., Mohsan, T., & Burhan, A. (2021). Forecasting stock prices using a data mining method: Evidence from emerging market. *International Journal of Finance & Economics*, *2021*, Article ID 10.1002/ijfe.2516.

[10]. Ding, S., & Li, R. J. (2021). Forecasting the sales and stock of electric vehicles using a novel self-adaptive optimized grey model. *Engineering Applications of Artificial Intelligence*, *100*, Article ID 104148.

[11]. Tian, L. W., Li, F., Yu, S., and Guo, Y. K. (2021). Forecast of LSTM-XGBoost in Stock Price Based on Bayesian Optimization. *Intelligent Automation and Soft Computing*, *29*(3), 855-868.

[12]. Das, S., Sahu, T. P., Janghel, R. R., & Sahu, B. K. (2022). Effective forecasting of stock market price by using extreme learning machine optimized by PSO-based group-oriented crow search algorithm. *Neural Computing & Applications*, *34*(1), 555-591.
[13]. Guo, W. C., Li, Z. G., Gao, C., & Yang, Y. (2022). Stock price forecasting based on improved time convolution network. *Computational Intelligence*, *2022*, Article ID 10.1111/coin.12519.
[14]. Kirisci, M., & Yolcu, O. C. (2022). A New CNN-Based Model for Financial Time Series: TAIEX and FTSE Stocks Forecasting. *Neural Processing Letters*, *2022*, Article ID 10.1007/s11063-022-10767-z.
[15]. Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. *Advances in Engineering Software*, *95*, 51-67.
[16]. Lu, K. Z., & Ma, Z. M. (2021). A modified whale optimization algorithm for parameter estimation of software reliability growth models. *Journal of Algorithms & Computation Technology*, *15*, Article ID 17483026211034442.
[17]. Moazenzadeh, R., Mohammadi, B., Safari, M. J. S., Chau, K. W. (2022). Soil moisture estimation using novel bio-inspired soft computing approaches. *Engineering Applications of Computational Fluid Mechanics*, *16*(1), 826-840.
[18]. Ghobaei-Arani, M., & Shahidinejad, A. (2022). A cost-efficient IoT service placement approach using whale optimization algorithm in fog computing environment. *Expert Systems with Applications*, *200*, Article ID 117012.
[19]. Holland, J. H. (1975). *Adaptation in Nature and Artificial Systems*, Ann Arbor, MI: The University of Michigan Press.
[20]. Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, Cambridge, Mass: MIT Press.
[21]. Koza J. R., Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J., & Lanza, G. (2005). *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*, New York: Springer.
[22]. Ciglarič, I., & Kidrič, A. (2006). Computer-aided derivation of the optimal mathematical models to study gear-pair dynamic by using genetic programming. *Structural and Multidisciplinary Optimization*, *32*(2), 153-160.
[23]. Koza, J. R., Streeter, M. J., & Keane, M. A. (2008). Routine high-return human-competitive automated problem-solving by means of genetic programming. *Information Sciences*, *178*(23), 4434-4452.
[24]. Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224-227.
[25]. Kim, K.-J., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications*, *19*(2), 125-132.
[26]. Kim, K.-J., & Lee, W. B. (2004). Stock market prediction using artificial neural networks with optimal feature transformation. *Neural Computing & Applications*, *13*, 25-260.
[27]. Tsang, P. M., Kwok, P., Choy, S.O., Kwan, R., Ng, S. C., Mak, J., Tsang, J., Koong, K., & Wong, T.-L. (2007). Design and implementation of NN5 for Hong Kong stock price forecasting. *Engineering Applications of Artificial Intelligence*, *20*(4), 453-461.
[28]. Chang, P.-C., & Liu, C.-H. (2008). A TSK type fuzzy rule based system for stock price prediction. *Expert Systems with Application*, *34*(1), 135-144.
[29]. Ince, H. & Trafalis, T. B. (2008). Short term forecasting with support vector machines and application to stock price prediction. *International Journal of General Systems*, *37*(6), 677-687.
[30]. Huang, C.-L., & Tsai, C.-Y. (2009). A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. *Expert Systems with Application*, *36*(2), 1529-1539.
[31]. Lai, R. K., Fan, C.-Y., Huang, W.-H., & Chang, P.-C. (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Application*, *36*(2), 3761-3773.
[32]. Hsu, C.-M. (2013). A hybrid procedure with feature selection for resolving stock/futures price forecasting problems. *Neural Computing & Applications*, *22*(3-4), 651-671.