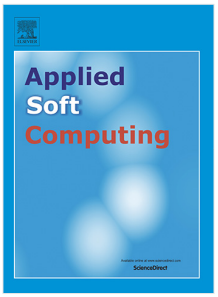
# Accepted Manuscript




A novel hybrid stock selection method with stock prediction

Fengmei Yang, Zhiwen Chen, Jingjing Li, Ling Tang

PII: S1568-4946(19)30146-2
DOI: https://doi.org/10.1016/j.asoc.2019.03.028
Reference: ASOC 5398

To appear in: *Applied Soft Computing Journal*

Received date : 15 August 2018
Revised date : 17 January 2019
Accepted date : 11 March 2019

Please cite this article as: F. Yang, Z. Chen, J. Li et al., A novel hybrid stock selection method with stock prediction, *Applied Soft Computing Journal* (2019), https://doi.org/10.1016/j.asoc.2019.03.028

# Highlights

- A novel stock selection method is proposed by introducing stock prediction.
- It includes two steps: stock prediction (for future markets) and stock scoring.
- Computational intelligences are used in the two steps of the novel model.
- Empirical study confirms its superiority over both traditional and similar models.

# A novel hybrid stock selection method with stock prediction

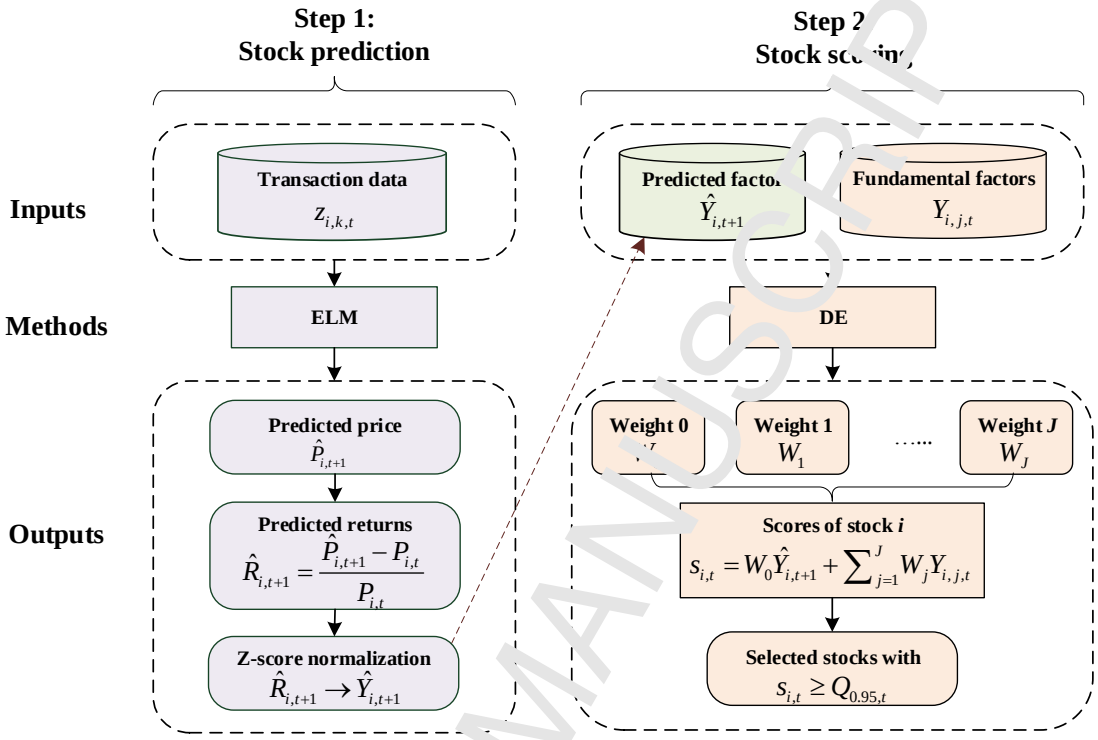



**Fig. 1. Overall framework of the proposed hybrid stock selection method with stock prediction**

# A novel hybrid stock selection method with stock prediction

**Fengmei Yang[a], Zhiwen Chen[a], Jingjing Li[b], Ling Tang[b,*]**

[a] School of Science, Beijing University of Chemical Technology, Beijing, 100029, China

[b] School of Economics and Management, Beihang University, Beijing 100191, China

**Abstract:** The success of stock selection is contingent upon the future performance of stock markets. We incorporate stock prediction into stock selection to specifically capture the future features of stock markets, thereby forming a novel hybrid (two-step) stock selection method involving stock prediction and stock scoring). (1) Stock returns for the next period are predicted using emerging computational intelligence (CI), i.e., extreme learning machine with a powerful learning capacity and a fast computing speed. (2) A stock scoring mechanism is developed as a linear combination of the predicted factor (generated in the first step) and the fundamental factors (popular in existing literature) based on CI-based optimization for weights, and top-ranked stocks are selected for an equally weighted portfolio. Using the A-share market of China as the study sample, the empirical results show that the novel hybrid approach, using highly weighted predicted factors, statistically outperforms both traditional methods (without stock prediction) and similar counterparts (with other model designs) in terms of market returns, which suggests the great contribution of stock prediction for improving stock selection.

***Keywords***: Stock selection; Stock prediction; Computational intelligence; Portfolio analysis

## 1. Introduction

Stock selection has long been identified as an important but challenging topic in the research area of financial market analysis [1, 2]. Like other similar research questions (e.g., stock prediction and stock portfolio recommendation), it is also in an attempt to facilitate the decision-making task for investment [3]. Nevertheless, it holds its own unique characteristic, in terms of a different specific task and hence different techniques. In particular, stock selection tends to distinguish 'good' stocks from 'bad' stocks, based on evaluation or scoring models [4, 5]; stock prediction reveals future information for stock prices and trends, based on forecasting models

*Corresponding author. School of Economics and Management, Beihang University, 37 Xueyuan Road, Haidian District, Beijing 100191, China. Email: lingtang@buaa.edu.cn (Ling Tang).

[6, 7, 8]; and stock portfolio recommendation makes optimal distributions of investment among stocks, based on optimization models [9, 10]. Accordingly, stock selection could serve as a preliminary step for stock portfolio recommendations to provide target stocks deserving study; and stock prediction could facilitate stock selection to capture the stock markets in the future [3, 11]. Therefore, this study especially focuses on stock selection, and introduce stock prediction for model improvements.

The core part of stock selection is the stock scoring mechanism for evaluating the value of a stock, which is the basis on which top-valued stocks (corresponding to high potential returns and low potential risk levels) are selected [12]. According to the existing literature, linear models, which describe the score of a stock as the sum of a set of weighted factors, are the predominantly used stock scoring models [13]. Accordingly, factor determination is an essential part of stock scoring; popular factors include yield factors, liquidity factors, risk factors, growth factors, momentum factors, etc., which effectively reflect the history and/or current features of a stock [2, 4, 12]. However, such information may lose efficacy in capturing the time-varying features of the associated stocks in the future, due to the extreme complexity and volatility of stock markets [14]. Furthermore, the success of stock selection is majorly contingent upon the future performance of stock markets; however, traditional stock selection models that consider only history and current information may have difficulty describing the future volatility of stock markets. Thus, the introduction of factors regarding future features of stock markets offers a new perspective for improving stock selection.

To explore the future features of stock markets, various forecasting algorithms have been employed, of which, computational intelligence (CI) (or artificial intelligence (AI)) has become increasingly dominant due to its powerful learning capability and high prediction accuracy. Typical CI techniques in stock market prediction (for stock prices, stock returns, market indexes, etc.) are artificial neural networks (ANNs) [15, 16, 17, 18] and support vector machines (SVMs) [19, 20, 21, 22]. The existing studies have investigated their effectiveness in providing valid information about future features of stock markets, which could in turn serve stock evaluation and selection [3, 23]. Among various CI learning paradigms, an emerging ANN (i.e., extreme learning machine (ELM), which effectively tackles the problems of time consumption and local optima in traditional CI models by using randomly fixed parameters) shows clear superiority in

financial market prediction in terms of much higher estimation accuracy and much faster computing speed [24, 25, 26]. Therefore, we implement the efficient CI tool in stock market prediction and further improve stock selection by incorporating predicted future features.

This paper proposes a novel hybrid stock selection method that incorporates stock prediction to effectively capture the future features of complex stock markets. The proposed method has two main steps: stock prediction and stock scoring. First, stock returns for the next period are predicted based on the emerging CI technique of ELM, which boasts fast computing speed and good generalization performance [27]. Second, the predicted factor and various fundamental factors (popular in existing models) are introduced into a typical linear stock scoring mechanism to evaluate the value of each candidate stock; highly valued stocks are selected to formulate an equally weighted portfolio. In the stock scoring mechanism, the weight terms of different factors are optimized using a typical CI programming technique, i.e., differential evolution (DE), which has a simple design but efficient performance [28]. Compared with existing stock selection models, our novel method makes the following contributions: (1) this method might be the first attempt to couple stock prediction with stock scoring, forming a novel stock selection method with the predicted factor regarding the future features of stock markets; (2) the proposed model is applied to the A-share market of China to verify its effectiveness and validity in stock selection and portfolio formulation, and typical stock selection models (without stock selection) and similar counterparts (with other forecasting models, factor designs, optimization algorithms or fitness functions) are also conducted for comparison.

The main aim of this paper is to propose a novel hybrid stock selection model with stock prediction for capturing the future features of stock markets and to verify its superiority over similar counterparts. The rest of this paper is organized as follows. Section 2 formulates the novel methodology in detail. Section 3 reports the empirical results and discusses the effectiveness of the proposed model. Section 4 concludes the paper and outlines potential directions for future research.

## 2. Literature review

Stock selection has become an increasingly hot issue in the field of finance research, as recent interesting studies listed in Table 1.

**Table 1** Recent studies on stock selection

| Authors | Methods | Factors | | Country |
|---|---|---|---|---|
| | | Fundamental factors | Technical factors | |
| Fu et al. (2018) [11] | Linear regression, deep neural network, and random forest | Return on equity (ROE), returns on assets (ROA), price-to-book (PB), etc. | Relative strength index (RSI), moving average (MA), volume (VOL), etc. | China |
| Zhang et al. (2018) [5] | SVM | Growth, and financial quality | Share price, volatility, turnover rate, etc. | China |
| Hajjami and Amin. (2018) [29] | Ordered weighted averaging | Earnings per share (EPS), return on equity (ROE), cash flow ratio (CF), etc. | Size ratio (SIZE) | Iran |
| Bhatia et al. (2018) [30] | Random forest, SVM, ANN, etc. | Book-to-price (B/P), sale-to-price (S/P), and return on equity (ROE). | Return rate, market value, and systematic risk | US |
| Nifty (2018) [31] | DE | Return on Assets (ROA) | Average true range (ATR), rate of change (ROC), moving average (MA), etc. | Indian |
| Liu et al. (2017) [32] | ANN | Earning-to-price (E/P), book-to-price (B/P), sales-to-price (S/P), etc. | Market capitalization, and momentum | US |
| Yu et al. (2016) [4] | DE, particle swarm optimization (PSO), genetic algorithm (GA), etc. | Current ratio (CR), inventory turnover rate (ITR), quick ratio (QR), etc. | Relative strength index (RSI), and raw stochastic value (RSV) | China |
| Yodmun et al. (2016) [33] | Fuzzy analysis | Price-to-earning (P/E), and price-to-book (P/B) | Price-to-intrinsic ($P/P_n$) | Indian |
| Peachavanish (2016) [34] | Cluster analysis | | Rate of Change (ROC), and exponential moving averages (EMA). | Thailand |
| Suzuki et al. (2016) [35] | Ensemble learning | | Spatial technical discriminant analysis (STDA), and spatial relative strength index (SRSI) | Japan and US |
| Amin et al. (2016) [36] | Ordered weighted averaging | Earnings per share (EPS), return on assets (ROA), return on equity (ROE), etc. | | Iran |

On one hand, diverse methods have been applied to stock selection, which fall into two main categories, i.e., statistical methods (e.g., linear regression [11], fuzzy analysis [33], cluster

analysis [34], and ordered weighted averaging [29, 36]) and CI techniques (e.g., ANN [11, 30, 32], SVM [5, 30], DE [4, 31], GA [4], PSO [4]). Of them, the CI techniques have become increasingly prevailing in the field of financial market analysis, due to their adaptive computer-assisted learning and powerful modelling performance [4]. Noticeably, a simple but effective CI algorithm, i.e., DE with fast convergence, strong robustness, simple operation and few parameters relative to other similar CI algorithms, have been shown as an effective tool for stock selection [4, 37]. Therefore, this study especially introduces such a useful CI technique in methodology formulation, and compare it with other popular CI techniques (e.g., PSO and GE).

On the other hand, various fundamental factors and technical factors have been employed to evaluate stocks. In particular, fundamental factors, such as ROE, current ratio CR and ROA, effectively reflect the nature of the business, finance, and competitiveness of a stock [5, 11, 30]. In comparison, technical factors, such as turnover rate, volume, and volatility, describe the market state of a stock [34, 35]. Accordingly, most studies considered both fundamental factors and technical factors in formulating stock selection models [29, 31, 32, 33]. However, these factors focus on the history and/or current features of stock markets, but fail in capturing the future features. Therefore, this paper makes the contribution to literature by incorporating predicted factors, in addition to these factors.

## 3. Methodology formulation

A novel hybrid stock selection method with stock prediction is formulated in this section. Section 3.1 presents the general framework of the novel model. Sections 3.2 and 3.3 elaborate on the two major steps, i.e., stock prediction and stock scoring, respectively. Sections 3.4 shows a toy example of the proposed stock selection model.

### 3.1 Overall framework

This study develops a novel stock selection model by introducing stock prediction to capture the future features of stock markets. Accordingly, the novel stock selection model has two main steps, i.e., stock prediction (for modeling future stock markets) and stock scoring (for evaluating

and selecting stocks), as shown in the overall framework illustrated in Fig. 1.

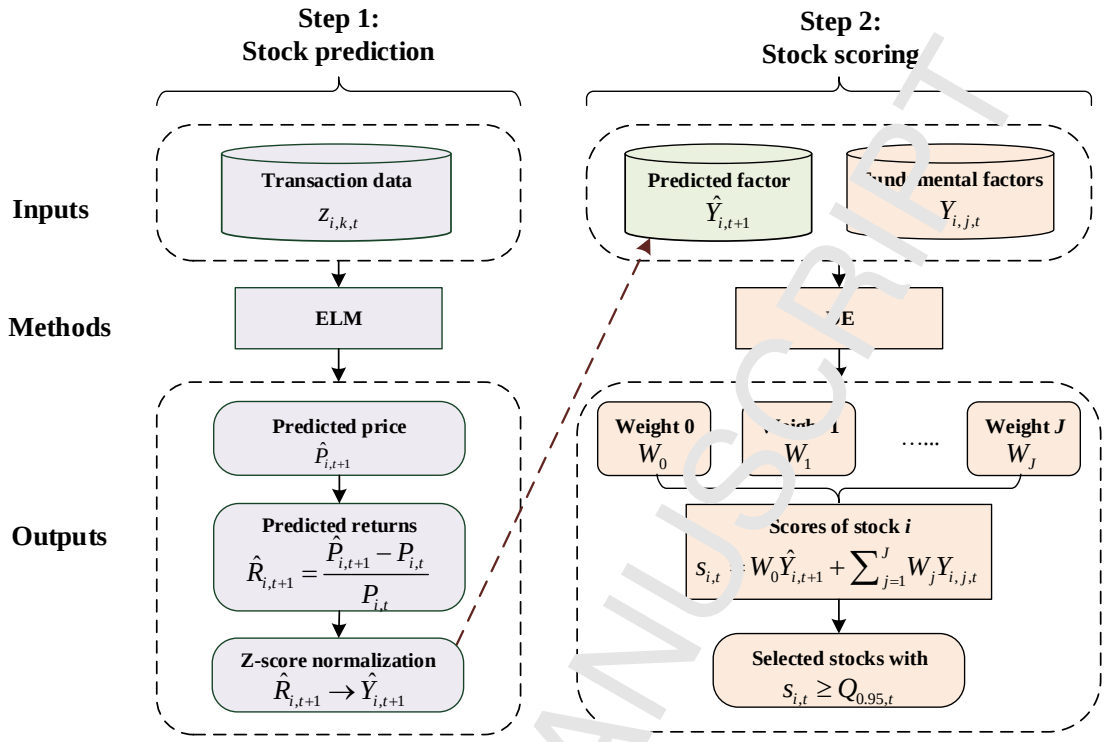



**Fig. 1** Overall framework of the proposed hybrid stock selection method with stock prediction

(1) ***Stock Prediction***: An efficient CI forecasting tool, ELM, is used to individually predict the prices of all candidate stocks for the next period, i.e., $P_{i,t+1}(i=1,\ldots,I)$, based on the associated transaction data $z_{i,k,t}(k=1,2,\ldots,16)$, where $P_{i,t}$ is the price of stock $i$ in period $t$ and $z_{i,k,t}$ is the $k$th element of transaction data. The corresponding predicted stock prices $\hat{P}_{i,t+1}$ for $P_{i,t+1}$ are converted into predicted stock returns $\hat{R}_{i,t+1} = (\hat{P}_{i,t+1} - P_{i,t})/P_{i,t}$ and normalized into dimensionless values $\hat{Y}_{i,t+1}$ (see Eq. (3)) as the output of the first step and an input of the second. To describe the complex stock market, various transaction data (price factors, liquidity factors, valuation factors, etc.) [5, 11] are introduced as the explanatory variables in the prediction model.

(2) ***Stock Scoring***: A stock scoring mechanism is formulated to evaluate all candidate stocks in a linear combination of the predicted factor $\hat{Y}_{i,t+1}$ and various fundamental factors $Y_{i,j,t}(j=1,\ldots,J)$ (see Table 2), i.e., $s_{i,t} = W_0\hat{Y}_{i,t+1} + \sum_{j=1}^{J} W_j Y_{i,j,t}$, where $s_{i,t}$ is the score of stock $i$ estimated in period $t$ and $W_j$ is the weight of the $j$th factor. A simple but efficient CI

optimization technique, DE, is implemented to determine the weights $W_0$ of the predicted factor $\hat{Y}_{i,t+1}$ and $W_j$($j$=1,…,$J$) on diverse fundamental factors $Y_{i,j,t}$. Finally, the top 5% of highly valued stocks (corresponding to the largest potentials of increases in future returns, i.e., the stocks with $s_{i,t} \geq Q_{0.95,t}$, where $Q_{0.95,t}$ is quantile 0.95 of all stock scores estimated in period $t$) are selected to formulate an equally weighted portfolio as the final output of the stock selection model.

The following two subsections respectively describe the above two steps (stock prediction and stock scoring) in detail together with the related techniques.

### 3.2 Stock prediction

The returns of all candidate stocks for the next period are predicted to reflect the future features of stock markets, thereby serving stock selection. In period $t$, the price of stock $i$ for the next period $t$+1 is estimated based on its associated transaction data $z_{i,k,t}$($k$=1,2,…,16) as follows:

$$\hat{P}_{i,t+1} = f\left(z_{i,1,t}, z_{i,2,t}, \ldots, z_{i,16,t}\right), \tag{1}$$

where $\hat{P}_{i,t}$ denotes the predicted price of stock $i$ for period $t$, and $z_{i,k,t}$ denotes the $k$th element of the transaction data. For explanatory variables, a total of 16 transaction data elements $z_{i,k,t}$($k$=1,…,16) of stock markets are introduced according to the previous studies [5, 11]: close, open, high, low, average price, market capitalization, return rate, volume, turnover, turnover rate, volatility, general capital, price-to-earnings (PE), price-to-book (PB), price-to-sales (PS) and price-to-cash flow (PCF). The predicted stock prices $\hat{P}_{i,t+1}$ are transformed into stock returns as follows:

$$\hat{R}_{i,t+1} = \frac{\hat{P}_{i,t+1} - P_{i,t}}{P_{i,t}}, \tag{2}$$

where $\hat{R}_{i,t}$ denotes the predicted returns of stock $i$ for period $t$. Then, standardize the predicted stock returns into dimensionless values via $Z$-score normalization [38] as the output of stock prediction or an important input of stock scoring (in Section 3.3) as follows:

$$\hat{Y}_{i,t+1} = \frac{\hat{R}_{i,t+1} - \overline{\hat{R}}_{t+1}}{\hat{D}_{t+1}}, \tag{3}$$

where $\overline{\hat{R}}_t = \sum_{i=1}^{N} \hat{R}_{i,t} / N$ is the average predicted returns cross all $N$ stocks in period $t$, and $\hat{D}_t = \sqrt{\sum_{i=1}^{N} \left( \hat{R}_{i,t} - \overline{\hat{R}}_t \right)^2 / N}$ is the standard deviation.

Regarding the forecasting technique $f(\bullet)$ in Eq. (1), an emerging CI technique, ELM with fast computing speed and good generalization performances, is introduced. Actually, ELM is a special case of single-hidden layer feedforward neural networks (SLFNs), which use randomly generated hidden neurons to avoid time-consuming iterative learning processes that usually result in local optima [24]. Given $N$ arbitrary distinct samples $(x_i, t_i)(i=1, ..., N)$, where $x_i \in R^n$ is the input and $t_i \in R^m$ is the output, a standard SLFN with $\tilde{N}$ hidden nodes can be represented as follows:

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i \cdot x_j + b_i) = o_j, \quad j = 1, 2, ..., N, \tag{4}$$

where $\beta_i$ denotes the output weight connecting with the $i$th hidden neuron, $g(\bullet)$ is the activation function, $w_i$ is the input weights connecting with the $i$th hidden neuron, $w_i \cdot x_j$ is the inner product of $w_i$ and $x_j$, $b_i$ is the bias of the $i$th hidden neuron, and $o_j$ is the output of the SLFN model for the $j$th simple. In this study, the number of hidden nodes $\tilde{N}$ is set using the trial-and-error method.

The goal of SLFNs is to minimize the errors of the outputs as follows:

$$\min_{\beta_i, w_i, x_i} \sum_{j=1}^{N} \left( \sum_{i=1}^{\tilde{N}} \beta_i g(w_i \cdot x_j + b_i) - t_j \right)^2. \tag{5}$$

To solve this problem, iteratively training algorithms based on gradients is a popular method used to determine the parameters $\beta_i$, $w_i$ and $b_i$. However, some tough problems may occur during such an iteratively learning process, e.g., parameter sensitivity, overfitting, high time consumption and local minima [39]. To address these problems, Huang et al. [40] innovatively introduced the interesting idea of using randomization to avoid tuning all parameters within the network. Suppose the error of the output is nearly zero as follows:

$$\sum_{j=1}^{\tilde{N}} \left\| o_j - t_j \right\| = 0, \tag{6}$$

Eq. (5) can be transformed into the following:

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i \cdot x_j + b_i) = t_j,\ j = 1,\ldots,N\ . \tag{7}$$

Let $H$ be the hidden layer output matrix of the neural network; Eq. (7) can be rewritten compactly as follows:

$$H\beta = T\ . \tag{8}$$

Randomly generate and fix the parameters $w_i$ and $b_i$ in the hidden neurons based on a given distribution (such as the Gaussian distribution), such that the optimization problem in Eq. (5) is equivalent to estimating a minimum norm least square solution for $\beta$ as follows:

$$\hat{\beta} = H^{\dagger}T\ , \tag{9}$$

where $H^{\dagger}$ is the Moore–Penrose generalized inverse of the matrix $H$, which could be computed, in practice, based on the generalized least squares [41]. Huang et al. [40] have proved that this solution of minimum norm least square is unique.

### 3.3 Stock scoring

By introducing the future factor $\hat{Y}_{i,t+1}$ predicted by the first step, a novel stock scoring mechanism is formulated. In addition, a set of fundamental factors $V_{i,j,t}$($j$=1,…,12) concerning stock profitability, leverage, liquidity, efficiency and growth that are popularly used in existing stock selection models [3, 12] are employed, as listed in Table 2. These fundamental factors $V_{i,j,t}$ are standardized into Z-score normalizations $Y_{i,j,t}$ for consistency [38] as follows:

$$Y_{i,j,t} = \frac{V_{i,j,t} - \overline{V}_{j,t}}{D_{j,t}}\ , \tag{10}$$

where $\overline{V}_{j,t} = \sum_{i=1}^{N} V_{i,j,t} / N$ is the average value of factor $j$ cross all $N$ stocks in period $t$, and $D_{j,t} = \sqrt{\sum_{i=1}^{N} \left(V_{i,j,t} - \overline{V}_{j,t}\right)^2 / N}$ is the standard deviation.

**Table 2** Factors used in the stock scoring mechanism

| Category | Abbr. | Descriptions | Ref. |
|---|---|---|---|
| **Profitability** | ROE | Return on equity (after tax) = net income after tax / shareholders' equity | [4] |
| | ROA | Return on asset (after tax) = net income after tax / total assets | [4] |
| | OPM | Operating profit margin = operating income / net sales | [12] |
| | NPM | Net profit margin = net income after tax / net sales | [12] |
| **Leverage** | DE | Debt-to-equity ratio = total liabilities / shareholders' equity | [12] |
| **Liquidity** | CF | Cash flow ratio = cash flow from operating activities/current liabilities | [12] |
| | CR | Current ratio = current assets / current liabilities | [4] |
| | QR | Quick ratio = quick assets / current liabilities | [4] |
| **Efficiency** | ITR | Inventory turnover rate = cost of goods sold / average inventory | [3] |
| | RTR | Receivables turnover rate = net credit sales / average accounts receivable | [3] |
| **Growth** | OIG | Operating income growth rate = (operating income at the current quarter – operating income at the previous quarter) / operating income at the previous quarter | [3] |
| | NIG | Net income growth rate = (net income after tax at the current quarter – net income after tax at the previous quarter / net income after tax at the previous quarter | [3] |
| **Predicted returns** | PR | Predicted returns = (predicted price at the next quarter – closing price at the current quarter) / closing price at the current quarter | This study |

The score $s_{i,t}$ of stock $i$ for period $t$ is described in a linear weighted combination of the predicted factor $\hat{Y}_{i,t+1}$ and fundamental factors $Y_{i,j,t}$:

$$s_{i,t} = W_0 \hat{Y}_{i,t+1} + \sum_{j=1}^{J} W_j Y_{i,j,t} , \tag{11}$$

where $W_j$ denotes the weight of the $j$th factor, balancing the importance of different factors in evaluating stocks, and $s_{i,t}$ is the score of stock $i$ estimated in period $t$, with a greater level corresponding to a higher potential for increases in future returns. Candidate stocks are ranked from highest to lowest and $r_{i,t} \in \{1,2,...,N\}$ denotes the ranking of stock $i$ for period $t$, i.e., $r_{i,t} \le r_{k,t}$ if $s_{i,t} \ge s_{k,t}$ where $i,k \in \{1,2,...,N\}$ represent any two separate stocks. Thus, a top-ranked stock is estimated to have a large potential of positive future returns, and an equally weighted portfolio is constructed based on the top 5% of ranked stocks, i.e., the stocks with $s_{i,t} \ge Q_{0.95,t}$, where $Q_{0.95,t}$ is quantile 0.95 of stock scores in period $t$. The performance of the formulated portfolio for the next period $t$+1 is evaluated in terms of the average returns of all selected stocks as follows:

$$\overline{R}_{t+1} = \frac{1}{m}\sum_{r_{i,t}=1}^{m} R_{t+1}\left(r_{i,t}\right), r_{i,t} = 1, 2, ..., m, \tag{12}$$

where $m$ is the total number of selected stocks, $R_{t+1}(r_{i,t})$ is the next-period returns of the stock with current ranking $r_{i,t}$ evaluated in period $t$, and $\overline{R}_{t+1}$ is the next-period returns of the portfolio constructed by the proposed model.

To effectively capture the relationships between factors, in terms of weights $w_j(j=0,1,\ldots,J)$, the *Spread* function is used as the fitness [2] as follows:

$$\begin{aligned} &\min_{w_j} F = -\frac{1}{T}\sum_{t=1}^{T} Spread_t \\ &Spread_t = \frac{\sum_{r_{i,t=1}}^{m} R_{t+1}\left(r_{i,t}\right) - \sum_{r_{i,t=1}}^{N} R_{t+1}\left(r_{i,t}\right)}{m} \end{aligned}, \tag{13}$$

where $T$ is the size of the training periods.

To solve Eq. (13), a simple but efficient CI optimization tool, i.e., the DE algorithm proposed by Price and Stone in 1997 [42], is employed. The reason of selecting DE as the optimization algorithm is that DE has been shown to have fast convergence, strong robustness, simple operation and few parameters, relative to other similar CI algorithms [4, 37]. As a population-based heuristic algorithm, DE searches for optimal solutions iteratively in four main stages, i.e., initialization, mutation, crossover and selection, until achieving the stop criteria [4, 28, 37].

(1) ***Initialization***: For a $D$-dimension optimization problem, DE randomly generates a population of $P$ initial feasible solutions in terms of chromosomes $X_{p,g=0} \in R^D$ as follows:

$$X_{p,g=0} = X_{LB} + Z \cdot \left(X_{UB} - X_{LB}\right), p = 1, 2, ..., P, \tag{14}$$

where $g \in \{0,1,\ldots,G\}$ is the iteration, $X_{LB}$ and $X_{UB} \in R^D$ are the lower and upper boundaries of the feasible domain, respectively, and $Z \in R^D$ is a random vector with random elements following a uniform distribution between 0 and 1.

(2) ***Mutation***: A simple but efficient mutation is employed in this study to create the donor vector $V_{p,g} \in R^D$ for individual $p$ as follows:

$$V_{p,g} = X_{r_1^p,g} + \alpha \cdot \left(X_{r_2^p,g} - X_{r_3^p,g}\right), \tag{15}$$

where $X_{r_1^p,g}$, $X_{r_2^p,g}$ and $X_{r_3^p,g} \in R^D$ are randomly sampled vectors from the current population excluding the individual $p$ at generation $g$, the random indices $r_1^p$, $r_2^p$ and $r_3^p \in \{1,2,\ldots,p-1,p+1,\ldots,P\}$ are mutually exclusive integers, and $a \in (0,1)$ is the scaled coefficient.

(3) ***Crossover***: The trial vector $U_{p,g}=\{u_{p,d,g}\}$ is produced based on the target vector $X_{p,g}=\{x_{p,d,g}\}$ and its donor vector $V_{p,g}=\{v_{p,d,g}\}$ to enhance the diversity of the population as follows:

$$u_{p,d,g} = \begin{cases} v_{p,d,g} & \text{if } r_{p,g} \le C_r \text{ or } r_d = a \\ x_{p,d,g} & \text{otherwise} \end{cases}, \tag{16}$$

where $r_{p,d} \in (0,1)$ is a random term, $r_d \in \{1,2,...,D\}$ is a random index, and $C_r \in (0,1)$ is the crossover rate.

(4) ***Selection***: Whether the target vector $X_{p,g}$ or its trail vector $U_{p,g}$ survive as the solution $X_{p,g+1}$ in the next generation for the individual $p$ is determined according to the fitness function as follows:

$$X_{p,g+1} = \begin{cases} U_{p,g} & \text{if } f\left(U_{p,g}\right) \le f\left(X_{p,g}\right) \\ X_{p,g} & \text{otherwise} \end{cases}, \tag{17}$$

where $f(\bullet)$ is the fitness function to be minimized (as defined in Eq. (13)). Stop if the fitness value stays the same for 1000 generations; otherwise, go to Step (2) for the next iteration.

### 3.4 A toy example

To illustrate the proposed stock selection model with stock prediction, a toy example is conducted. In the toy example, two stocks of China's A-share stock market are considered, i.e., 000002.SZ and 600000.SH (marked as $i$=S1 and S2, respectively). For simplicity, only two transaction factors, i.e., close and PE ($k$=F1 and F2, respectively) are employed in stock prediction while one fundamental factor, i.e., QR (with the $Z$-score normalized value marked as $Y$) is used in stock scoring. For the period of Q1 2015 ($t$=0), the corresponding transaction data $z$, stock price $P$ and the Z-score normalized fundamental factor $Y$ are as follows:

$$z_{i,k,t} = [z_{\mathrm{S1,F1,0}} \quad z_{\mathrm{S1,F2,0}} \quad z_{\mathrm{S2,F1,0}} \quad z_{\mathrm{S2,F2,0}}] = [12.83 \quad 9.69 \quad 12.33 \quad 13.56], \tag{18}$$

$$P_{i,t} = [P_{\mathrm{S1,0}} \quad P_{\mathrm{S2,0}}] = [12.83 \quad 12.33], \tag{19}$$

$$Y_{i,t} = \begin{bmatrix} Y_{\mathrm{S1},0} & Y_{\mathrm{S2},0} \end{bmatrix} = \begin{bmatrix} -1 & 1 \end{bmatrix}. \tag{20}$$

Based on the general framework (Fig. 1), a better stock can be selected from the two candidate stocks, with the detailed process illustrated in Fig. 2.

First, introduce the transaction data $z$ (i.e., Eq. (18)) as the inputs to the ELM, the predicted prices of the two stocks for the next period $t$=1 can be obtained:

$$\hat{P}_{i,t+1} = \begin{bmatrix} \hat{P}_{\mathrm{S1},1} & \hat{P}_{\mathrm{S2},1} \end{bmatrix} = \begin{bmatrix} 15.27 & 10.48 \end{bmatrix}. \tag{21}$$

The predicted returns of two stocks can be obtained as:

$$\hat{R}_{i,t+1} = \begin{bmatrix} \hat{R}_{\mathrm{S1},1} & \hat{R}_{\mathrm{S2},1} \end{bmatrix} = \begin{bmatrix} \dfrac{\hat{P}_{\mathrm{S1},1} - P_{\mathrm{S1},0}}{P_{\mathrm{S1},0}} & \dfrac{\hat{P}_{\mathrm{S2},1} - P_{\mathrm{S2},0}}{P_{\mathrm{S2},0}} \end{bmatrix} = \begin{bmatrix} 0.19 & -0.15 \end{bmatrix}. \tag{22}$$

Standardized the predicted returns in a form of $Z$-score normalization to get the dimensionless values:

$$\begin{aligned} \hat{Y}_{i,t+1} &= \begin{bmatrix} \hat{Y}_{\mathrm{S1},1} & \hat{Y}_{\mathrm{S2},1} \end{bmatrix} = \frac{\begin{bmatrix} \hat{R}_{\mathrm{S1},1} & \hat{R}_{\mathrm{S2},1} \end{bmatrix} - \bar{\hat{R}}_1}{\hat{D}_1} \\ &= \frac{\begin{bmatrix} \hat{R}_{\mathrm{S1},1} & \hat{R}_{\mathrm{S2},1} \end{bmatrix} - (\hat{R}_{\mathrm{S1},1} + \hat{R}_{\mathrm{S2},1})/2}{\sqrt{\left(\left(\hat{R}_{\mathrm{S1},1} - \bar{\hat{R}}_1\right)^2 + \left(\hat{R}_{\mathrm{S2},1} - \bar{\hat{R}}_1\right)^2\right)/2}}, \\ &= \frac{[0.19, -0.15] - 0.02}{0.17} = \begin{bmatrix} 1 & -1 \end{bmatrix} \end{aligned} \tag{23}$$

Given the weight terms $W_0$=0.6 and $W_1$=0.4 on the two predicted and history information, respectively, the values of the two stocks can be finally calculated:

$$\begin{aligned} s_{i,t} &= \begin{bmatrix} W_0\hat{Y}_{\mathrm{S1},1} + W_1 Y_{\mathrm{S1},1} & W_0\hat{Y}_{\mathrm{S2},1} + W_1 Y_{\mathrm{S2},1} \end{bmatrix} \\ &= [0.6\times 1 + 0.4\times(-1) \quad 0.6\times(-1) + 0.4\times(1)] = \begin{bmatrix} 0.2 & -0.2 \end{bmatrix} \end{aligned}. \tag{24}$$

Therefore, the stock S2, (i.e., 000002.SZ) with a higher score is selected from the two candidate stocks.

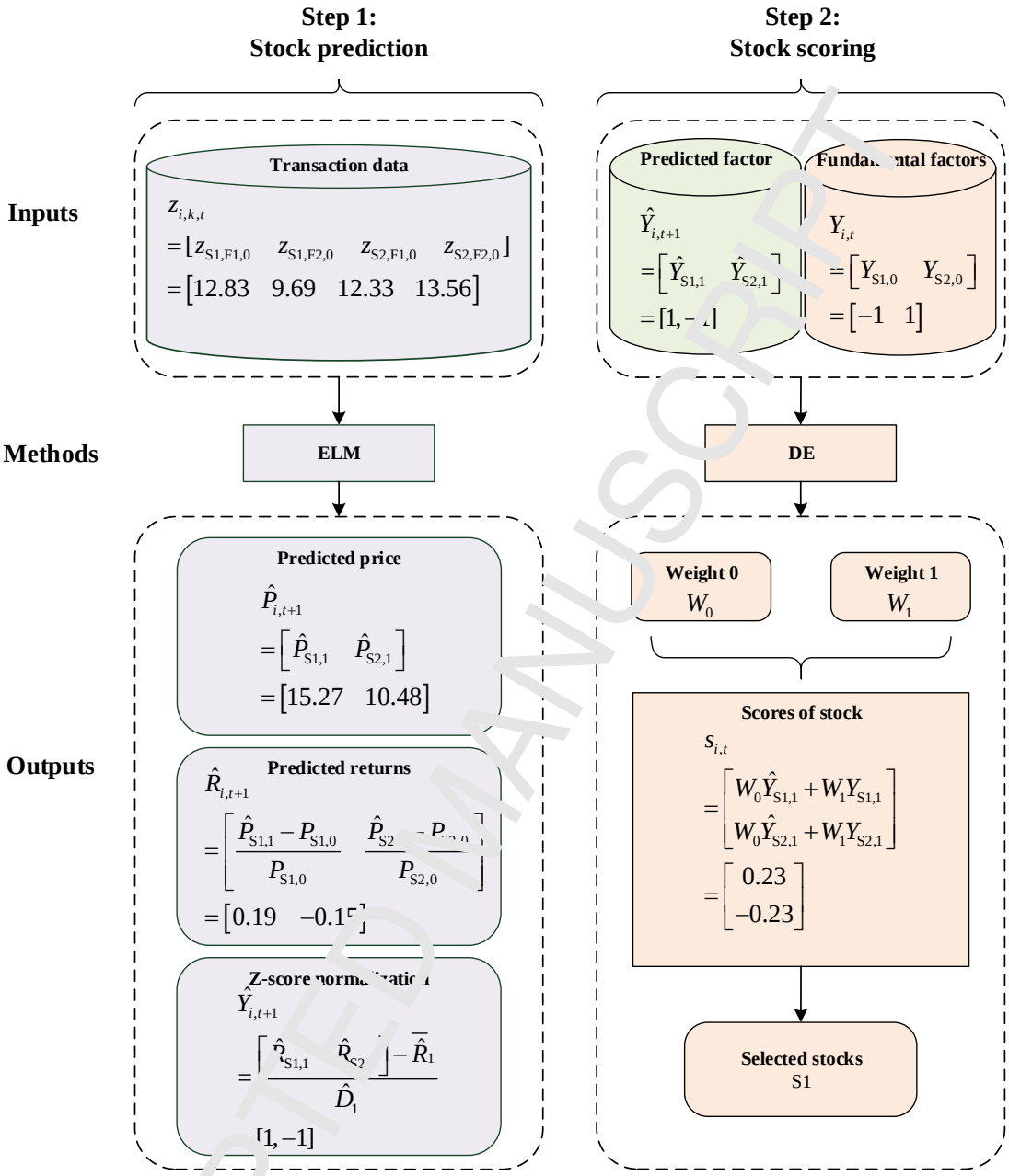



**Fig. 2** A toy example for the proposed stock selection model with stock prediction

## 4. Empirical study

For illustration and verification, the novel hybrid stock selection method is applied for portfolio construction in the A-share market of China. Section 4.1 designs the empirical study, Section 4.2 investigates the effectiveness of the proposed model relative to average market performance and A-share index, Section 4.3 discusses the superiority of the proposed model over various typical stock selection models, and Section 4.4 summarizes the empirical study.

## 4.1 Experimental design

This section presents the experimental design in terms of data descriptions, benchmark models and evaluation criteria.

**(1) Data descriptions**

China's A-share stock market is focused on due to its increasingly important role in global financial markets in terms of trading volume and capitalization [43]. Overall, a total of 2,473 stocks are considered candidates for portfolios, excluding stocks in the financial industry with a distinct balance sheet structure [2, 44] and those labeled as Special Treatment (ST) that might be unsteady [45]. The studying period is from Q1 2006 to Q4 2016, which is further divided into a training subperiod (the first 80% of the whole period) and a testing subperiod (the remaining 20%). In the existing models of stock prediction, various transaction data (particularly price factors, liquidity factors, and valuation factors) have been popularly used as important factors (e.g., [5, 30, 11]), because they could effectively reflect the market information and return structure of a stock [23]. Accordingly, this study especially employs such an informative data source in the stock prediction. In particular, all available transaction data (even beyond the study period) prior to the target periods are used to sufficiently train the forecasting method. Quarterly data (including transaction data for stock prediction and fundamental factors for stock scoring) are derived from the Wind Database (http://www.wind.com.cn).

**(2) Benchmark methods**

In a CI-based stock selection method, factor design, optimization algorithm and fitness function have been considered as the important model designs [4], and the proposed method makes the major contribution by introducing stock prediction. Thus, the four model designs, i.e., forecasting model, factor design, optimization algorithm and fitness function, are especially focused on to formulate baseline models. In particular, to verify the superiority of the novel stock selection model (marked as M0), we formulate a set of similar counterparts using other forecasting models (M1), factor designs (M2), optimization algorithms (M3) and fitness functions (M4) as benchmark models for comparison, as listed Table 3.

In each type of benchmark methods, the non-target designs follow a similar way to the proposed method, to keep the fairness of the comparison; in contrast, for the target design,

various techniques popularly used in existing related studies are introduced in comparison with that of the proposed method. For forecasting methods, two typical CI techniques, i.e., back propagation neural networks (BPNN; a typical ANN) and support vector regression (SVR) that have been widely used in stock prediction [16, 45], stock portfolio construction [9, 10] and stock selection [3, 23, 32], are considered. For factor design, two existing designs are conducted in comparison with our novel design (considering both predicted and fundamental factors; as marked by A0): design A1 with only the fundamental factors in Table 2 [1, 12]; and design A2 with only the predicted factors [3, 23]. For optimization algorithms, two dominant CI algorithms in stock selection [1, 2, 4, 12], i.e., PSO and GA, are employed. As for fitness function, four effective criteria for stock evaluation are introduced, i.e., information coefficient (*IC*) [4, 44], cumulative return (*CR*) [1, 12, 32], intra-frac ile hit rate (*IFHR*) [4, 44] and winning rate (*WIN*) [32]:

$$IC = -\frac{1}{T}\sum_{t=1}^{T} -\frac{\mathrm{cov}\left(r_{i,t}, r_{i,t}^{'}\right)}{\sqrt{\mathrm{var}\left(r_{i,t}\right)\mathrm{var}\left(r_{i,t}^{'}\right)}}, \tag{25}$$

$$CR = -\prod_{t=1}^{T}\left(1+\overline{R}_{t+1}\right), \tag{26}$$

$$IFHR = -\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{r_{i,t}=1}^{m}\mathrm{sgn}\left(R_{t+1}\left(r_{i,t}\right)-M_{t+1}\right)}{2m}+\frac{\sum_{r_{i,t}=N-m+1}^{N}\mathrm{sgn}\left(R_{t+1}\left(r_{i,t}\right)-M_{t+1}\right)}{2m}\right) \tag{27}$$

$$WIN = -\frac{1}{T}\sum_{t=1}^{T}\mathrm{sgn}\left(\overline{R}_{t+1}\right), \tag{28}$$

where $T$ is the size of the training period, $r_{i,t}$ is the score ranking of stock $i$ estimated in period $t$, $r_{i,t}^{'}$ is the ranking of actual returns for the next period $t$+1, the functions cov(•) and var(•) are the covariance and variance estimations, respectively, $m$ is the number of selected stocks, $\overline{R}_{t+1}$ is the next-period returns of the portfolio constructed in period $t$, $R_{t+1}(r_{i,t})$ is the next-period returns of the stock with the current ranking $r_{i,t}$ evaluated in period $t$, $M_t$ is the average returns of all candidate stocks for period $t$, and the indicator function sgn($x$) is set to 1 when $x \geq 0$ and 0 otherwise.

**Table 3** Benchmark models for comparison

| Type | Descriptions | Forecasting model | Factor design | Optimization algorithm | Fitness function |
|---|---|---|---|---|---|
| **M0** | Novel method | ELM | A0 | DE | *Spread* |
| **M1** | Benchmarks with other forecasting models | BPNN | A0 | DE | *Spread* |
| | | SVR | A0 | DE | *Spread* |
| **M2** | Benchmarks with other factor designs | ELM | A1 | DE | *Spread* |
| | | ELM | A2 | DE | *Spread* |
| **M3** | Benchmarks with other optimization algorithms | ELM | A0 | PSO | *Spread* |
| | | ELM | A0 | GA | *Spread* |
| **M4** | Benchmarks with other fitness functions | ELM | A0 | DE | *IC* |
| | | ELM | A0 | DE | *CR* |
| | | ELM | A0 | DE | *IFHR* |
| | | ELM | A0 | DE | *WIN* |

**(3) Evaluation criteria**

To evaluate the performance of stock prediction, root mean squared error (*RMSE*) [46] and mean absolute percent error (*MAPE*) [46] are selected for level accuracy, while directional statistic (*Dstat*) is selected for directional accuracy [46] as follows:

$$RMSE = \sqrt{\frac{1}{N \cdot T'} \sum_{i=1}^{N} \sum_{t=1}^{T'} \left( P_{i,t} - \hat{P}_{i,t} \right)^2}, \tag{29}$$

$$MAPE = \frac{1}{N \cdot T'} \sum_{i=1}^{N} \sum_{t=1}^{T'} \left| \frac{P_{i,t} - \hat{P}_{i,t}}{P_{i,t}} \right|, \tag{30}$$

$$Dstat = \frac{1}{N \cdot T'} \sum_{i=1}^{N} \sum_{t=1}^{T'} a_{i,t}, \tag{31}$$

where $T'$ is the size of the testing period, $N$ is the total number of stocks, $P_{i,t}$ is the price of stock $i$ in period $t$, $\hat{P}_{i,t}$ is the corresponding predicted stock price, and $a_{i,t}=1$ if $\left(P_{i,t} - P_{i,t-1}\right)\left(\hat{P}_{i,t} - P_{i,t-1}\right) > 0$ and $a_{i,t}=0$ otherwise. To evaluate the effectiveness of models, the computation time that covers both the training and testing processes is calculated.

The main aim of stock selection is to select promising stocks to formulate profitable portfolios with high investment returns. Therefore, we evaluate the performance of stock selection methods in terms of average returns (*AR*) across the testing period. Moreover, investment risk should also be considered in portfolio construction, such that risk-adjusted returns, i.e., *SharpeRatio* [47], is adopted for measuring the excess returns per unit deviation of

a given portfolio as follows:

$$AR = \frac{1}{T'}\sum_{t=1}^{T'} \overline{R}_{t+1}, \tag{32}$$

$$SharpeRatio = \frac{E\left[\overline{R}_t - R_t^f\right]}{\sigma}, \tag{33}$$

where $R_t^f$ is the returns of a risk-free asset in period $t$, $E\left[\overline{R}_t - R_t^f\right]$ is the expected portfolio returns beyond the risk-free returns, and $\sigma$ is the standard deviation of excess returns.

**(4) Parameter settings**

The parameters of different techniques involved in both the proposed and benchmark models are specified according to previous related studies, as listed in Table 4. In particular, the parameters of optimization algorithms are determined mainly based on a recent stock selection study using various AI optimization techniques (i.e., [4]). For forecasting algorithms, the parameters in BPNN and SVR are set according to Refs. [16] and [46], respectively. For ELM, the number of hidden neurons is set to 20% of the size of training sets, and the activation function is set to the sigmoidal function based on the trial-and-error method, according to Refs. [48] and [49].

**Table 4** Parameter settings of different techniques involved in the proposed and benchmark models

| Optimization algorithms | Parameter | Value | Forecasting algorithms | Parameter | Value |
|---|---|---|---|---|---|
| DE | $P$ | 30 | ELM | $n$ | 0.1$N$ |
| | $Cr$ | 0.5 | | $TF$ | sig |
| | $\beta$ | 0.6 | BPNN | $n$ | 10 |
| GA | $P$ | 50 | | $ep$ | 3000 |
| | $cr$ | 0.6 | | $lr$ | 0.1 |
| | $mr$ | 0.5 | | $mc$ | 0.4 |
| PSO | $P$ | 40 | SVR | $K$ | RBF |
| | $c1$=$c2$ | 1.495 | | $\gamma$ | $\{2^{-15}, 2^{-13},\ldots,2^{13},2^{15}\}$ |
| | $w$ | [0.4,0.9] | | $C$ | $\{2^{-15}, 2^{-13},\ldots,2^{13},2^{15}\}$ |

Notes: For optimization algorithms, $P$ refers to the population size. In DE, $Cr$ and $\beta$ are the crossover rate and the scale factor, respectively. In GA, $cr$ is the crossover rate and $mr$ is the mutation rate. In PSO, $c1$ and $c2$ refer to the local exploration coefficient and the global exploration coefficient, respectively, and $w$ is the intra-weight. For forecasting methods, $n$ refers to the number of hidden neurons in ELM and BP. In ELM, $N$ represents the size of training set, $TF$ is the transfer function, and sig is the sigmoidal function. In BPNN, $ep$ is the number of iterations, $lr$ is the momentum constant, and $mc$ is the value of learning rate. In SVR, $K$ refers to the kernel function, RBF stands for the radial basis function, and $\gamma$ and $C$ are the kernel parameter and the penalty parameter, respectively, determined by the grid search method with grids of $\{2^{-15}, 2^{-13},\ldots,2^{13},2^{15}\}$.

Due to the randomness of the initial solutions and the random parameters, CI techniques (including DE) generate a different optimized value each time. Therefore, all models run 30 times for each case, and the results are averaged as the final output. The empirical study is conducted using MATLAB 2017b software on a computer with CPU 2.50 GHz.

## 4.2 Model effectiveness

The proposed model is employed to select stocks in the A-share market of China and then formulate equally weighted portfolios. Figs. 3 and 4 illustrate the corresponding average return and accumulative return percentages during the testing period, respectively, compared with the figures of the market average returns across all candidate stocks (marked as R1) and the A-share index (R2).

The comparison results show that the portfolios formulated by the novel method can obtain satisfactory results, with higher average returns (Fig. 3) and accumulative returns (Fig. 4) than market performance (without stock selection). The proposed model defeats the market average returns (R1) in 7 of 9 periods and the A-share index (R2) in 8 of 9 periods. Moreover, the

proposed model obtains positive returns in all cases, with exceptions of Q3 2015 and Q1 2016, which are largely impacted by the stock market crash in China [50]. However, the proposed model significantly outperforms both market average returns (R1) and the A-share index (R2) in terms of accumulative returns in all cases without exception. These results imply that the proposed stock selection model can be used as an effective tool for constructing profitable portfolios with affluent market returns.

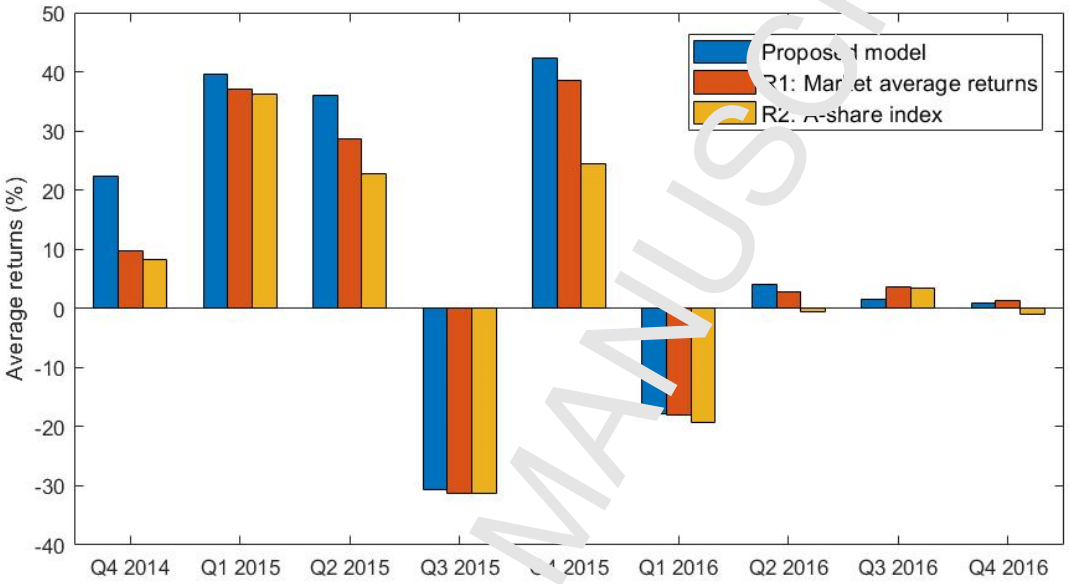



**Fig. 3** Comparison of the proposed model and market performances in terms of average returns (%)

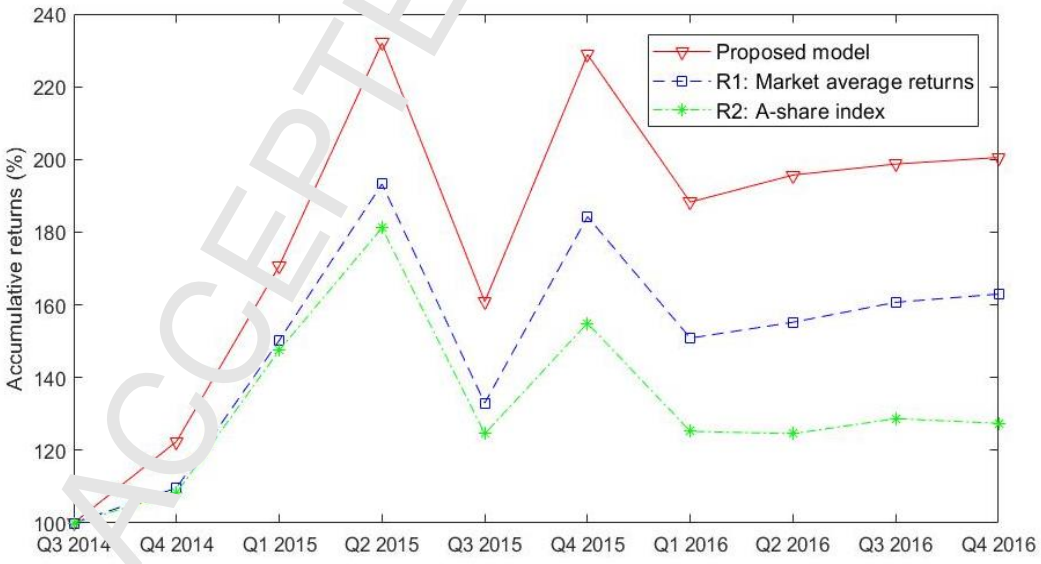



**Fig. 4** Comparison between the proposed model and market performances in term of accumulative returns (%)

To investigate the importance of the predicted factor (particularly introduced by the novel method), Fig. 5 depicts boxplots of the estimated factor weights in stock selection. The results reveal that the involved predicted factor (i.e., PR) greatly contributes to stock selection in terms of large weights (averaging approximately 0.515). Among all considered factors, the predicted factor ranks second according to average weights, closely following net income growth (NIG), which has been highlighted in previous related studies [4]. This finding supports the novel idea of introducing stock prediction into stock selection, which could provide useful (i.e., highly weighted) information regarding future stock markets.

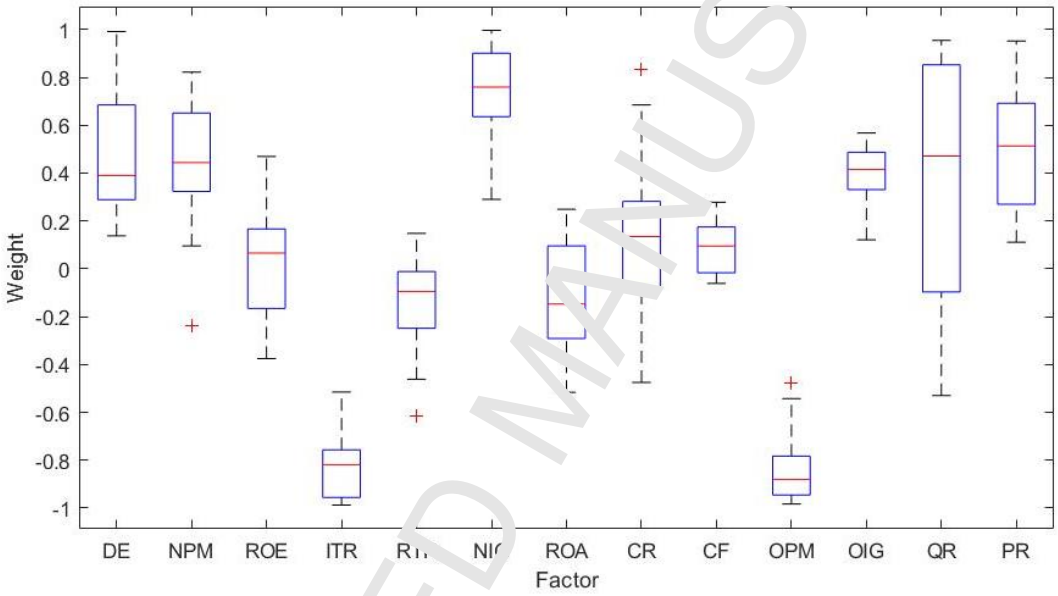



**Fig. 5** Factor importance exploration in terms of estimated weights in stock scoring mechanism

## 4.3 Superiority over benchmark models

A series of benchmark models are conducted to statistically prove the superiority of the proposed model in terms of forecasting models, factor designs, optimization algorithms and fitness functions.

**(1) Comparisons with benchmark forecasting methods (M1)**

The proposed model introduces an emerging CI technique, ELM, to achieve accurate and fast stock prediction. To verify the effectiveness of ELM in stock prediction, two classical IC technologies, i.e., BPNN (the original form of ELM) and SVR, are conducted as benchmark forecasting methods. Table 5 reports the comparison results, where Prob.(R1) and Prob.(R2) are the ratios that the target model achieves higher returns than the market average returns (R1)

and A-share index (R2) in testing period, respectively, Max. and Min. refer to the maximum and minimum returns of the generated portfolios, respectively, and HitRate is the ratio of obtaining positive returns.

**Table 5** Comparison results of stock selection models with different forecasting methods

| **Panel A: Model performance** | | | |
|---|---|---|---|
| Model | ELM | BPNN | SVR |
| *AR* | **0.1090** | 0.0855 | 0.0872 |
| *SharpeRatio* | **0.3894** | 0.3085 | 0.3117 |
| *Max.* | 0.4277 | 0.4050 | **0.4314** |
| *Min.* | -0.3068 | -0.3192 | **-0.2929** |
| *Prob*.(*R*1) | **0.6445** | 0.4037 | 0.5185 |
| *Prob*.(*R*2) | **0.7741** | 0.7444 | 0.7519 |
| *HitRatio* | **0.7595** | 0.7556 | 0.5741 |
| **Panel B: Statistical tests** (H0: $AR_{\mathrm{ELM}} \leq AR_{\mathrm{Benchmark}}$) | | | |
| Model | ELM | BPNN | SVR |
| Normality test (*p*-value) | 0.2067 | 0.1987 | 0.4371 |
| *t*- stat | N.A. | 17.7068 | 12.2034 |
| *p*-value | N.A. | **<0.0001** | **<0.0001** |
| **Panel C: Stock prediction** | | | |
| Model | ELM | BPNN | SVR |
| Computing time | **59.6290** | 16098.2450 | 32417.9660 |
| *Dstat* | **0.5254** | 0.5070 | 0.5047 |
| *MAPE* | **0.2281** | 0.2480 | 0.3667 |
| *RMSE* | **19.2819** | 33.3161 | 45.9879 |
| **Panel D: Diebold-Mariano test** (H0: Benchmark is superior to ELM in prediction performance) | | | |
| Benchmark | ELM | BPNN | SVR |
| *DM*-stat | N.A. | -17.1937 | -4.9740 |
| *p*-value | N.A. | **<0.0001** | **<0.0001** |

The results show that the proposed model with ELM performs the best in terms of *AR*, *SharpeRatio*, *Prob*.(*R*1), *Prob*.(*R*2) and *HitRatio* (Panel A in Table 5). To statistically prove the superiority of the proposed stock selection model, one-tailed *t* tests are conducted, where the null hypothesis is the inferiority of the proposed model to other benchmarks in terms of *AR* (Panel B in Table 5). The testing results demonstrate that all *AR* values of considered models follow a normal distribution at the significance level of 5% (meeting the basic assumption of *t* tests), and the proposed model with ELM defeats both benchmarking models with BPNN and SVR at the confidence level of 95% (with all *p*-values far below 5%).

The superiority of the novel stock selection model in producing higher market returns can be explained by the powerful performance of stock prediction by ELM. Panel C further compares the prediction performances of ELM, BPNN and SVR, and indicates ELM to be the most powerful in terms of both accuracy (evaluated by *Dstat*, *MAPE* and *RMSE*) and speed (evaluated by computing time). Further, the Diebold-Mariano test [51] is conducted to test the superiority of ELM, and the results show that ELM defeats the benchmark models of BPNN and SVR in prediction accuracy, at the confidence level of 95% (Panel D in Table 5). Therefore, with the most valid information regarding future stock markets based on ELM, the novel model outperforms its counterparts BPNN and SVR, thereby yielding the most profitable portfolios and affluent market returns.

**(2) Comparisons with benchmark factor designs (M2)**

The innovation of the proposed model exactly lies in the stock-scoring factors, which cover both predicted factors (regarding future stock markets) and fundamental factors (reflecting the historical performance) of candidate stocks. To confirm such a novel design, the proposed model is compared with similar counterparts with design A1 (considering only the fundamental factors in Table 2, similar to traditional stock selection models [1, 12]) and A2 (using only the predicted factors, following some current studies [3, 23]). Table 6 reports the related results, which indicate that the proposed model with design A0 (finely coupling predicted factors with fundamental factors) appears to be the best model in terms of *AR*, *SharpeRatio*, *Max.*, *Min.* and *Prob.*(*R*1). Similarly, the one-tailed *t* tests statistically support the success of design A0 under a confidence level of 95% (with all *p*-values far below 5%).

Interestingly, all stock selection models with designs A0, A1 and A2 obtain higher returns than the market performances without stock selection (R1) and A-share index (R2) in terms of *AR*, *SharpeRatio*, *Max.*, and *Prob.*(*R*1), implying the great contribution of useful information within both fundamental and predicted factors for improving stock selection. When comparing designs A1 and A2, one cannot defeat the other in terms of all criteria. However, they are both totally defeated by the novel design A0. These results underline the importance of both fundamental factors and predicted factors, which our proposed model uses to produce the best results.

**Table 6** Comparison results of stock selection models with different factor designs

| **Panel A: Model performance** | | | | | |
|---|---|---|---|---|---|
| Model | A0 | A1 | A2 | R1 | R2 |
| *AR* | **0.1090** | 0.0857 | 0.0859 | 0.0804 | 0.0479 |
| *SharpeRatio* | **0.3894** | 0.3096 | 0.3161 | 0.3055 | 0.1869 |
| *Max.* | **0.4277** | 0.4150 | 0.3911 | 0.3858 | 0.3627 |
| *Min.* | **-0.3068** | -0.3302 | -0.3652 | -0.3120 | -0.3124 |
| *Prob.*(*R*1) | **0.6444** | 0.4481 | 0.5481 | N.A. | 0.0000 |
| *Prob.*(*R*2) | 0.7741 | 0.7593 | 0.7556 | **1.0000** | N.A. |
| *HitRatio* | 0.7593 | 0.7741 | 0.7444 | **0.7778** | 0.5556 |
| **Panel B: Statistical tests (**H0: $AR_{A0} \leq AR_{Benchmark}$**)** | | | | | |
| Model | A0 | A1 | A2 | R1 | R2 |
| Normality test (*p*-value) | 0.3067 | 0.0521 | 0.3419 | N.A. | N.A. |
| *t*-stat | N.A. | 20.6365 | 14.2446 | 26.6422 | 56.9491 |
| *p*-value | N.A. | **<0.0001** | **<0.0001** | **<0.0001** | **<0.0001** |

**(3) Comparisons with benchmark optimization algorithms (M3)**

The proposed method using DE-based optimization is compared with two popular optimization algorithms, i.e., GA and PSO, as shown in Table 7. It can be easily seen from Table 7 that the proposed model performs the best in terms of *AR* and *SharpeRatio* for both the training period and the testing period, and the one-tailed *t* tests statistically confirm such a conclusion at the confidence level of 95%. This result again verifies the effectiveness of DE (though with a simple form) in stock selection. Interestingly, Das and Suganthan [28] similarly observed the superiority of the DE algorithm over PSO and GA.

**Table 7** Comparison results with stock selection models with different optimization algorithms

| **Model** | **Training period** | | **Testing period** | | **Statistical tests (**H0: $AR_{DE} \leq AR_{Benchmark}$**)** | | |
|---|---|---|---|---|---|---|---|
| | *AR* | *SharpeRatio* | *AR* | *SharpeRatio* | Normality test (*p*-value) | *t*-stat | *p*-value |
| DE | **0.1191** | **0.1279** | **0.1090** | **0.3894** | 0.3067 | N.A. | N.A. |
| PSO | 0.1097 | 0.1105 | 0.1027 | 0.3663 | 0.5000 | 3.8263 | **0.0002** |
| GA | 0.0906 | 0.1043 | 0.1015 | 0.3643 | 0.2007 | 4.3942 | **<0.0001** |

**(4) Comparisons with benchmark fitness functions (M4)**

Four other dominant fitness functions in portfolio construction, i.e., *IC*, *CR*, *IFHR* and *WIN*, are introduced and compared with the *Spread* function used in the proposed model. Panel A in Table 8 compares the performance of different models with different fitness functions in terms of *AR* and *SharpeRatio*. The results demonstrate that the proposed model with *Spread* performs

the best in terms of *AR* and *SharpeRatio*. The results of one-tailed *t* tests, where the null hypothesis is that the *AR* of the proposed model is not higher than those of benchmarks, further confirm the superiority of the proposed model, as all *p*-values are far less than the significance level of 5% (see Panel B in Table 8). Moreover, stock selection models with different fitness functions obtain higher returns than the market average performance without stock selection (R1) and A-share index (R2) in most cases, implying the importance of stock selection.

**Table 8** Comparison results with stock selection models with different fitness functions

| **Panel A: Model performance** | | | | | |
|---|---|---|---|---|---|
| Model | *Spread* | *IC* | *ICR* | *IFHR* | *WIN* |
| *AR* | **0.1090** | 0.0793 | 0.0981 | 0.1036 | 0.0949 |
| *SharpeRatio* | **0.3894** | 0.3081 | 0.3602 | 0.3888 | 0.3439 |
| **Panel B: Statistical tests** (H0: $AR_{Spread} \leq AR_{Benchmark}$) | | | | | |
| Model | *Spread* | *IC* | *ICR* | *IFHR* | *WIN* |
| Normality test (*p*-value) | 0.3067 | 0.3949 | 0.1311 | 0.3651 | 0.5000 |
| *t*-stat | N.A. | 18.5781 | 9.0940 | 3.1366 | 7.0071 |
| *p*-value | N.A. | **<0.0001** | **<0.0001** | **0.0013** | **<0.0001** |

## 4.4 Summarizations

According to the above results, it can be concluded that the proposed stock selection model using stock prediction is statistically powerful and efficient relative to market performances (without stock selection) and various benchmarking counterparts (with other model designs) in terms of investment returns and model robustness.

The portfolio formulated by the proposed stock selection model obtains far higher returns than the market average performances (i.e., the equally weighted portfolio on all candidate stocks without selection) and the A-share index of China. Noticeably, the predicted factor (introduced in this study) greatly contributes to stock selection, according to its high weight, which supports the novel idea of incorporating stock prediction into stock selection to provide useful (i.e., highly weighted) information.

Additionally, the thorough comparison with similar benchmark models (with other forecasting models, factor designs, optimization algorithms and fitness functions) verifies the superiority of the novel model. In particular, the benchmarking factor designs A1 (considering only the fundamental factors in Table 2, similar to traditional stock selection models) and A2

(using only the predicted factors) are totally defeated by the proposed model (covering both A1 and A2), confirming the novel design of coupling stock prediction and stock scoring.

## 5. Conclusions

This paper proposes a novel hybrid stock selection model by incorporating stock prediction to effectively capture the future features of complex stock markets. The novel model majorly contributes to the literature from two main perspectives. First, it might be the first attempt to couple stock prediction with stock selection to form a novel stock selection method with a predicted factor capturing future stock markets. Second, it is applied to the A-share market of China and compared with typical stock selection models (without stock prediction) and similar counterparts (with other forecasting models, factor designs, optimization algorithms and fitness functions) to verify its effectiveness and validity.

Using the A-share market of China as the study sample, the empirical results indicate that the proposed stock selection model can be used as a powerful and efficient tool for generating profitable portfolios and affluent market returns. First, it obtains far higher returns than the market average performance and A-share index, in which the novel idea of introducing stock prediction provides quite useful (highly weighted) information regarding future stock markets. Second, the proposed model statistically outperforms a series of benchmark models with other forecasting models, factor designs, optimization algorithms and fitness functions, under a confidence level of 95%. In particularly, the superiority over the factor design considering only fundamental factors (similar to traditional stock selection models) and the one using only the predicted factors confirms the effectiveness of incorporating stock prediction into stock selection.

However, the proposed hybrid stock selection model can be further improved from the following four perspectives. First, the proposed model can be extended to other tough tasks in quantitative asset management, such as portfolio formulation and market-timing determination. Second, besides the regular sliding procedure with a fixed training sample size window used in stock selection, other popular strategies (e.g., the data splitting strategy) can be employed to test the robustness of the proposed model. Third, the proposed model should be applied to other

capital markets to verify its generalizability and universality. Finally, this paper considered only transaction data as the predictive factors in stock prediction, and other related data could also be introduced. We will look into these interesting issues in the near future.

## 6. References


[1] C. F. Huang, T. N. Hsieh, B. R. Chang, & C. H. Chang, A comparative study of stock scoring using regression and genetic-based linear models, in: 2011 IEEE International Conference on Granular Computing (GrC), Taiwan, China, 2011, pp. 268-273.

[2] Y. L. Becker, P. Fei, & A. M. Lester, Stock selection: An innovative application of genetic programming methodology, in: R. Riolo, T. Soule, & B. Worzel (Eds.), Genetic Programming Theory and Practice IV, Springer US, 2007, pp. 315-334.

[3] C. F. Huang, A hybrid stock selection model using genetic algorithms and support vector regression, Appl. Soft. Comput. 12 (2) (2012) 807-818.

[4] L. Yu, L. Hu, & L. Tang, Stock selection with a novel sigmoid-based mixed discrete-continuous differential evolution algorithm, IEEE Trans. Knowl. Data Eng. 28 (7) (2016) 1891-1904.

[5] R. Zhang, Z. A. Lin, S. Chen, J. Lin, & X. Liang, Multi-factor stock selection model based on kernel support vector machine, J. Math. Res. 10(5) (2018) 9-18.

[6] G. S. Atsalakis, & K. P. Valavanis, Surveying stock market forecasting techniques–Part II: Soft computing methods, Expert Syst. Appl., 36(3) (2009) 5932-5941.

[7] X. Li, H. Xie, T. L. Wong, & F. L. Wang, Market impact analysis via sentimental transfer learning, in: 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), Jeju, South Korea, 2017, pp. 451-452.

[8] W. Chen, Y. Cai, K. Lai, & H. Xie, A topic-based sentiment analysis model to predict stock market price movement using Weibo mood, Web Intell. 14(4) (2016) 287-300.

[9] Y. Luo, B. S. Kristal, C. Schweikert, & D. F. Hsu, Combining multiple algorithms for portfolio management using combinatorial fusion, in: 2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), Oxford, United Kingdom, 2017, pp. 361-366.

[10] Y. Zhao, C. Stasinakis, G. Sermpinis, & Y. Shi, Neural network copula portfolio optimization for exchange traded funds, Quant. Financ. 18(5) (2018) 761-775.

[11] X. Fu, J. Du, Y. Guo, M. Liu, T. Dong, & X. Duan, A machine learning framework for stock selection, arXiv preprint arXiv:1806.01743, 2018.

[12] C. F. Huang, B. R. Chang, D. W. Cheng, & C. H. Chang, Feature selection and parameter optimization of a fuzzy-based stock selection model using genetic algorithms, Int. J. Fuzzy Syst. 14 (1) (2012) 65-75.

[13] A. E. Levin, Stock selection via nonlinear multi-factor models, in: NIPS'95 Proceedings of the 8th International Conference on Neural Information Processing Systems, Denver, USA, 1996, pp. 966-972.

[14] V. L. Bernard, & J. K. Thomas, Evidence that stock prices do not fully reflect the implications of current earnings for future earnings, J. Account. Econ. 13 (4) (1990) 305-340.

[15] Y. Du, Application and analysis of forecasting stock price index based on combination of ARIMA model and BP neural network, in: 2018 Chinese Control And Decision Conference (CCDC), Shenyang, China, 2018, pp. 2854-2857.

[16] M. Qiu, Y. Song, & F. Akagi, Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market, Chaos Solitons Fractals 85 (2016) 1-7.

[17] H. J. Kim, & K. S. Shin, A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets, Appl. Soft. Comput. 7 (2) (2007) 569-576.

[18] L. Xi, H. Muzhou, M. H. Lee, J. Li, D. Wei, H. Hai, & Y. Wu, A new constructive neural network method for noise processing and its application on stock market prediction, Appl. Soft. Comput. 15 (2014) 57-66.

[19] X. Li, H. Xie, Y. Song, S. Zhu, Q. Li, & F. L. Wang, Does summarization help stock prediction? A news impact analysis, IEEE Intell. Syst. 30(3) (2015) 26-34.

[20] X. Li, H. Xie, L. Chen, J. Wang, & X. Deng, News impact on stock price return via sentiment analysis, Knowl.-Based Syst. 69 (2014) 14-23.

[21] J. Wang, R. Hou, C. Wang, & L. Shen, Improved v-support vector regression model based

on variable selection and brain storm optimization for stock price forecasting, Appl. Soft. Comput. 49 (2016) 164-178.

[22] A. Kazem, E. Sharifi, F. K. Hussain, M. Saberi, & O. K. Hussain, Support vector regression with chaos-based firefly algorithm for stock market price forecasting, Appl. Soft. Comput. 13 (2) (2013) 947-958.

[23] T. S. Quah, & B. Srinivasan, Improving returns on stock investment through neural network selection, Expert Syst. Appl. 17 (4) (1999) 295-301

[24] F. Wang, Y. Zhang, Q. Rao, K. Li, & H. Zhang, Exploring mutual information-based sentimental analysis with kernel-based extreme learning machine for stock prediction, Soft Comput. 21 (12) (2017) 3193-3205.

[25] S. P. Das, & S. Padhy, Unsupervised extreme learning machine and support vector regression hybrid model for predicting energy commodity futures index, Memet. Comput. 9 (4) (2017) 333-346.

[26] X. Li, H. Xie, R. Wang, Y. Cai, J. Cao, F. Wang, ... & X. Deng, Empirical analysis: stock market prediction via extreme learning machine, Neural Comput. Appl. 27(1) (2016) 67-78.

[27] R. Dash, P. K. Dash, & R. Bisoi, A self adaptive differential harmony search based optimized extreme learning machine for financial time series prediction, Swarm Evol. Comput. 19 (2014) 25-42.

[28] S. Das, & P. N. Suganthan, Differential evolution: A survey of the state-of-the-art, IEEE Trans. Evol. Comput. 15 (1) (2011) 4-31.

[29] M. Hajjami, & G. R. Amin, Modelling stock selection using ordered weighted averaging operator, Int. J. Intell. Syst. 33(11) (2018) 2283-2292.

[30] M. Bhatia & A. Madaan, Stock portfolio performance by weighted stock selection, 2018.

[31] G. Keerthana, Stock selection using differential evolution algorithm, in: National Conference on Emerging Technologies for Sustainable Engineering & Management (NCETSEM'18), 2018.

[32] Y. C. Liu, & I. C. Yeh, Using mixture design and neural networks to build stock selection decision support systems, Neural Comput. Appl. 28 (3) (2017) 521-535.

[33] S. Yodmun, & W. Witayakiattilerd, Stock selection into portfolio by fuzzy quantitative

analysis and fuzzy multicriteria decision making, Adv. Oper. Res. (2016) 1–14.

[34] R. Peachavanish, Stock selection and trading based on cluster analysis of trend and momentum indicators, In: Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong, 2016, pp. 317–321.

[35] T. Suzuki, & Y. Ohkura, Financial technical indicator based on chaotic bagging predictors for adaptive stock selection in Japanese and American markets, Physica A 442 (2016) 50-66.

[36] G. R. Amin, & M. Hajjami, Application of optimistic and pessimistic owa and dea methods in stock selection, Int. J. Intell. Syst. 31(12) (2016) 1270-1283.

[37] H. K. Kim, J. K. Chong, K. Y. Park, & D. A. Lowther, Differential evolution strategy for constrained global optimization and application to practical engineering problems, IEEE Trans. Magn. 43 (4) (2007) 1565-1568.

[38] C. Cheadle, M. P. Vawter, W. J. Freed, & K. G. Becker, Analysis of microarray data using Z score transformation, J. Mol. Diagn. 5 (2) (2003) 73-81.

[39] L. Yu, Z. Yang, & L. Tang, A novel multistage deep belief network based extreme learning machine ensemble learning paradigm for credit risk assessment, Flex. Serv. Manuf. J. 28 (4) (2016) 576-592.

[40] G. B. Huang, Q. Y. Zhu, & C. K. Siew, Extreme learning machine: Theory and applications, Neurocomputing 70 (1-3) (2006) 489-501.

[41] K. S. Banerjee, Generalized inverse of matrices and its applications. Technometrics, 15 (1973) 197-197.

[42] K. Price, R. M. Storn, & J. A. Lampinen, Differential Evolution: A Practical Approach to Global Optimization, Springer Science & Business Media, Berlin, 2006.

[43] Y. Sun, & X. Wang, Asymmetric effects of Chinas monetary policy on the stock market: Evidence from a nonlinear VAR mode, Asian Econ. Financ. Rev. 8 (6) (2018) 745-761.

[44] Y. L. Becker, H. Fox, & P. Fei, An empirical study of multi-objective algorithms for stock ranking, in: R. Riolo, T. Soule, & B. Worzel (eds), Genetic Programming Theory and Practice V, Springer US, 2008, pp. 239-259.

[45] B. Tian, Q. Gong, Y. Yang, Z. Shang, and Y. Feng, Stock selection using support vector machine in Chinese securities exchange, J. Harbin Inst. Technol. 14 (3) (2007) 378-384.

[46] C. J. Lu, Hybridizing nonlinear independent component analysis and support vector regression with particle swarm optimization for stock index forecasting, Neural Comput. Appl. 23 (7-8) (2013) 2417-2427.

[47] A. W. Lo, The statistics of Sharpe ratios, Financ. Anal. J. 58 (4) (2002) 36-52.

[48] G. Feng, G. B. Huang, Q. Lin, & R. K. L. Gay, Error minimized extreme learning machine with growth of hidden nodes and incremental learning, IEEE Trans. Neural Networks, 20(8) (2009) 1352-1357.

[49] L. Tang, Y. Wu, & L. Yu, A randomized-algorithm-based decomposition-ensemble learning methodology for energy price forecasting, Energy 157 (2018) 526-538.

[50] Z. X. Xiao, The 2015 and early 2016 stock price crash in China: Cause and lessons, in: ISSGBM International Conference on Information, Dubai, United Arab Emirates, 2016, pp. 81-84.

[51] F. X. Diebold & R. S. Mariano, Comparing predictive accuracy, J. Bus. Econ. Stat. 13(3) (1995) 253-263.