Applied Soft Computing Journal 100 (2021) 106943

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

Mean–variance portfolio optimization using machine learning-based stock price prediction

Wei Chen [a,*], Haoyu Zhang [a], Mukesh Kumar Mehlawat [b], Lifen Jia [a]

[a] School of Management and Engineering, Capital University of Economics and Business, Beijing, China
[b] Department of Operational Research, University of Delhi, Delhi, India

ARTICLE INFO

Article history:
Received 16 June 2020
Received in revised form 22 November 2020
Accepted 24 November 2020
Available online 5 December 2020

Keywords:
Portfolio selection
Stock prediction
eXtreme Gradient Boosting
Firefly algorithm
Mean–variance model

ABSTRACT

The success of portfolio construction depends primarily on the future performance of stock markets. Recent developments in machine learning have brought significant opportunities to incorporate prediction theory into portfolio selection. However, many studies show that a single prediction model is insufficient to achieve very accurate predictions and affluent returns. In this paper, a novel portfolio construction approach is developed using a hybrid model based on machine learning for stock prediction and mean–variance (MV) model for portfolio selection. Specifically, two stages are involved in this model: stock prediction and portfolio selection. In the first stage, a hybrid model combining eXtreme Gradient Boosting (XGBoost) with an improved firefly algorithm (IFA) is proposed to predict stock prices for the next period. The IFA is developed to optimize the hyperparameters of the XGBoost. In the second stage, stocks with higher potential returns are selected, and the MV model is employed for portfolio selection. Using the Shanghai Stock Exchange as the study sample, the obtained results demonstrate that the proposed method is superior to traditional ways (without stock prediction) and benchmarks in terms of returns and risks.

1. Introduction

Constructing a portfolio by proper stock selection has been considered an essential task for individual and institutional investors. In this context, the portfolio's improvement and optimization have become one of the most concerning issues in modern financial research and investment decision-making [1]. As we know, portfolio selection's success is significantly contingent upon the performance of future stock markets. Reasonable and accurate forecasts have the potential to generate high investment returns and hedge risks [2].

The efficient market hypothesis [3] is the leading theory of finance and economics. This hypothesis put forward that it is impossible to predict the price of financial assets. However, many studies confirmed that stock prices and returns are predictable to some extent [4]. Up to now, existing studies have mainly focused on (i) statistical methods. Statistical methods aim at predicting by analyzing the past price characteristics, such as autoregressive conditional heteroscedasticity (ARCH) [5], autoregressive integrated moving average (ARIMA) [6], generalized autoregressive conditional heteroscedasticity (GARCH) [7]; and (ii) machine learning methods. Common methods are neural networks (NNs) [8,9], support vector regression (SVR) [10,11], ensemble learning [12,13], etc. A detailed review of stock prediction approaches can be found in [14,15]. Some comparative studies highlighted that machine learning has a stronger ability to deal with non-linear and non-stationary problems than statistical models [16].

In the machine learning communities (even in machine learning competitions like[1]), there are various ensemble learning algorithms, whose purpose is to reduce prediction bias and variance and achieve better predictive performance than a single algorithm [17]. Ensemble learning algorithms mainly include Adaptive Boosting (AdaBoost) [18], Gradient Boosted Decision Tree (GBDT) [13], and eXtreme Gradient Boosting (XGBoost) [19]. Among them, XGBoost has received much attention for its low computational complexity, high prediction accuracy, and outstanding efficiency. In fact, XGBoost is an improved GBDT, which is composed of many decision trees and is applied for classification and regression. Recently, XGBoost has been applied to the forecasting of financial markets. For instance, Nobre and Neves [19] proposed an expert system using the XGBoost as a binary classifier. The results demonstrated that the system could achieve higher average returns in financial markets. Dey et al. [20] used

* Corresponding author.
E-mail address: chenwei@cueb.edu.cn (W. Chen).

[1] The website is www.kaggle.com.

https://doi.org/10.1016/j.asoc.2020.106943
1568-4946/

the XGBoost to classify the stock trend in different periods, and the superiority of the XGBoost is verified by comparing it with non-ensemble algorithms. However, to the best of our knowledge, XGBoost is widely applied as a classifier for predicting the financial markets, but few studies use it for regression problems. In this paper, an XGBoost-based model is developed for stock price forecasting.

Based on the previous research in the literature, it is indicated that XGBoost has several hyperparameters that need to be set artificially, and the setting of hyperparameters has an essential impact on the prediction accuracy [21]. Moreover, even if researchers know each hyperparameter's role, it is hard to accurately determine which hyperparameters need to be adjusted and how to adjust it to obtain the optimal model [22]. Therefore, it is crucial to adopt an appropriate optimization approach to choose optimal hyperparameters. Traditional hyperparameter optimization methods, such as grid search, generally select the hyperparameters with optimal performance by comparing different combinations. However, this type of method belongs to the enumeration, which cannot be used for large-scale calculations with high precision [23]. In recent years, meta-heuristic algorithms, such as particle swarm optimization (PSO) [24], artificial bee colony (ABC) [25], salp swarm algorithm (SSA) [26] and firefly algorithm (FA) [27,28], have been used to optimize hyperparameters. At present, the standard FA, with few parameters, simple operation, and strong robustness, has been shown as a useful tool in hyperparameter optimization problems; see, for instance, Kuo et al. [29], Ibrahim and Khatib [30], Chahnasir et al. [31], Payne et al. [32], Mehr et al. [33]. However, the standard FA has the problem of quickly falling into local optimum when solving complex optimization problems. To deal with this, several variants of the traditional FA have been developed to optimize the hyperparameters. Kazem et al. [34] proposed a stock price prediction model, in which the hyperparameters of SVR is optimized by the chaotic FA. Zhang et al. [35] developed a modified FA to optimize SVR's hyperparameters for stock price prediction, in which the opposition-based chaotic strategy and dynamic adjustment strategy are introduced. Unfortunately, few current studies utilize the FA and its variants to determine the hyperparameters of the XGBoost. This paper thus proposes an improved Firefly algorithm (IFA) to optimize the hyperparameters of XGBoost.

Portfolio optimization is the distribution of wealth among several assets, in which two parameters, namely, expected returns and risks, are very important. Investors usually tend to maximize returns and minimize risks. However, high returns typically bring increased risks. The mean–variance (MV) model proposed by Markowitz [36] is one of the best models to solve the portfolio optimization problem. In this model, returns and risks are quantified by means and variances, respectively, facilitating investors to make tradeoffs between maximizing expected return and minimizing risk. After Markowitz's work, the standard MV model has been improved and extended in the following aspects: (i) Concerning multi-period portfolio selections [37–39]. (ii) Introducing the alternative risk measurements. For instance, safety-first model [40], mean-semivariance model [41], mean absolute deviation model [42], mean semiabsolute deviation models [43], etc. (iii) Including several real-world constraints, such as cardinality constraints, transaction costs [44–47]. More researches can be found in comprehensive surveys [48,49]. However, the above studies pay more attention to the improvement and expansion of the MV model, but neglect the selection of high-quality assets before forming the optimal portfolio. Actually, in the investment process, high-quality asset input is a reliable guarantee for the construction of an optimal portfolio. In recent years, some studies have been conducted by integrating the asset preselection with the portfolio selection models. For example, Paiva et al. [50] put forward an investment decision model that uses the SVM to classify assets and combines with the MV model to form an optimal portfolio. Wang et al. [16] proposed a hybrid method combining long short-term memory with the MV model, which optimizes the formation of the portfolio in combination with asset preselection. These researches showed that the combination of stock prediction and portfolio selection might provide a new perspective for financial analysis. In this study, the MV model is adopted for portfolio selection to determine each asset's proportion.

In sum, this paper proposes a novel approach to portfolio construction by combining machine learning-based model for stock prediction with the MV model for portfolio selection. Specifically, two main phases are involved in this model: stock prediction and portfolio selection. In the first phase, a hybrid model combining XGBoost with IFA is proposed to predict stock prices for the next period, where the IFA is developed to optimize the hyperparameters of the XGBoost. In the second phase, stocks with higher potential returns are selected according to each stock's predicted prices, and the MV model is employed for allocating the investment proportion of the portfolio. The proposed method provides several new features compared with existing studies on portfolio construction in general (see Table 1). The main contributions of this study can be summarized as follows.

- We develop a hybrid model named IFAXGBoost for stock price prediction, in which an improved firefly algorithm (IFA) is designed to optimize the XGBoost. In the proposed IFA, we dynamically divide the whole firefly group into an elite subgroup and a normal subgroup, and the chaotic search strategy and the PSO-based search strategy are designed accordingly.
- Unlike most studies for portfolio construction, we incorporate the hybrid model IFAXGBoost into the mean–variance model to capture the future features of stock markets, thus forming a novel portfolio construction approach.
- Using historical data from the Shanghai Stock Exchange, we compare the proposed method's performance with traditional ways (without stock prediction) and benchmarks (with other prediction models). The experimental results show the effectiveness and superiority of this method.

Section 2 introduces the forecasting methods we use. In Section 3, the MV model, is presented. Section 4 covers the details of the experiments performed. In Section 5, we report the empirical results. Finally, Section 6 summarizes our main work and outlines future work.

2. Prediction method

2.1. XGBoost method

The XGBoost is the abbreviation of "eXtreme Gradient Boosting," proposed by Chen [55] in 2016. This method's characteristics are low in computational complexity, fast in running speed, and high accuracy. The objective function of the XGBoost is to combine the standard penalty term with the loss function term to obtain the optimal solution, in which the regular penalty term can reduce the variance of the model, thus preventing over-fitting. The additive model of the tree ensemble model can be described by:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{1}$$

Table 1
Comparison with existing work.

Attribute	[19]	[9]	[51]	[52]	[53]	[44]	[45]	[54]	[28]	[35]	[47]	[50]	[16]	Proposed approach
Portfolio selection	✓	✓	×	×	×	✓	✓	×	×	×	✓	✓	✓	✓
Environment	Crisp	Crisp	Crisp	Crisp	Crisp	Fuzzy	Uncertain	Crisp	Crisp	Crisp	Crisp	Crisp	Crisp	Crisp
XGBoost	✓	×	×	×	×	×	×	×	×	×	×	×	×	✓
Hyperparameter optimization	✓	×	×	×	×	×	×	×	×	×	×	×	×	✓
Sharpe ratio	✓	×	×	×	×	×	×	×	×	×	×	×	✓	✓
Cardinality	×	×	×	×	×	×	✓	×	×	×	✓	✓	✓	✓
Financial market	BEM	US stock market	S&P 500, FTSE-100, NSE, SGX, Hang Seng, SSE	SSE	China stock market	–	–	NASDAQ	S&P-500, FTSE-100, Nikkei-225	SSE	S&P-100	Ibovespa	FTSE 100	SSE
Solution approach	MOO-GA	Neural network	Support vector machine	Adaptive SVR algorithm	PSO	FA-SA	FA-GA	Chaos based -FA	FA-MSVR	MFA	NSGA-II	Support vector machine	LSTM	IFA
Stock prediction	✓	✓	✓	✓	✓	×	×	✓	✓	✓	×	✓	✓	✓
Transaction cost	✓	×	×	×	×	✓	✓	×	✓	×	✓	✓	✓	✓

Acronyms:- BEM: Brazile exchange market, NSE: National stock exchange (India), SGX: Singapore exchange limited, SSE: Shanghai stock exchange, MOO-GA: Multiobjective optimization genetic algorithm, GA: Genetic algorithm, SVR: Support vector machine regression, PSO: Particle swarm optimization, FA-SA: Firefly algorithm and simulated annealing algorithm, FA-GA: Firefly algorithm and simulated annealing algorithm, FA: Firefly algorithm, FA-MSVR: Firefly algorithm and multi-output support vector regression, MFA: Modified firefly algorithm, NSGA-II: Non-dominated sorting genetic algorithm LSTM: long short-term memory IFA: Improved firefly algorithm.

where F is the set of regression trees, f is a tree in the tree space F, K is the number of trees and x_i represents the i^{th} eigenvector. The objective function to be optimized is as follows:

$$L(j) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \tag{2}$$

where $l(y_i, \hat{y}_i)$ is the loss function, and Ω is regular punishment that can be calculated as follows:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2, \tag{3}$$

where:

$\gamma = L1$ regularization coefficient,
$\lambda = L2$ regularization coefficient,
$T =$ number of leaves,
$\omega =$ leaf weight.

It is difficult to learn all parameters of a tree at once, so the additive strategy is adopted in Eq. (1), which is extended as follows:

$$\hat{y}_i^{(0)} = 0, \tag{4}$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i), \tag{5}$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i), \tag{6}$$

$$\cdots$$

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i). \tag{7}$$

The t^{th} objective function is described as follows:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t). \tag{8}$$

The second Taylor expansion of Eq. (8) is performed to obtain Eq. (9), where the first derivative g_i and the second derivative h_i are Eqs. (10) and (11), respectively.

$$L(t) \approx \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i)) \right] + \Omega(f_t), \tag{9}$$

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}), \tag{10}$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}). \tag{11}$$

2.2. Hyperparameter optimization method

In traditional machine learning methods, hyperparameter selection is based on experience, leading to unreliable results and increasing the prediction's randomness. For the sake of solving this problem, an improved firefly algorithm (IFA) is developed to select optimal hyperparameters of the XGBoost. In the following, the standard FA is first reviewed, and then the IFA is presented.

2.2.1. Standard firefly algorithm

In the standard FA, the brightness is used to assess the quality of fireflies, which is influenced by the fitness value of the firefly's position. In terms of the minimization problem, the brightness is inversely proportional to the fitness value. Furthermore, fireflies update their positions according to attraction, which is generally proportional to the distance between them. Let $X_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$ be the i^{th} firefly in the D-dimensional space, where $i = 1, 2, \ldots, N$ and N is the population size. The distance between any two fireflies X_i and X_j is given as follows:

$$r_{ij} = \left\| X_i - X_j \right\| = \sqrt{\sum_{d=1}^{D} (x_{id} - x_{jd})^2}, \tag{12}$$

where x_{id} and x_{jd} are the positions of firefly i and firefly j under the d^{th} dimension.

The firefly's attraction is defined as:

$$\beta(r_{ij}) = \beta_0 e^{(-\gamma {r_{ij}}^2)}, \tag{13}$$

where β_0 is the attraction at $r = 0$ and γ is the optical absorption coefficient.

For two fireflies, the less bright one moves toward the other. Assume that x_j is brighter (better) than x_i. The updated position of firefly i to firefly j is defined as follows:

$$x_{id}(t+1) = x_{id}(t) + \beta_0 e^{-\gamma {r_{ij}}^2}(x_{jd}(t) - x_{id}(t)) + \alpha(rand - \frac{1}{2}), \tag{14}$$

where t denotes the number of algorithm iterations, rand is a uniform random number within [0, 1], and α is the step size factor between [0, 1]

2.2.2. Improved firefly algorithm

The standard FA has typical problems of quickly falling into local optimum and premature convergence [56]. To overcome this demerit, we first dynamically divide the whole firefly group into an elite subgroup and a normal subgroup. The chaotic search strategy is then used for the elite subgroup to improve the local searchability. At the same time, the PSO-based search strategy is employed for the normal subgroup to enhance the global search capability and speed up the global convergence.

(1) Movement of the elite subgroup

The convergence speed and solution quality of the FA are affected by the population quality. The introduction of chaotic sequences can improve the swarm's diversity and accelerate the convergence of the global solution [57,58]. However, if the chaotic sequences are applied to all solutions, the complexity of calculation will increase. Therefore, to overcome the above drawbacks, we propose a new strategy by incorporating the chaotic sequences into elite fireflies' movement strategy. The proposed strategy is described in detail below.

(1) Initialize the position of fireflies randomly, and calculate the value of objective function.

(2) Determine the threshold p to obtain elite fireflies. Since fireflies' position changes significantly in the early stage of optimization; more attention should be paid to exploring to avoid falling into local optima. While in the later stage of optimization, after enough exploration, the search pays more attention to the exploitation to achieve a better solution. Based on this consideration, the threshold p for controlling elite fireflies is determined as:

$$p = \left\lfloor N * e^{\frac{t}{t_{max}}-1} \right\rfloor, \tag{15}$$

where $\lfloor\ \rfloor$ represents rounding down, t and t_{max} are the current number of iterations and the total number of iterations, respectively.

(3) The elite fireflies conduct the chaotic search in the search space, and the position is updated as:

$$x_{elite}(t+1) = x_{min} + c_1(x_{max} - x_{min}), \tag{16}$$

$$c_1 = \mu c_2(1 - c_2), \tag{17}$$

where x_{max} and x_{min} indicate the upper and lower bounds of the elite fireflies, respectively, c_2 is a randomly chosen number within [0, 1] and μ is the control parameter of chaotic state.

(2) Movement of the ordinary subgroup

In the standard FA, fireflies' movement is only based on the brightness, which causes fireflies to quickly lose the advantage of the optimal solution obtained, thereby reducing the search efficiency. Therefore, it is necessary to introduce other information besides brightness. In the PSO [59], each particle's movement takes advantage of its current best position p_{best} and the global best position g_{best} of the whole population. In other words, a particle moves toward not only its best previous position but also the best particle. Inspired by the PSO's exploitation mechanism, this paper introduces the global best solution g_{best} into the FA. In this way, the movement of fireflies is based on both brightness and the global optimal solution. The movement of the ordinary subgroup is expressed as follows:

$$x_{id}(t+1) = x_{id}(t) + \beta_0 e^{-\gamma r_{ij}^2}(x_{jd}(t) - x_{id}(t)) + \omega(rand - \frac{1}{2})(g_{best} - x_i), \tag{18}$$

where ω is a coefficient that reduces as the number of iterations increases:

$$\omega = \alpha(1 - \frac{t}{t_{max}}). \tag{19}$$

Finally, the flowchart and pseudo code of the IFA are described as Fig. 1 and Algorithm 1, respectively.

Algorithm 1 Improved firefly algorithm

```
Calculate the fitness values of each firefly
Initialize the population of fireflies (solutions) and a set of parameters
while t < MaxGeneration do
    Sort all fireflies according to the value of the objective function
    Obtain threshold p according to Eq. (15)
    for i = p to N do
        Update the positions of elite fireflies according to Eqs. (16) and (17)
        Evaluate new solutions and update brightness
    end for
    for i = 1 to p do
        for j = 1 to p do
            if I_j > I_i then
                Move firefly i toward j in d-dimension according to Eqs. (18) and (19)
            end if
            Evaluate new solutions and update brightness
        end for
    end for
    Rank the fireflies and find the current best
end while
```

Remark. The time complexity of the IFA is $O(n^2)$, which is the same as the standard FA. Note that n represents population size and has a significant influence on complexity, where the complexity rises with the increase of the population size. The reason for the high complexity is that the IFA compares each firefly to the whole population.

3. Mean–variance model

The MV model proposed by Markowitz [36] lays the basis for portfolio selection. In this model, the investment return and risk are quantified by expected return and variance, respectively. Rational investors always pursue the lowest risk under a specific expected return or the highest return under a particular risk, choosing an appropriate portfolio to maximize expected utility. The MV model aims to make a trade-off between maximizing returns and minimizing risks, which is expressed by a typical multi-objective optimization formula:

$$\begin{aligned} &\min \quad \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \sigma_{ij} \\ &\max \quad \sum_{i=1}^{n} x_i \mu_i \\ &\text{s.t.} \quad \begin{cases} \sum_{i=1}^{n} x_i = 1, \\ 0 \le x_i \le 1, \forall\, i = 1, \ldots, n, \end{cases} \end{aligned} \tag{20}$$

where σ_{ij} is covariance between asset i and asset j, x_i and x_j represent the proportion of the initial value, and μ_i is the expected return on asset i.

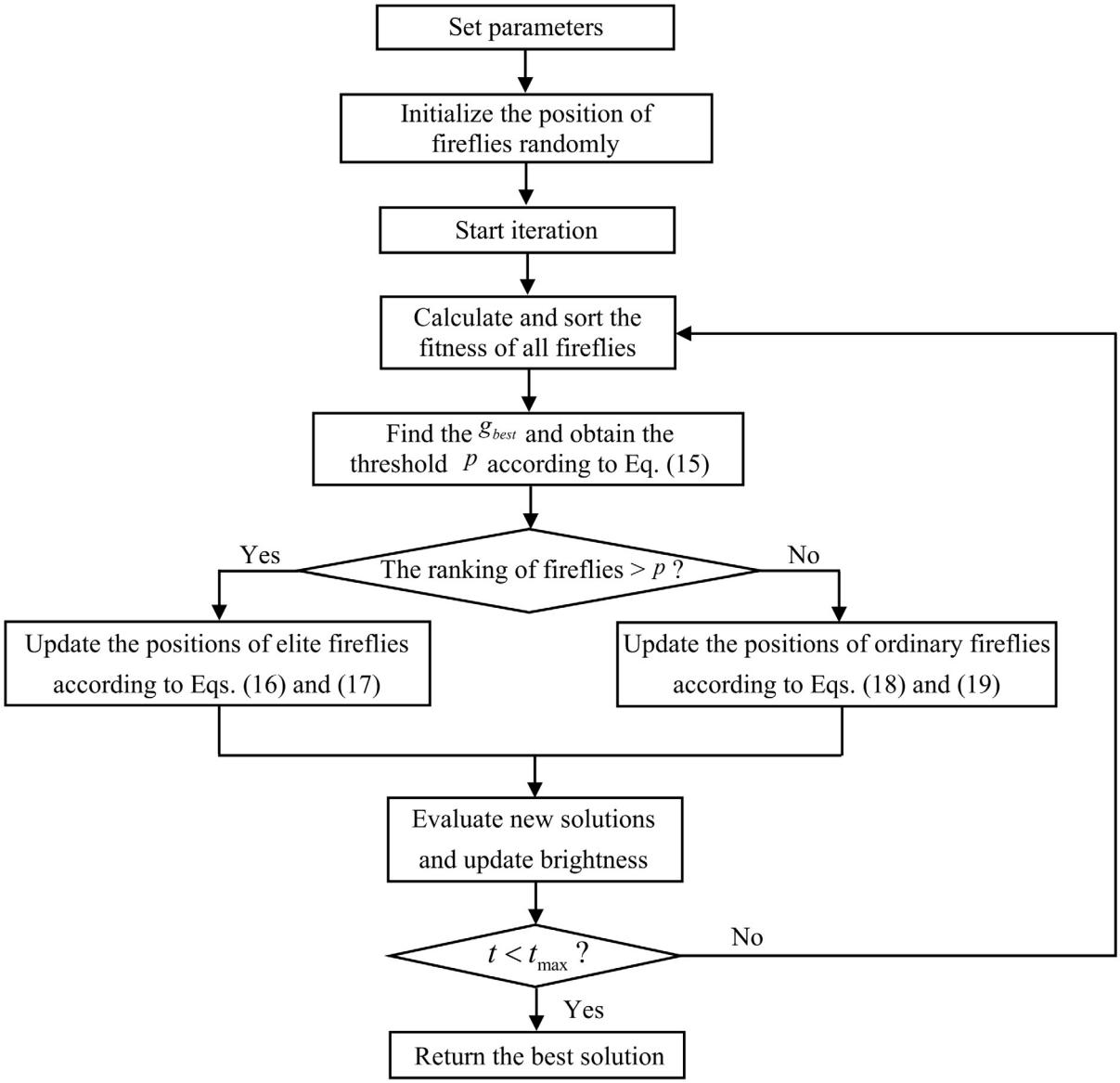

Fig. 1. The flowchart of the IFA.

Chang et al. [60] introduced the risk aversion coefficient to change the multi-objective formula into the mono-objective formulation:

$$\begin{aligned} \min \quad & \lambda \left[\sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j \sigma_{ij} \right] - (1-\lambda) \left[\sum_{i=1}^{n} x_i \mu_i \right] \\ \text{s.t.} \quad & \begin{cases} \sum_{i=1}^{n} x_i = 1, \\ 0 \le x_i \le 1, \forall\, i = 1, \ldots, n. \end{cases} \end{aligned} \tag{21}$$

Among them, the risk aversion coefficient is between 0 and 1. $\lambda = 0$ indicates that investor is very risk-averse and pursues the maximization of return without considering risk. In contrast, $\lambda = 1$ means that investors seek to minimize risk without considering a return. The value between the two extremes balances the expected return maximization and risk minimization. As a result, investors can choose one of these possible solutions based on their risk preference to form the appropriate portfolio.

4. Experimental design

4.1. Data and input indicator

4.1.1. Data

The predictability of stock prices is related to the volatility over time. Naturally, stable stocks would be easier to predict than relatively noisy ones. The Shanghai Stock Exchange (SSE) 50 index is selected according to the turnover and total market value, with the characteristics of a large scale, relatively stable, and adequate liquidity [61]. Therefore, this paper randomly selects 24 stocks in the SSE 50 index as candidate assets, large enough for individual investors to choose stocks before forming portfolios. The ticker symbol of 24 stocks are 600,000, 600,018, 600,028, 600,029, 600,031, 600,048, 600,050, 600,089, 600,100, 600,156, 600,547, 600,637, 600,690, 600,837, 600,887, 600,999, 601,006, 601,088, 601,111, 601,166, 601,186, 601,688, 601,766, 601,988. Data from November 2009 to November 2019 are collected for analysis and divided into a training set and a testing set as a ratio of 8:2. The training set is used to train the model and adjust the hyperparameters to get good generalization. The test set is

Table 2
Summary statistics for candidate assets.

Stock	Mean	Std.	Max.	Min.	Range
600,000	12.69	3.68	24.19	7.11	17.08
600,018	5.14	1.67	10.65	2.38	8.27
600,028	6.52	1.78	14.14	4.1	10.04
600,029	6.59	2.54	15.98	2.26	13.72
600,031	11.95	7.66	41.05	4.63	36.42
600,048	11.37	3.41	27.69	4.86	22.83
600,050	5.16	1.40	10.54	2.98	7.56
600,089	11.01	4.52	25.72	5.52	20.2
600,100	13.24	5.83	32.56	6.43	26.13
600,156	6.44	2.64	17.45	3.12	14.33
600,547	34.81	14.05	91.88	15.2	76.68
600,637	20.88	12.15	77.23	5.7	71.53
600,690	16.09	5.60	32.2	7.48	24.72
600,837	12.72	4.03	30.22	7.08	23.14
600,887	26.41	7.52	51.29	12.84	38.45
600,999	16.80	5.72	38.3	8.19	30.11
601,006	8.11	1.45	14.32	5.63	8.69
601,088	20.78	5.10	37.99	12.85	25.14
601,111	8.06	2.75	16.86	3.23	13.63
601,166	17.13	6.30	41.73	8.65	33.08
601,186	8.85	3.96	27.09	3.7	23.39
601,688	15.20	5.35	33.29	7.27	26.02
601,766	7.83	3.86	35.88	3.34	32.54
601,988	3.43	0.56	5.6	2.45	3.15

Table 3
Input indicators summary.

Attribute	Details	Attribute	Details
1	$\ln\left(\frac{close\ price_i}{close\ price_{i-1}}\right)$	11	$\ln\left(\frac{high\ price_{i-3}}{open\ price_{i-3}}\right)$
2	$\ln\left(\frac{close\ price_{i-1}}{close\ price_{i-2}}\right)$	12	$\ln\left(\frac{low\ price_i}{open\ price_i}\right)$
3	$\ln\left(\frac{close\ price_{i-2}}{close\ price_{i-3}}\right)$	13	$\ln\left(\frac{low\ price_{i-1}}{open\ price_{i-1}}\right)$
4	$\ln\left(\frac{close\ price_{i-3}}{close\ price_{i-4}}\right)$	14	$\ln\left(\frac{low\ price_{i-2}}{open\ price_{i-2}}\right)$
5	$\ln\left(\frac{high\ price_i}{open\ price_i}\right)$	15	$\ln\left(\frac{low\ price_{i-3}}{open\ price_{i-3}}\right)$
6	$\ln\left(\frac{high\ price_i}{open\ price_{i-1}}\right)$	16	True range
7	$\ln\left(\frac{high\ price_i}{open\ price_{i-2}}\right)$	17	Average true range (period = 14)
8	$\ln\left(\frac{high\ price_i}{open\ price_{i-3}}\right)$	18	Momentum index (period = 10)
9	$\ln\left(\frac{high\ price_{i-1}}{open\ price_{i-1}}\right)$	19	Relative strength index (period = 14)
10	$\ln\left(\frac{high\ price_{i-2}}{open\ price_{i-2}}\right)$		

used to evaluate the performance of the final model. Table 2 demonstrates the summary statistics of the close prices for the 24 stocks.

4.1.2. Input indicators

In existing stock forecasting studies, various transaction data (especially price factors) are used as important factors, because they can better reflect the stock market information [2]. Basak et al. [4] suggested that the average true range (ATR), moving average (MA), on balance volume (OBV), etc. are correlated with changes in the stock market. Zhou et al. [13] selected relative strength index (RSI), momentum index (MTM), etc. as feature subsets. Moreover, financial time-series forecasting is always explained by the lagged observations. For example, Paiva et al. [50] used several lagged variables and technical indicators to predict future stock prices. In this paper, we select 19 indicators as the input of stock prediction, including 15 lagged return observations and 4 technical indicators. To reduce the prediction errors, all indicators are uniformly mapped to the interval $[-1, 1]$. Table 3 summarizes the selected input indicators.

4.2. Proposed model: Ifaxgboost+mv

In the financial market, individual investors usually would like to know which measures should be adopted to help them possess the best portfolio of maximal potential return with minimal risk. Therefore, this paper mainly focuses on selecting stocks with high returns to form a portfolio from the perspective of prediction. The process has two critical phases, namely stock prediction (for obtaining the future prices of stocks) and portfolio selection (for forming optimal portfolio), with the overall flowchart shown in Fig. 2.

(1) **Stock prediction:** XGBoost is employed for predicting the prices of all candidate assets for the next period. Furthermore, hyperparameters of the XGBoost are optimized by the developed IFA during the training process. Mean square error (MSE) is calculated and input into the IFA as a fitness function. The main hyperparameters, such as subsample, learning_rate, reg_alpha, reg_lambda, colsample_bytree, colsample_bylevel, gamma, min_child_weight, are updated in the minimization process until the prediction error MSE meets the requirements. The prediction model named IFAXGBoost is applied to predict the stock price for the next period once the optimal hyperparameters are obtained.

(2) **Portfolio selection:** This phase aims to determine the capital proportion allocated to each asset. In this paper, the MV model is used to conduct the portfolio's asset allocation based on the selected high-quality asset and establish the optimal portfolio with unequal proportions. To be specific, the Monte Carlo method is applied to generate different portfolios. That is, randomly create a set of weights and then calculate the mean return and variance of portfolios under the weight. Repeat the process 50,000 times. From a statistical point of view, 50,000 random portfolios cover most possible portfolios with different weights and are sufficiently representative [16]. Besides, the best one is chosen from these 50,000 portfolios, according to the MV model. To find a balance between return and risk, the Sharpe ratio is used to make better decisions, and available resources are allocated to the portfolio with the largest Sharpe ratio.

4.3. Benchmarks with other alternative models

In addition to the MV model, equal-weight portfolio (1/N) has also been studied by some scholars [2,62]. The following alternative models are based on the IFAXGBoost+MV and used for comparison.

(1) Alternative model: IFAXGBoost+1/N

The design of this model is similar to the structure of IFAXGBoost+MV, using the same prediction model to select assets, but the asset proportion is evenly distributed. Specifically, in the first phase, the IFAXGBoost is used to predict the price of assets. In the second phase, the top k assets with higher expected returns are selected and allocated the same proportion. The goal of the alternative model is to test the effectiveness of the MV model with the same asset selection.

(2) Alternative model: Machine learning+MV or 1/N

These models are designed to determine whether different prediction methods impact the construction of the optimal portfolio. To be specific, stock prices are predicted using the XGBoost, LSTM, or SVR, and stocks with higher potential returns are selected for the next stage. Then, the composition of the portfolio is determined by the MV or 1/N method.

(3) Alternative model: Random+MV or 1/N

Unlike the IFAXGBoost+MV in the stock prediction phase, the Random+MV or 1/N model is carried out randomly and does not rely on any prediction. Specifically, we randomly select a number of assets from all samples and then apply MV or 1/N method to determine the portfolio. The purpose of these alternative models is to examine the necessity for stock prediction using machine learning.

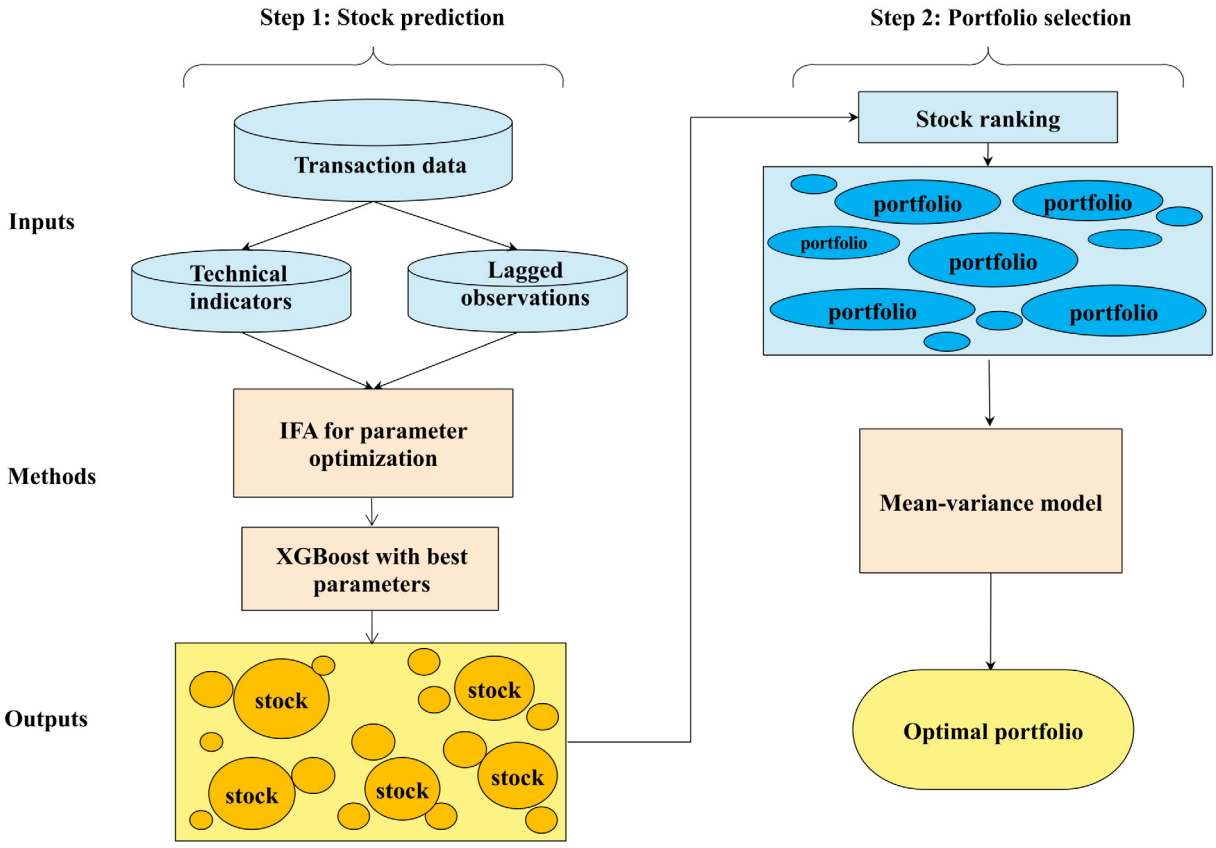

Fig. 2. Overall flowchart of IFAXGBoost+MV.

5. Experimental results

5.1. Performance of the proposed IFA

In this section, a set of classical functions are selected to evaluate the performance of the IFA, including three unimodal functions and three multimodal functions. Table 4 presents the detailed descriptions of these functions. In the following experiments, we compare the IFA with the standard FA, particle swarm optimization (PSO) [59], salp swarm algorithm (SSA) [63], gravitational search algorithm (GSA) [64], and artificial fish swarm algorithm (AFSA) [65]. The detailed parameters are given in Table 5.

Table 6 shows the maximum, minimum, mean, standard deviation, range, and average time of the different algorithms based on 20 independent runs. It can be seen that in all cases, the IFA is superior to other algorithms in terms of the maximum, minimum, mean. In most cases, IFA is better than other algorithms according to the standard deviation and range. However, the IFA requires more calculation time than the PSO, SSA, and GSA. Also, Fig. 3 provides the convergence curves of the IFA, FA, PSO, SSA, GSA and AFSA. We can see that the IFA has faster convergence speed and higher accuracy than other algorithms. To sum up, the convergence accuracy, stability, and robustness of the IFA are superior to the FA, PSO, SSA, GSA and AFSA.

5.2. Stock price prediction results

Reasonable and accurate forecasts have the potential to generate high investment returns and hedge risks. We compare the proposed IFAXGBoost with combined models, including FAXGBoost (hyperparameters of XGBoost optimized by FA), PSOXGBoost (hyperparameters of XGBoost optimized by PSO) and SSAXGBoost (hyperparameters of XGBoost optimized by SSA), and single models including XGBoost, LSTM, ELM, SVR and ANN. To investigate the accuracy of stock prediction methods, as in [23] and [35], four indexes, namely mean absolute percentage error (MAPE), mean square error (MSE), mean absolute error (MAE), and root mean square error (RMSE), are used in this paper. MSE and RMSE often represent the dispersion of results, while MAPE and MAE indicate the deviation of results. These indexes are defined as follows:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i}, \tag{22}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{23}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{24}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{25}$$

where $\hat{y}_i$ represents the predicted price, y_i indicates the actual price, and n is the total number of samples. Moreover, as in [66], the training time is used as an indicator to reflect the complexity of different prediction models.

From Tables 7 to 9, we can see that the error indexes of the IFAXGBoost are the smallest among all prediction models. For example, compared with the FAXGBoost, the MAPE, MSE, MAE, and RMSE are reduced by 2.402, 0.0004, 0.0035, and 0.0047, respectively, demonstrating the effectiveness of the improvement strategy of the IFA. Compared with the XGBoost, the MAPE, MSE, MAE, and RMSE are reduced by 6.01, 0.003, 0.0178, and 0.0215,

Table 4
The description of functions in the experiment.

Name	Function	Range	Global optimum
Sphere	$f_1(x) = \sum_{i=1}^{D} {x_i}^2$	$[-100, 100]$	0
Schwefel 2.22	$f_2(x) = \sum_{i=1}^{D} \lvert x_i \rvert + \prod_{i=1}^{D} \lvert x_i \rvert$	$[-10, 10]$	0
Sum square	$f_3(x) = \sum_{i=1}^{D} i{x_i}^2$	$[-10, 10]$	0
Rastrigin	$f_4(x) = \sum_{i=1}^{D} ({x_i}^2 - 10\cos 2\pi x_i + 10)$	$[-5.12, 5.12]$	0
Ackley	$f_5(x) = -20\exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} {x_i}^2}\right) - \exp(\frac{1}{D}\sum_{i=1}^{D}\cos 2\pi x_i) + 20 + e$	$[-32, 32]$	0
Griewank	$f_6(x) = \frac{1}{4000}\sum_{i=1}^{D} {x_i}^2 - \prod_{i=1}^{D}\cos(\frac{x_i}{\sqrt{i}}) + 1$	$[-600, 600]$	0

Table 5
Parameter settings of different algorithms.

Parameters	Population size	Dimension of function	Maximum generation
	60	30	100
IFA	β_0	γ	α_0
	0.8	0.0001	0.3
FA	β_0	γ	α_0
	0.8	0.0001	0.3
PSO	Inertia constant ω	Cognitive constant c_1	Social constant c_2
	0.8	1.5	1.5
SSA	Constant c_2	Constant c_3	
	[0, 1]	[0, 1]	
GSA	Constant G_0	Constant α	
	100	20	
AFSA	Visual parameters α	Step parameters β	Constant δ
	0.01	0.001	40

Table 6
Comparison of benchmark functions.

Functions	Algorithms	Maximum	Minimum	Mean	Std.	Range	Time (s)
Sphere	IFA	**6.02e−16**	**2.85e−19**	**8.33e−17**	**1.47e−16**	**6.02e−16**	240
	FA	0.56456	0.31269	0.45392	0.05742	0.25187	217.2
	PSO	980.040	6.81399	183.380	255.916	973.226	**7.2**
	SSA	1183.08	152.543	540.519	301.025	1030.54	11.4
	GSA	68968.2	38954.6	57461.1	7627.61	30013.6	138
	AFSA	6078.47	3124.79	5016.66	1161.53	2953.68	1432.2
Schwefel 2.22	IFA	**3.29e−17**	**3.96e−20**	**7.18e−18**	**9.27e−18**	**3.29e−17**	439.8
	FA	3.14913	2.73415	2.94920	0.13482	0.41497	437.4
	PSO	11.6648	0.03378	1.79326	3.00335	11.6310	18
	SSA	15.6728	1.84997	11.7304	5.63977	13.8228	**6.6**
	GSA	73.6206	41.8232	52.1662	8.26753	31.7974	156.0
	AFSA	102.310	91.0685	97.0715	5.02428	11.2417	1864.2
Sum square	IFA	**2.31e−17**	**1.73e−21**	**3.34e−18**	**5.78e−18**	**2.31e−17**	217.8
	FA	1.91345	1.33919	1.63745	0.16477	0.57425	273.6
	PSO	10.1369	0.00461	2.14215	2.67684	10.1323	**7.8**
	SSA	47.3368	7.26818	21.0227	10.7745	40.0686	12.6
	GSA	798.116	282.106	493.112	121.951	516.010	147.6
	AFSA	2108.52	1360.18	1771.28	203.430	748.340	1554
Rastrigin	IFA	**1.06e−14**	**0**	**2.57e−15**	**2.19e−15**	**1.06e−14**	246.6
	FA	186.856	129.501	160.166	16.6047	57.3548	301.2
	PSO	74.9568	0.18701	23.7953	21.9883	74.7697	**10.2**
	SSA	4.36474	0.97336	2.05947	0.77977	3.39137	10.8
	GSA	375.902	228.156	314.460	38.7628	147.746	143.4
	AFSA	356.465	289.224	334.102	15.9569	67.2409	1655.4
Ackley	IFA	**1.54e−08**	**1.13e−11**	**2.16e−09**	**3.58e−09**	**1.54e−08**	243
	FA	16.7882	13.1109	15.5079	0.98239	3.67728	297.0
	PSO	7.60448	0.03696	3.24591	2.37827	7.56752	**9.6**
	SSA	9.11950	6.71744	7,42947	0.99412	2.40206	11.4
	GSA	20.0653	17.8782	19.3777	0.58647	2.18706	145.8
	AFSA	20.0474	19.9629	20.0226	0.03426	0.08449	978.6
Griewank	IFA	**0.94119**	**0.00015**	**0.30975**	0.29299	0.94103	379.8
	FA	646.786	463.599	572.193	51.0674	183.186	336.0
	PSO	10.9518	0.09625	2.15254	2.65646	10.8555	**10.2**
	SSA	11.2893	3.15138	8.64558	3.44704	8.13795	11.4
	GSA	663.006	459.482	566.064	64.6618	203.523	141.6
	AFSA	3.64344	3.16135	3.40205	**0.20722**	**0.48208**	976.2

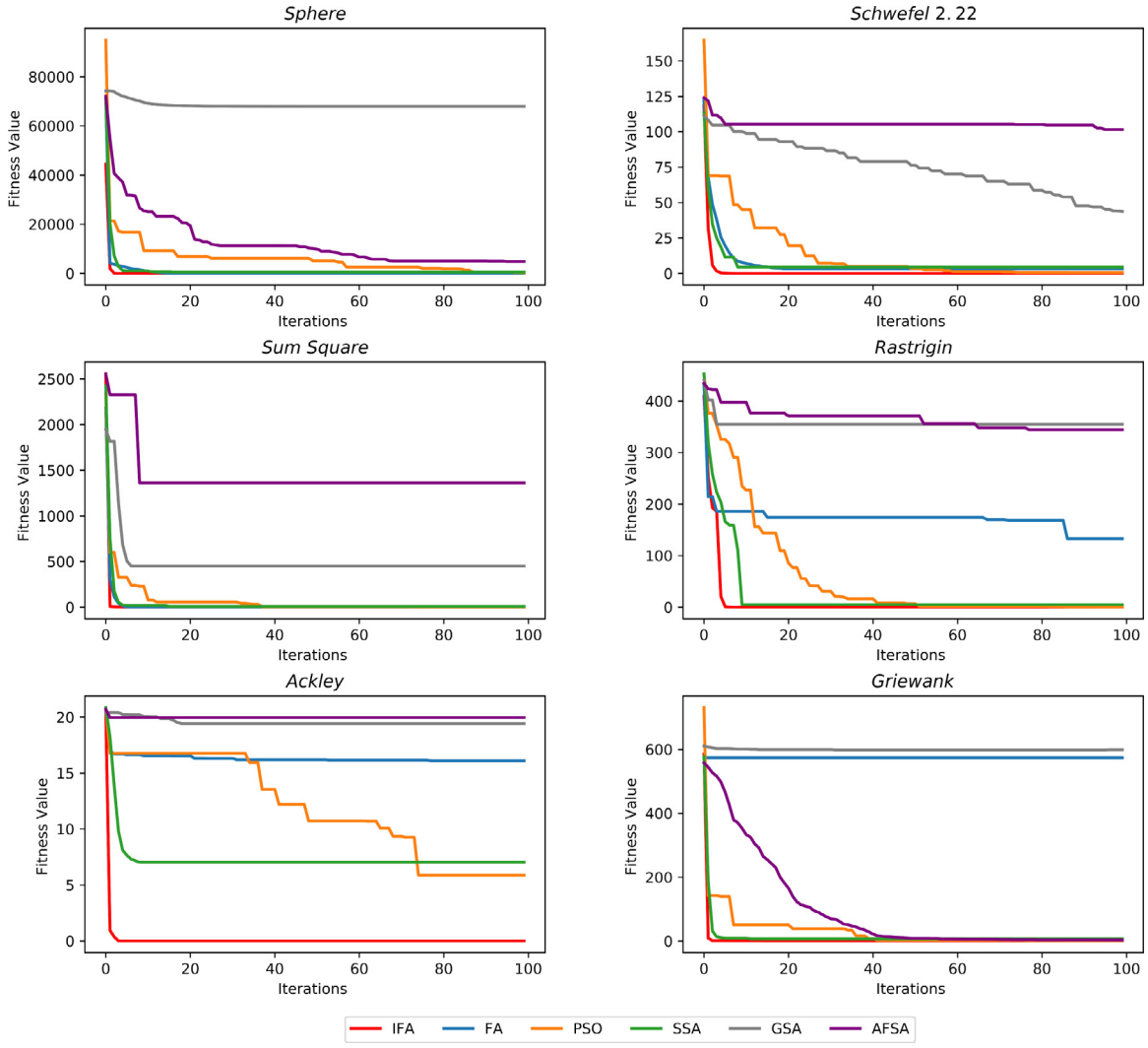

Fig. 3. The convergence characteristics of different algorithms.

Table 7
Comparison of IFAXGBoost, FAXGBoost and PSOXGBoost.

Stock	IFAXGBoost					FAXGBoost					PSOXGBoost				
	MAPE	MSE	MAE	RMSE	Time (s)	MAPE	MSE	MAE	RMSE	Time (s)	MAPE	MSE	MAE	RMSE	Time (s)
600,000	18.5540	0.0027	0.0418	0.0524	1155.51	19.3440	0.0028	0.0441	0.0538	1281.30	20.4087	0.0031	0.0456	0.0556	514.87
600,018	14.1082	0.0069	0.0623	0.0736	1690.66	14.4579	0.0073	0.0698	0.0850	1472.04	14.4657	0.0070	0.0687	0.0836	782.99
600,028	19.2996	0.0018	0.0324	0.0428	1449.16	20.7184	0.0024	0.0360	0.0491	1414.57	20.8317	0.0025	0.0372	0.0495	733.72
600,029	12.4610	0.0048	0.0537	0.0694	1585.17	13.2132	0.0054	0.0570	0.0736	1675.70	13.0103	0.0051	0.0555	0.0711	773.91
600,031	19.6052	0.0010	0.0264	0.0318	1577.08	20.3144	0.0011	0.0273	0.0342	1643.54	24.2381	0.0015	0.0319	0.0382	577.73
600,048	12.6359	0.0034	0.0471	0.0587	1271.01	13.2466	0.0037	0.0486	0.0608	1381.16	12.7312	0.0035	0.0477	0.0590	523.89
600,050	14.9606	0.0050	0.0576	0.0710	823.56	14.4151	0.0056	0.0575	0.0748	821.84	16.6843	0.0075	0.0670	0.0866	547.91
600,089	33.9649	0.0015	0.0300	0.0393	1585.02	47.3970	0.0024	0.0381	0.0493	1595.27	45.5177	0.0024	0.0374	0.0493	773.79
600,100	24.2758	0.0017	0.0310	0.0414	1581.76	28.8860	0.0026	0.0366	0.0516	1545.89	27.8571	0.0024	0.0357	0.0494	674.67
600,156	66.0393	0.0037	0.0411	0.0613	1514.53	73.3475	0.0041	0.0460	0.0646	1471.99	79.5371	0.0046	0.0454	0.0676	488.53
600,547	22.4554	0.0023	0.0378	0.0484	1108.28	26.1129	0.0026	0.0415	0.0516	1080.06	23.8798	0.0026	0.0402	0.0510	510.61
600,637	32.3155	0.0011	0.0256	0.0343	991.75	40.4817	0.0018	0.0313	0.0432	1006.04	34.8962	0.0013	0.0274	0.0366	593.03
600,690	13.8130	0.0041	0.0477	0.0643	1095.14	14.9120	0.0044	0.0516	0.0670	901.87	15.0307	0.0048	0.0533	0.0690	576.52
600,837	27.6460	0.0037	0.0474	0.0609	1124.58	31.3699	0.0043	0.0509	0.0663	1240.57	33.0029	0.0047	0.0545	0.0687	680.46
600,887	14.8002	0.0053	0.0579	0.0727	955.33	15.9756	0.0059	0.0616	0.0769	1303.24	15.6084	0.0057	0.0596	0.0751	671.01
600,999	24.9024	0.0039	0.0519	0.0620	1025.47	25.0075	0.0039	0.0529	0.0627	927.93	24.5283	0.0038	0.0522	0.0615	624.31
601,006	15.0479	0.0030	0.0463	0.0551	980.37	14.9320	0.0029	0.0453	0.0546	925.34	14.6751	0.0031	0.0453	0.0560	608.70
601,088	15.3223	0.0028	0.0412	0.0528	1175.61	16.6750	0.0032	0.0442	0.0565	1178.09	16.5083	0.0031	0.0437	0.0555	679.60
601,111	12.6768	0.0043	0.0532	0.0656	1218.12	12.6088	0.0041	0.0525	0.0645	1138.27	13.4588	0.0049	0.0566	0.0702	720.45
601,166	11.4587	0.0013	0.0302	0.0365	1046.07	13.9372	0.0017	0.0343	0.0421	955.04	14.1385	0.0016	0.0339	0.0399	670.69
601,186	7.7528	0.0007	0.0216	0.0270	892.46	10.6040	0.0014	0.0303	0.0386	800.23	9.2715	0.0012	0.0265	0.0342	582.84
601,688	12.4458	0.0044	0.0535	0.0668	1651.49	13.9070	0.0051	0.0580	0.0715	1592.63	12.5883	0.0045	0.0534	0.0670	729.27
601,766	12.2179	0.0006	0.0204	0.0261	1236.32	14.1649	0.0009	0.0252	0.0310	1152.85	16.7428	0.0013	0.0275	0.0353	644.54
601,988	10.2852	0.0029	0.0426	0.0546	1340.20	10.6571	0.0033	0.0447	0.0582	1317.20	10.7130	0.0034	0.0442	0.0581	701.72
Avg.	19.5435	0.0030	0.0416	0.0528	1253.11	21.9452	0.0035	0.0452	0.0576	1242.61	19.5995	0.0036	0.0454	0.0578	641.07

Table 8
Comparison of SSAXGBoost, XGBoost and SVR.

Stock	SSAXGBoost					XGBoost					SVR				
	MAPE	MSE	MAE	RMSE	Time (s)	MAPE	MSE	MAE	RMSE	Time (s)	MAPE	MSE	MAE	RMSE	Time (s)
600,000	18.7252	0.0029	0.0438	0.0539	704.06	19.4847	0.0038	0.0476	0.0622	0.327	22.8989	0.0044	0.0530	0.0664	0.124
600,018	15.3385	0.0082	0.0736	0.0905	961.55	16.9369	0.0103	0.0845	0.1018	0.494	13.2312	0.0071	0.0642	0.0846	0.180
600,028	21.3667	0.0024	0.0373	0.0488	921.55	22.0892	0.0033	0.0417	0.0580	0.535	30.2968	0.0081	0.0492	0.0900	0.138
600,029	13.7526	0.0060	0.0605	0.0771	916.57	14.9087	0.0089	0.0712	0.0947	0.501	20.6076	0.0057	0.0662	0.0756	0.157
600,031	21.0408	0.0012	0.0282	0.0344	861.90	24.4162	0.0014	0.0316	0.0380	0.489	25.3572	0.0016	0.0329	0.0409	0.111
600,048	13.0461	0.0040	0.0485	0.0628	745.95	18.8528	0.0075	0.0722	0.0868	0.370	15.3858	0.0070	0.0576	0.0840	0.075
600,050	17.1647	0.0060	0.0640	0.0777	702.78	17.5527	0.0085	0.0699	0.0923	0.337	24.1904	0.0132	0.0980	0.1149	0.099
600,089	45.5801	0.0022	0.0374	0.0472	906.49	45.3379	0.0020	0.0371	0.0457	0.463	79.4496	0.0047	0.0612	0.0687	0.126
600,100	28.2045	0.0024	0.0370	0.0494	731.26	26.7376	0.0027	0.0362	0.0528	0.450	51.0036	0.0048	0.0601	0.0695	0.137
600,156	82.6897	0.0051	0.0473	0.0711	729.68	78.8494	0.0067	0.0591	0.0823	0.304	114.557	0.0080	0.0773	0.0897	0.081
600,547	24.9897	0.0026	0.0405	0.0511	659.87	22.8197	0.0038	0.0435	0.0619	0.342	28.1979	0.0053	0.0525	0.0734	0.086
600,637	40.3576	0.0016	0.0303	0.0401	644.12	36.4610	0.0014	0.0284	0.0374	0.280	93.4172	0.0049	0.0616	0.0705	0.045
600,690	15.3900	0.0048	0.0524	0.0694	722.27	24.5663	0.0106	0.0868	0.1032	0.302	14.4138	0.0058	0.0512	0.0768	0.089
600,837	36.6201	0.0052	0.0579	0.0724	793.80	32.2967	0.0049	0.0546	0.0702	0.312	51.5690	0.0054	0.0613	0.0738	0.094
600,887	17.7203	0.0062	0.0647	0.0787	734.81	20.2491	0.0100	0.0786	0.1002	0.332	17.4289	0.0081	0.0698	0.0904	0.087
600,999	25.1544	0.0041	0.0544	0.0640	744.73	23.3568	0.0050	0.0574	0.0707	0.305	18.0374	0.0025	0.0418	0.0502	0.092
601,006	14.9304	0.0031	0.0460	0.0554	737.14	28.8123	0.0102	0.0918	0.1011	0.300	26.6909	0.0091	0.0870	0.0957	0.090
601,088	17.0726	0.0034	0.0456	0.0585	769.07	25.1895	0.0074	0.0714	0.0862	0.301	16.5613	0.0052	0.0492	0.0727	0.109
601,111	14.3416	0.0052	0.0596	0.0718	732.09	14.6090	0.0053	0.0602	0.0733	0.299	11.4462	0.0042	0.0486	0.0649	0.091
601,166	14.2486	0.0018	0.0349	0.0428	842.65	24.8959	0.0054	0.0638	0.0738	0.275	20.1787	0.0036	0.0503	0.0602	0.072
601,186	12.5179	0.0017	0.0348	0.0411	806.90	16.4512	0.0034	0.0463	0.0588	0.297	23.0714	0.0058	0.0670	0.0764	0.084
601,688	15.5370	0.0057	0.0633	0.0755	1005.3	16.5816	0.0073	0.0700	0.0858	0.318	19.0236	0.0092	0.0823	0.0962	0.107
601,766	14.9853	0.0011	0.0260	0.0328	963.22	21.3001	0.0020	0.0365	0.0447	0.262	14.8056	0.0015	0.0290	0.0387	0.032
601,988	10.4294	0.0035	0.0445	0.0590	775.10	20.4609	0.0108	0.0853	0.1042	0.275	26.4989	0.0154	0.1086	0.1244	0.124
Avg.	22.9668	0.0038	0.0472	0.0594	796.37	25.5507	0.0060	0.0594	0.0744	0.353	32.4200	0.0063	0.0616	0.0770	0.101

Table 9
Comparison of LSTM, ELM and ANN.

Stock	LSTM					ELM					ANN				
	MAPE	MSE	MAE	RMSE	Time (s)	MAPE	MSE	MAE	RMSE	Time (s)	MAPE	MSE	MAE	RMSE	Time (s)
600,000	25.5598	0.0081	0.0619	0.0897	21.756	25.6666	0.0055	0.0590	0.0747	0.094	24.3916	0.0065	0.0576	0.0805	11.412
600,018	12.3320	0.0055	0.0601	0.0743	30.156	17.9025	0.0131	0.0863	0.1147	0.129	23.7162	0.0171	0.1161	0.1308	16.408
600,028	29.8726	0.0080	0.0515	0.0895	29.256	30.2479	0.0115	0.0511	0.1077	0.135	30.2268	0.0052	0.0554	0.0725	17.425
600,029	11.6938	0.0046	0.0521	0.0678	29.173	13.2660	0.0055	0.0553	0.0747	0.117	19.5489	0.0132	0.0927	0.1149	16.957
600,031	22.5866	0.0021	0.0336	0.0462	24.143	20.8977	0.0014	0.0300	0.0384	0.117	32.1178	0.0032	0.0487	0.0569	17.235
600,048	14.1086	0.0051	0.0527	0.0711	27.195	13.8842	0.0044	0.0517	0.0668	0.135	27.1125	0.0133	0.1032	0.1154	13.081
600,050	14.7126	0.0062	0.0574	0.0790	23.006	18.3512	0.0095	0.0720	0.0979	0.104	29.1270	0.0175	0.1156	0.1325	13.901
600,089	49.7376	0.0031	0.0389	0.0556	28.443	50.4193	0.0030	0.0411	0.0552	0.088	52.1739	0.0023	0.0418	0.0489	15.833
600,100	23.2798	0.0016	0.0307	0.0395	32.226	38.4277	0.0033	0.0467	0.0580	0.117	37.1215	0.0034	0.0463	0.0584	17.759
600,156	86.3625	0.0056	0.0350	0.0751	26.096	88.9497	0.0063	0.0571	0.0795	0.086	54.6501	0.0022	0.0346	0.0478	13.439
600,547	25.9425	0.0067	0.0527	0.0817	25.513	26.1420	0.0077	0.0500	0.0880	0.086	31.8298	0.0061	0.0594	0.0778	15.148
600,637	50.3675	0.0032	0.0348	0.0564	24.502	87.5423	0.0053	0.0587	0.0728	0.072	36.5604	0.0014	0.0275	0.0385	12.310
600,690	18.0326	0.0072	0.0645	0.0849	26.969	14.9387	0.0053	0.0539	0.0731	0.071	20.7830	0.0082	0.0731	0.0908	13.378
600,837	36.6479	0.0056	0.0523	0.0747	27.224	38.4141	0.0042	0.0511	0.0655	0.080	40.8014	0.0068	0.0689	0.0828	14.407
600,887	17.0115	0.0073	0.0642	0.0855	28.204	16.2234	0.0071	0.0626	0.0846	0.079	21.1549	0.0101	0.0809	0.1007	14.953
600,999	17.9670	0.0040	0.0440	0.0629	29.324	18.7511	0.0034	0.0424	0.0587	0.070	27.6133	0.0062	0.0668	0.0789	14.961
601,006	30.0729	0.0110	0.0941	0.1050	29.506	24.7404	0.0079	0.0788	0.0894	0.082	35.0847	0.0136	0.1096	0.1169	15.147
601,088	22.0254	0.0072	0.0593	0.0850	29.825	17.7052	0.0060	0.0515	0.0777	0.083	17.9113	0.0045	0.0509	0.0676	14.932
601,111	12.6172	0.0058	0.0537	0.0764	30.685	12.9304	0.0065	0.0556	0.0810	0.080	12.6421	0.0046	0.0574	0.0684	15.286
601,166	23.0482	0.0049	0.0576	0.0699	35.006	16.3489	0.0026	0.0409	0.0510	0.087	28.7367	0.0065	0.0722	0.0811	16.110
601,186	16.0329	0.0033	0.0452	0.0570	31.664	19.5319	0.0045	0.0560	0.0673	0.082	17.9257	0.0039	0.0503	0.0627	16.208
601,688	12.4341	0.0049	0.0514	0.0699	30.727	14.3915	0.0062	0.0611	0.0788	0.083	15.1949	0.0060	0.0637	0.0779	15.804
601,766	27.5028	0.0027	0.0452	0.0520	31.024	15.1569	0.0012	0.0273	0.0346	0.076	19.1109	0.0016	0.0334	0.0409	16.194
601,988	30.0167	0.0187	0.1207	0.1367	32.795	25.6264	0.0142	0.1048	0.1195	0.086	19.8566	0.0088	0.0806	0.0943	17.094
Avg.	23.6349	0.0059	0.0547	0.0744	28.517	27.7690	0.0061	0.0560	0.0754	0.093	28.1413	0.0072	0.0669	0.0807	15.224

respectively, indicating the availability of hyperparameter optimization using the IFA. Compared with the other hybrid models (PSOXGBoost and SSAXGBoost), the prediction errors are also reduced, which means that the prediction accuracy after hyperparameter tuning is better than before. Also, the hybrid models, i.e., IFAXGBoost, FAXGBoost, PSOXGBoost, and SSAXGBoost perform better than the single XGBoost, which indicates that it is useful to optimize the model by adjusting hyperparameters. However, hybrid models need longer running times than single models. The reason is that the hyperparameters of hybrid models are tuned using swarm intelligence algorithms, while those of single models are set artificially.

5.3. Optimal portfolio selection results

This section will determine the proper cardinality of the portfolio and evaluate the effectiveness and superiority of the proposed IFAXGBoost+MV.

5.3.1. Determination the cardinality of portfolio

Holding too many different stocks is difficult to manage for individual investors. As a result, many studies considered a portfolio with less than ten stocks. Paiva et al. [50] argued that an average of seven assets per day is appropriate for each portfolio. Wang et al. [16] discovered that a portfolio with ten assets performs better than those with other amounts. Therefore, we

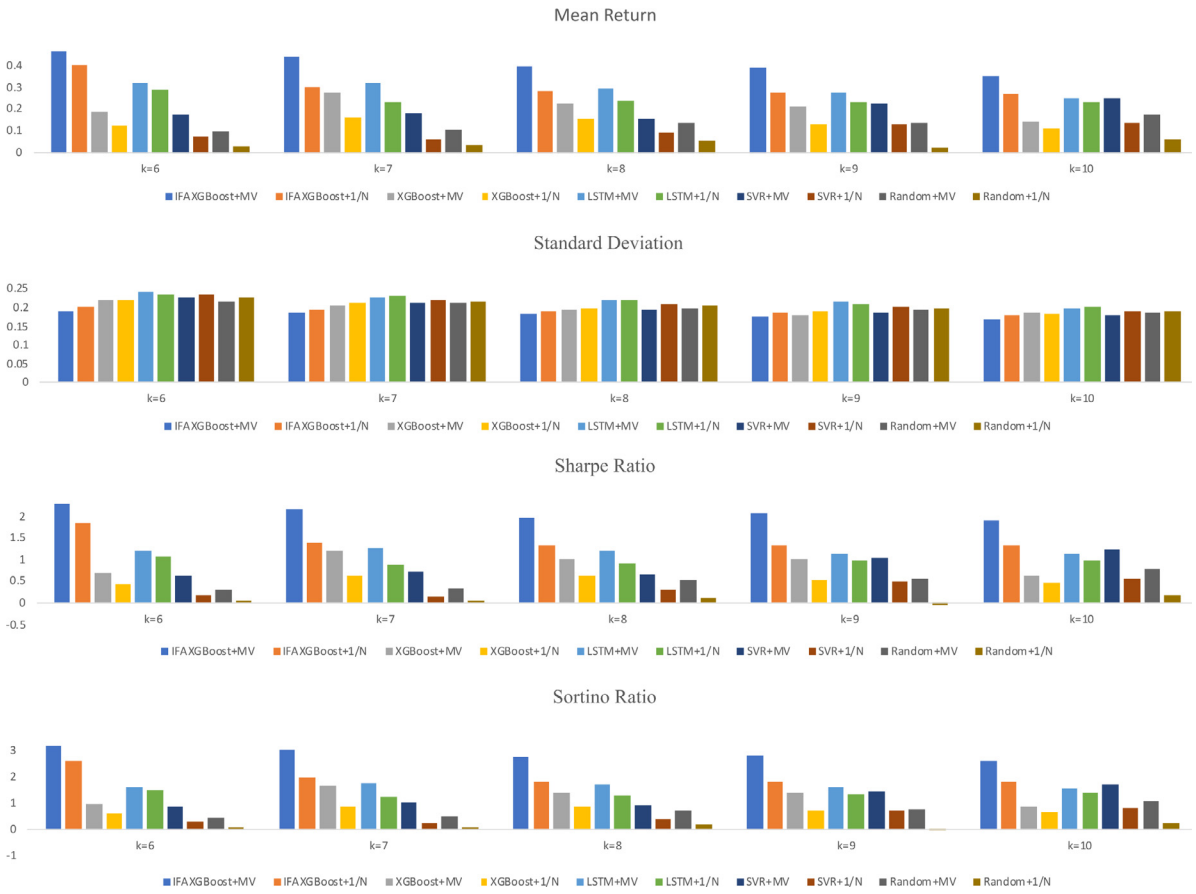

Fig. 4. Annualized performances of different portfolio's cardinalities.

Table 10
Performance characteristics without transaction cost.

Model	IFAXG-Boost+MV	IFAXG-Boost+1/N	XG-Boost+MV	XG-Boost+1/N	LSTM+MV	LSTM+1/N	SVR+MV	SVR+1/N	Random+MV	Random+1/N
Panel A: Daily return characteristics										
Mean return	0.0014	0.0010	0.0009	0.0005	0.0011	0.0008	0.0007	0.0002	0.0004	0.0001
Maximum	0.0593	0.0527	0.0563	0.0637	0.0655	0.0691	0.0582	0.0458	0.0415	0.0444
Minimum	−0.0698	−0.0530	−0.0744	−0.0663	−0.0574	−0.0572	−0.0592	−0.0662	−0.0613	−0.0674
Range	0.1292	0.1057	0.1307	0.1300	0.1229	0.1263	0.1050	0.1244	0.1029	0.1118
Skewness	−0.0825	−0.0457	−0.4562	−0.3810	−0.2044	−0.0444	−0.2254	−0.2166	−0.6127	−0.4172
Kurtosis	4.5911	2.7562	5.3739	4.0368	2.5240	2.1481	1.6780	2.4564	2.6394	2.2962
Panel B: Daily risk characteristics										
Standard dev.	0.0118	0.0122	0.0128	0.0133	0.0143	0.0145	0.0134	0.0138	0.0133	0.0135
Downside dev.	0.0083	0.0087	0.0094	0.0098	0.0103	0.0103	0.0096	0.0099	0.0098	0.0098
1-percent VaR	0.0260	0.0274	0.0288	0.0303	0.0320	0.0328	0.0303	0.0317	0.0304	0.0313
1-percent CVaR	0.0300	0.0316	0.0331	0.0348	0.0369	0.0377	0.0348	0.0364	0.0349	0.0359
5-percent VaR	0.0179	0.0191	0.0200	0.0212	0.0223	0.0229	0.0212	0.0223	0.0213	0.0220
5-percent CVaR	0.0229	0.0242	0.0254	0.0268	0.0283	0.0290	0.0268	0.0281	0.0269	0.0277
Panel C: Annualized risk-return metrics										
Mean return	0.4354	0.3004	0.2767	0.1623	0.3156	0.2284	0.1826	0.0624	0.1026	0.0330
Standard dev.	0.1879	0.1950	0.2044	0.2120	0.2270	0.2305	0.2124	0.2189	0.2113	0.2154
Downside dev.	0.1332	0.1386	0.1501	0.1559	0.1628	0.1639	0.1522	0.1576	0.1561	0.1570
Sharpe ratio	2.1576	1.3862	1.2067	0.6240	1.2582	0.8608	0.7186	0.1478	0.3439	0.0140
Sortino ratio	3.0441	1.9503	1.6437	0.8488	1.7541	1.2106	1.0030	0.2053	0.4655	0.0193
Panel D: Statistical tests for daily returns										
t-statistic	2.9190	2.0640	1.7544	0.9437	2.2995	1.4785	1.0094	0.1880	0.2610	N.A.
p-value	0.0037	0.0396	0.0800	0.3458	0.0219	0.1400	0.3133	0.8510	0.7942	N.A.

choose stocks with cardinality $k = 6, 7, 8, 9, 10$ to form portfolios. Moreover, annualized mean return, annualized standard deviation, annualized Sharpe ratio, and annualized Sortino ratio are selected to evaluate portfolios' performance. The Sharpe ratio is a comprehensive indicator that can consider both return and risk. Also, the Sortino ratio is similar to the sharpe ratio, except that it can distinguish between good and bad fluctuations. In this paper, the return of a risk-free asset is set to 0.03, according to the China treasuring bill rate in recent 10 years.

Table 11
Performance characteristics with transaction cost (0.5‰).

Model	IFAXGBoost+MV	IFAXGBoost+1/N	XGBoost+MV	XGBoost+1/N	LSTM+MV	LSTM+1/N	SVR+MV	SVR+1/N	Random+MV	Random+1/N
Panel A: Daily return characteristics										
Mean return	0.0009	0.0005	0.0005	0.00008	0.0006	0.0003	0.0002	−0.0003	0.00004	−0.0003
Maximum	0.0588	0.0522	0.0558	0.0632	0.0650	0.0686	0.0577	0.0453	0.0454	0.0439
Minimum	−0.0703	−0.0535	−0.0749	−0.0668	−0.0579	−0.0577	−0.0597	−0.0667	−0.0618	−0.0679
Range	0.1292	0.1057	0.1307	0.1300	0.1229	0.1263	0.1050	0.1244	0.1072	0.1118
Skewness	−0.0825	−0.0457	−0.4686	−0.3810	−0.2048	−0.0444	−0.2281	−0.2166	−0.4598	−0.4172
Kurtosis	4.5911	2.7562	5.4415	4.0368	2.5259	2.1481	1.7002	2.4564	2.6009	2.2962
Panel B: Daily risk characteristics										
Standard dev.	0.0118	0.0122	0.0128	0.0133	0.0143	0.0145	0.0134	0.0138	0.0131	0.0135
Downside dev.	0.0083	0.0087	0.0094	0.0098	0.0103	0.0103	0.0096	0.0099	0.0096	0.0098
1-percent VaR	0.0265	0.0279	0.0293	0.0308	0.0325	0.0333	0.0308	0.0322	0.0304	0.0317
1-percent CVaR	0.0305	0.0321	0.0336	0.0353	0.0374	0.0382	0.0353	0.0369	0.0349	0.0364
5-percent VaR	0.0184	0.0196	0.0205	0.0217	0.0228	0.0234	0.0217	0.0228	0.02147	0.0225
5-percent CVaR	0.0234	0.0247	0.0259	0.0273	0.0288	0.0295	0.0273	0.0286	0.0269	0.0282
Panel C: Annualized risk-return metrics										
Mean return	0.2650	0.1416	0.1229	0.0203	0.1607	0.0831	0.0467	−0.0634	0.0097	−0.0929
Standard dev.	0.1879	0.1948	0.2036	0.2118	0.2270	0.2305	0.2119	0.2188	0.2087	0.2151
Downside dev.	0.1332	0.1384	0.1498	0.1556	0.1628	0.1639	0.1518	0.1575	0.1524	0.1567
Sharpe ratio	1.2414	0.5731	0.4561	−0.0454	0.5756	0.2302	0.0790	−0.4272	−0.0967	−0.5714
Sortino ratio	1.7500	0.8068	0.6200	−0.0619	0.8025	0.3238	0.1103	−0.5934	−0.1325	−0.7840
Panel D: Statistical tests for daily returns										
t-statistic	2.9218	2.0640	1.7625	0.9437	2.0042	1.4788	1.0453	0.1880	0.6891	N.A.
p-value	0.0037	0.0396	0.0786	0.3458	0.0456	0.1399	0.2964	0.8510	0.4911	N.A.

The different models' annualized performances under cardinality $k = 6, 7, 8, 9, 10$ are shown in Fig. 4. It can be seen that the IFAXGBoost+MV is superior to the alternative models in terms of mean return, standard deviation, Sharpe ratio and Sortino ratio. Specifically, when the cardinality $k = 7$, the return of the IFAXGBoost+MV is 0.43, compared to 0.32 for the LSTM+MV, 0.30 for the IFAXGBoost+1/N, 0.27 for the XGBoost+MV, 0.23 for the LSTM+1/N, 0.18 for the SVR+MV, 0.16 for the XGBoost+1/N, 0.10 for the Random+MV, 0.06 for the SVR+1/N and 0.03 for the Random+1/N. As the annualized standard deviation of risk measurement, the IFAXGBoost+MV has the lowest risk when the cardinality $k = 6, 7, 8, 9, 10$. Considering the sharpe ratio, it is clear that the IFAXGBoost+MV achieves the highest level of 1.98 for $k = 8$, with the IFAXGBoost+1/N coming in second with 1.32. With regard to the sortino ratio, the IFAXGBoost provides favorable results. For example, when the cardinality $k = 10$, the IFAXGBoost+MV model has the highest sortino ratio as 2.58, followed by IFAXGBoost+1/N (1.79), SVR+MV (1.77), LSTM+MV (1.69), LSTM+1/N (1.57), Random+MV (1.06), XGBoost+MV (0.85), SVR+1/N(0.77), XGBoost+1/N (0.63) and Random+1/N (0.21).

In general, when the cardinality $k = 7$, most models perform well in terms of mean return, standard deviation, Sharpe ratio, and Sortino ratio. Hence, in this paper, we choose the portfolio with $k = 7$ for subsequent analysis, which is consistent with Paiva et al.'s research [50].

5.3.2. Superiority over benchmark models

Transaction cost is an essential factor that cannot be ignored in quantitative investment research. In China's stock market, transaction costs mainly include transfer fees, stamp duty, commissions, and other components. Stamp duty adopts single transactions, and additional fees mostly take the bilateral transaction. This paper uses unilateral transaction costs and sets the transaction costs at 0, 0.5‰ and 1‰. Tables 10 to 12 summarize the performance of each model for different transaction costs, in which Panel A, B, C, and D describe daily return characteristics, daily risk characteristics, annualized risk-return metrics, and statistical tests for daily returns, respectively.

(1) Panel A: comparison of daily return characteristics

Following Panel A of Table 10, it is evident that the IFAXGBoost+MV has the highest daily mean return: 0.0014, the LSTM+MV follows, with a return of 0.0011, and then the IFAXGBoost+1/N, with 0.0010. After considering the transaction cost, as shown in Tables 11 and 12, the IFAXGBoost+MV achieves the first place in terms of daily mean return.

(2) Panel B: comparison of daily risk characteristics

In Panel B of Tables 10 to 12, we use standard deviation, downside deviation, value at risk (VaR), and conditional value at risk (CVaR) to measure risk. We can see that: (1) When transaction costs are ignored, the IFAXGBoost+MV achieves the lowest risk, with a standard deviation of 0.0118, downside deviation of 0.0083, 1-percent VaR of 0.0260, 1-percent CVaR of 0.0300, 5-percent VaR of 0.0179, and 5-percent CVaR of 0.0229. (2) When transaction costs are considered, the IFAXGBoost+MV has the lowest standard deviation, downside deviation, 1-percent VaR, 1-percent CVaR, 5-percent VaR, 5-percent CVaR.

(3) Panel C: comparison of annualized risk return indicators

From Panel C of Table 10, the IFAXGBoost+MV has the highest annualized return of 0.44. In addition, the Sharpe ratio of the IFAXGBoost+MV reaches the highest level as 2.1576, and the IFAXGBoost+1/N ranks the second with 1.3862. As for the Sortino ratio, the IFAXGBoost+MV performs better, followed by IFAXGBoost+1/N. After considering transaction costs of 0.5‰ and 1‰, as shown in Tables 11 and 12, the IFAXGBoost+MV has the highest annualized return, Sharpe ratio and Sortino ratio.

(4) Panel D: statistical tests for daily returns

To prove the superiority of the proposed portfolio selection model statistically, one-tailed t-tests are carried out, and the Random+1/N model is taken as the benchmark model. The testing results, where the null hypothesis is that the difference between each model and the benchmark model is negative in terms of daily returns, further confirm the superiority of the proposed IFAXGBooost+MV model, as the p-value is far less than the significance level of 5% (see Panel D in Tables 10 to 12).

Moreover, Figs. 5 to 7 are presented to visualize the results of Tables 10 to 12 through box plots of daily returns. It is observed that, when transaction costs are not considered, the IFAXGBoost+MV shows the lowest volatility, followed by the

Table 12
Performance characteristics with transaction cost (1‰).

Model	IFAXGBoost+MV	IFAXGBoost+1/N	XGBoost+MV	XGBoost+1/N	LSTM+MV	LSTM+1/N	SVR+MV	SVR+1/N	Random+MV	Random+1/N
Panel A: Daily return characteristics										
Mean return	0.0004	0.00004	−0.00003	−0.0004	0.0001	−0.0002	−0.0003	−0.0008	−0.0005	−0.0008
Maximum	0.0583	0.0517	0.0553	0.0627	0.0645	0.0680	0.0448	0.0572	0.0449	0.0434
Minimum	−0.0708	−0.0540	−0.0754	−0.0673	−0.0584	−0.0581	−0.0602	−0.0672	−0.0544	−0.0684
Range	0.1292	0.1057	0.1307	0.1300	0.1229	0.1262	0.1050	0.1244	0.0993	0.1118
Skewness	−0.0825	−0.0457	−0.4683	−0.2273	−0.2048	−0.0444	−0.3810	−0.2166	−0.3590	−0.4172
Kurtosis	4.5911	2.7562	5.4381	4.0368	2.5259	2.1481	1.6805	2.4564	1.8954	2.2962
Panel B: Daily risk characteristics										
Standard dev.	0.0118	0.0122	0.0128	0.0133	0.0143	0.0145	0.0134	0.0138	0.0127	0.0135
Downside dev.	0.0083	0.0087	0.0094	0.0098	0.0103	0.0103	0.0096	0.0099	0.0092	0.0098
1-percent VaR	0.0270	0.0284	0.0298	0.0313	0.0330	0.0338	0.0313	0.0327	0.0301	0.0322
1-percent CVaR	0.0310	0.0326	0.0341	0.0358	0.0379	0.0387	0.0358	0.0374	0.0345	0.0369
5-percent VaR	0.0189	0.0201	0.0210	0.0222	0.0233	0.0239	0.0222	0.0233	0.0214	0.0230
5-percent CVaR	0.0239	0.0252	0.0264	0.0278	0.0293	0.0300	0.0278	0.0291	0.0267	0.0287
Panel C: Annualized risk-return metrics										
Mean return	0.1177	0.0106	−0.0079	−0.0966	0.0233	−0.0453	−0.0813	−0.1744	−0.1327	−0.1972
Standard dev.	0.1879	0.1949	0.2038	0.2119	0.2270	0.2303	0.2121	0.2186	0.2027	0.2151
Downside dev.	0.1332	0.1385	0.1499	0.1557	0.1628	0.1638	0.1523	0.1577	0.1474	0.1569
Sharpe ratio	0.4667	−0.0993	−0.1863	−0.5978	−0.0295	−0.3271	−0.5247	−0.9349	−0.8029	−1.0556
Sortino ratio	0.6580	−0.1397	−0.2532	−0.8131	−0.0412	−0.4600	−0.7307	−1.2958	−1.1037	−1.4477
Panel D: Statistical tests for daily returns										
t-statistic	2.9208	2.0640	1.7597	0.9437	1.9738	1.4785	1.0077	0.1880	0.4182	N.A.
p-value	0.0037	0.0396	0.0791	0.3458	0.0490	0.1400	0.3141	0.8510	0.6760	N.A.

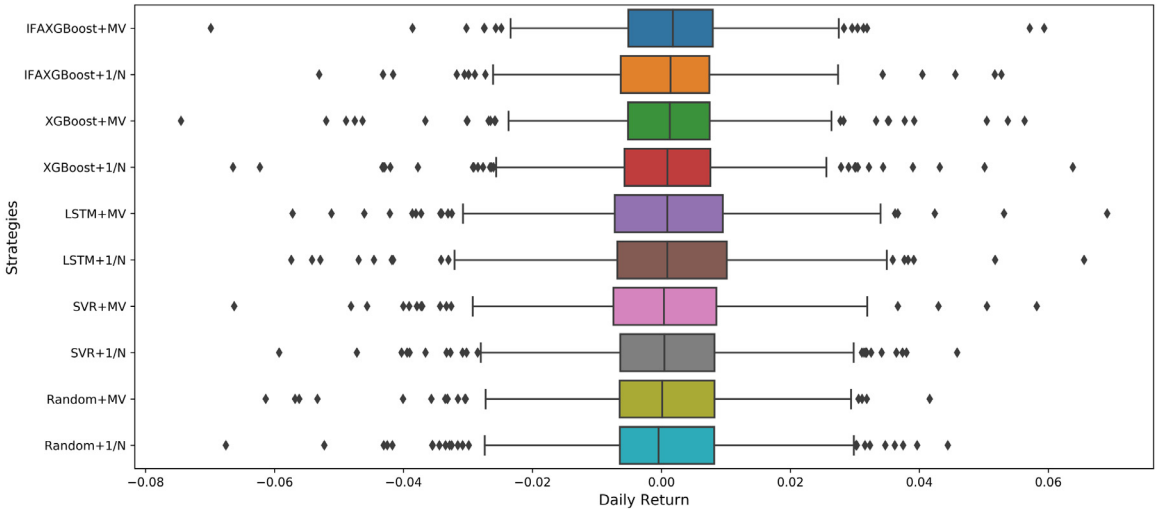

Fig. 5. Box plot of the daily returns without transaction costs.

XGBoost+MV; and when transaction costs are considered, the range of Random+MV is the lowest, but it has more outliers.

To sum up, the IFAXGBoost+MV performs significantly better than the models based on XGBoost, LSTM, SVR, and random selection in terms of return characteristics, risk characteristics, and risk-return metrics.

5.3.3. Visualization of models performance

To better show the superiority of the proposed IFAXGBoost+MV, we visualize the results in different forms. Fig. 8 demonstrates the accumulative returns of each model, excluding transaction costs. The cumulative return of the IFAXGBoost+MV is the highest compared with the benchmarks. For example, the cumulative return of the IFAXGBoost+MV is 0.94, compared to 0.62 for the IFAXGBoost+1/N, 0.56 for the XGBoost+MV, 0.31 for the XGBoost+1/N, 0.65 for the LSTM+MV, 0.46 for the LSTM+1/N, 0.36 for the SVR+MV, 0.12 for the SVR+1/N, 0.28 for the Random+MV, 0.06 for the Random+1/N. After considering the transaction costs 0.5‰ and 1‰, as can be seen from Figs. 9 and 10, respectively, the cumulative return of each model decreases significantly, but the IFAXGBoost+MV maintains the highest cumulative return.

Furthermore, inspired by the studies [16,50], we need to prove whether the good performance of the IFAXGBoost+MV only occurs within a certain period. Therefore, we compare the return–risk ratio of each model every quarter. As shown in Fig. 11, the IFAXGBoost+MV achieves positive return–risk ratio, except for Q1 2018 and Q2 2018, which are heavily influenced by the decline of the stock market in China. Also, Figs. 12 and 13 show the quarterly return–risk ratio with transaction costs 0.5‰ and 1‰ , respectively. From Fig. 12, we can observe that the IFAXGBoost+MV achieves positive returns in the six of nine survey quarters. After considering transaction costs 1‰, only five survey quarters show that the IFAXGBoost+MV has good performance.

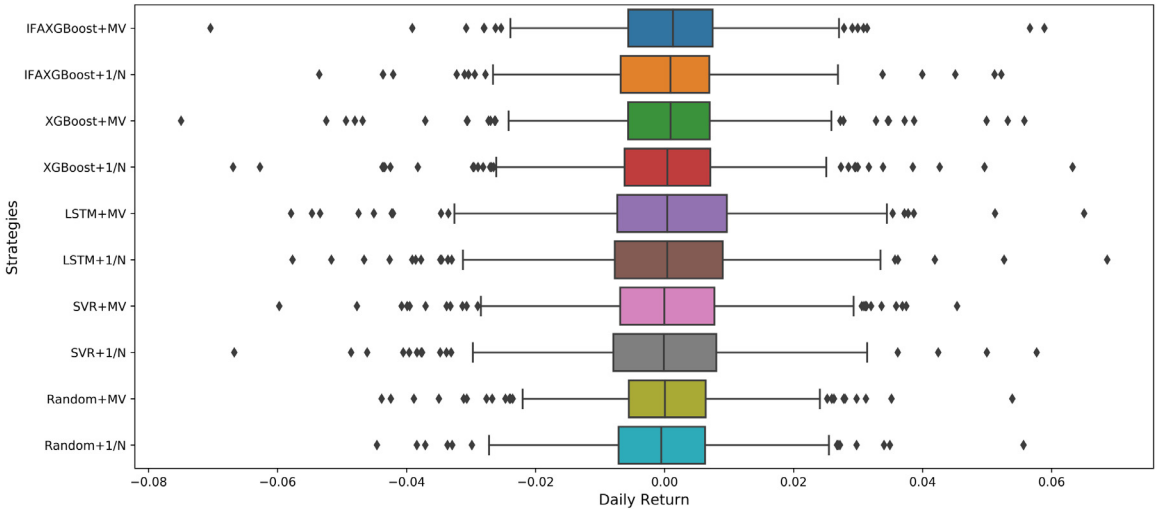

Fig. 6. Box plot of the daily returns with transaction cost (0.5‰).

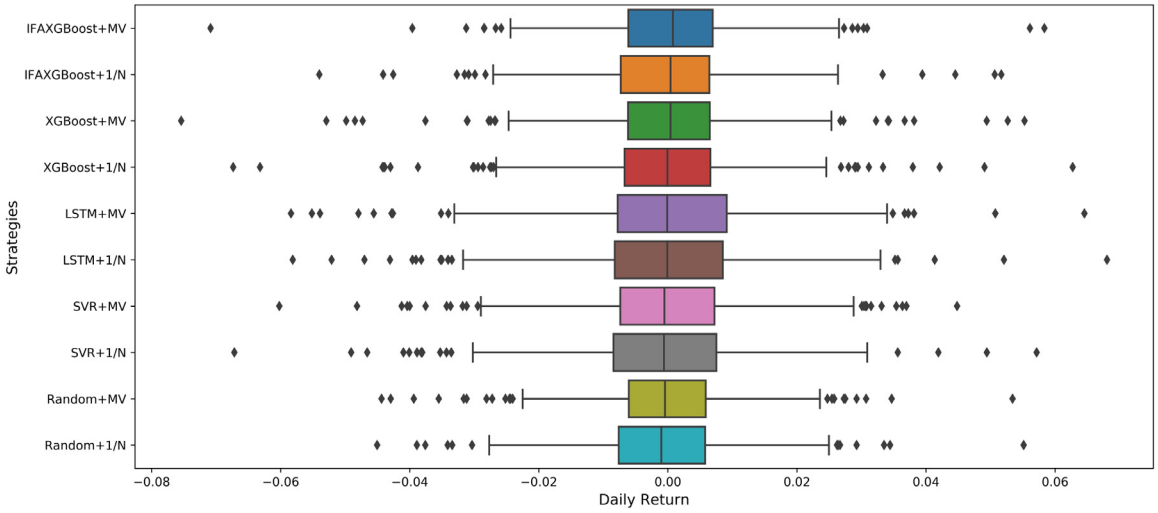

Fig. 7. Box plot of the daily returns with transaction cost (1‰).

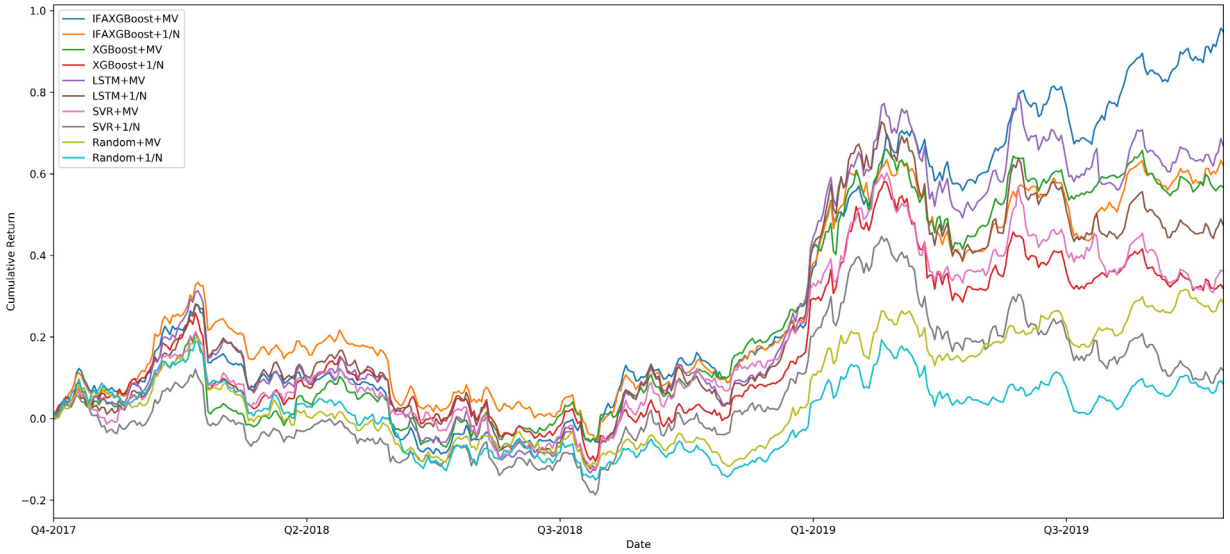

Fig. 8. Cumulative returns excluding transaction costs.

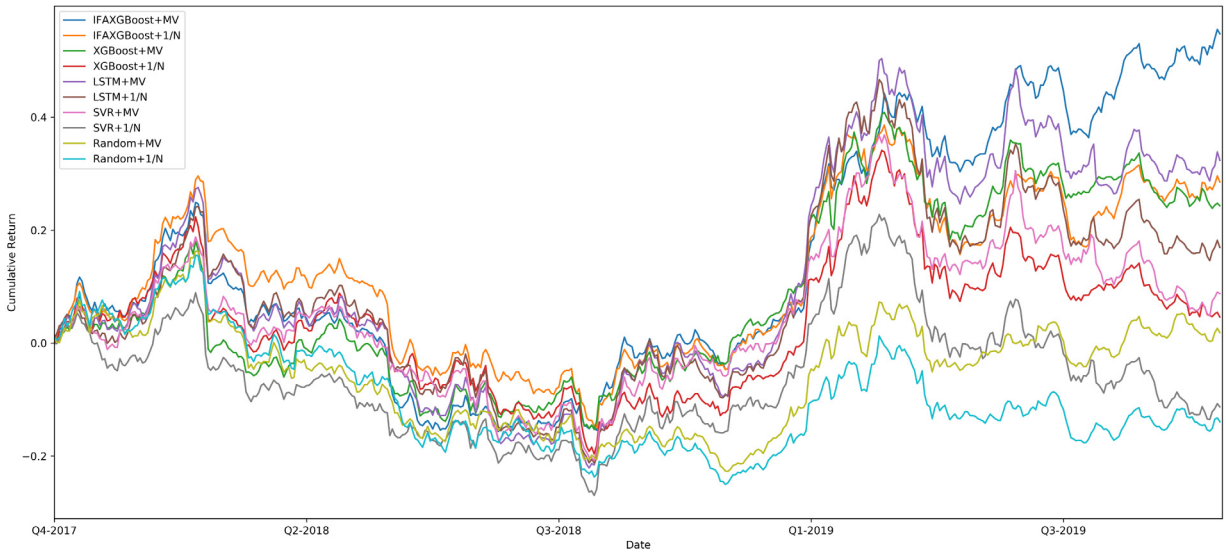

Fig. 9. Cumulative returns including transaction cost (0.5‰).

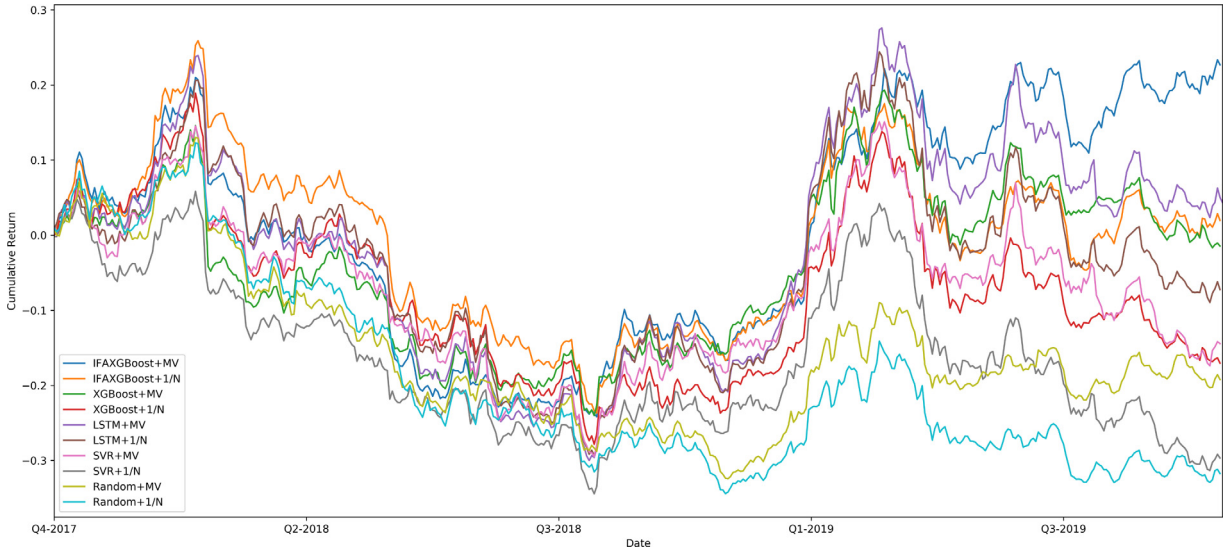

Fig. 10. Cumulative returns including transaction cost (1‰).

Overall, the IFAXGBoost+MV achieves favorable return–risk ratio in most quarters.

6. Discussion and conclusions

6.1. Discussion for key findings

This paper proposes a novel portfolio construction approach based on stock forecasting. The prediction model is developed using a hybrid method based on the XGBoost and IFA, and the MV model is used for portfolio selection. In this paper, we have several findings.

First of all, to improve the FA's optimization ability, the IFA is developed, which dynamically divides the whole firefly group into an elite subgroup and an ordinary subgroup, and the chaotic search strategy and the PSO-based search strategy are designed accordingly. After comparing the IFA against FA, PSO, SSA, AFSA as well as GSA, the advantage of IFA is verified by a set of unimodal and multimodal test functions.

Secondly, although the XGBoost has received much attention for its outstanding efficiency, it is not easy to accurately predict the complexity and diversity of the input data. To improve the prediction accuracy and avoid the negative influence of hyperparameter selection, the IFAXGBoost is developed for predicting the prices of candidate assets, which are used for portfolio selection. Further, IFAXGBoost is evaluated statistically. The outcomes of the IFAXGBoost are compared with the FAXGBoost, PSOXGBoost, SSAXGBoost, XGBoost, LSTM, SVR, ANN, and ELM. We can find that: (1) the IFAXGBoost can provide higher-quality asset input for the portfolio selection; and (2) the hybrid models based hyperparameter optimization can achieve higher prediction accuracy than the single models.

Finally, prediction results are incorporated into the optimal portfolio formation; stocks with higher potential returns are selected for constructing the portfolio. The results show the following: (1) for individual investors, holding seven assets is appropriate, (2) the IFAXGBoost+MV, IFAXGBoost+1/N, XGBoost+MV, XGBoost+MV, LSTM+MV, LSTM+1/N, SVR+MV, and SVM+1/N are

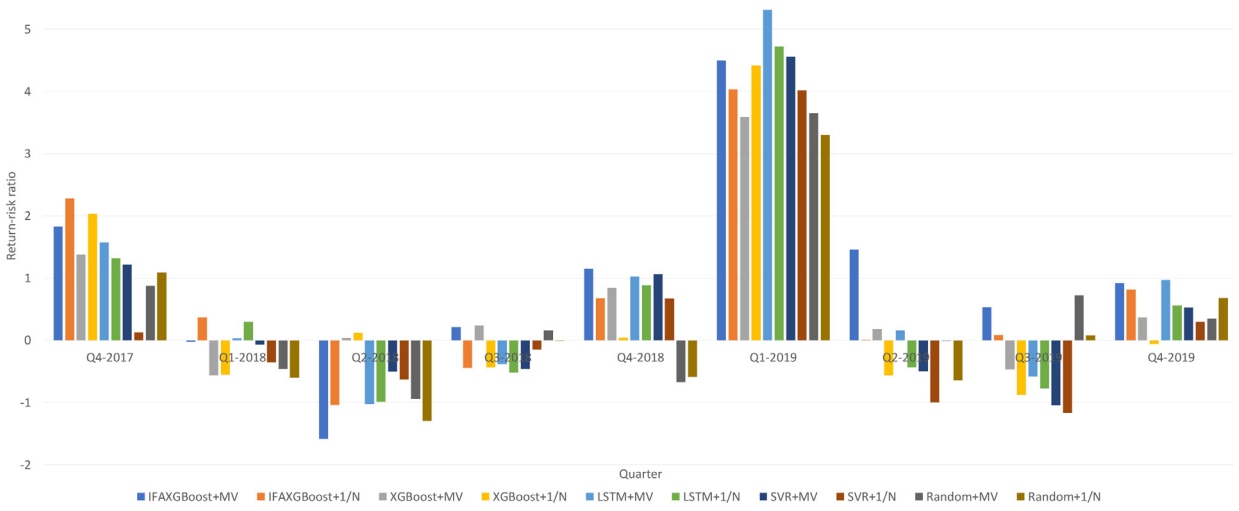

Fig. 11. Return–risk ratio of each quarter without transaction costs.

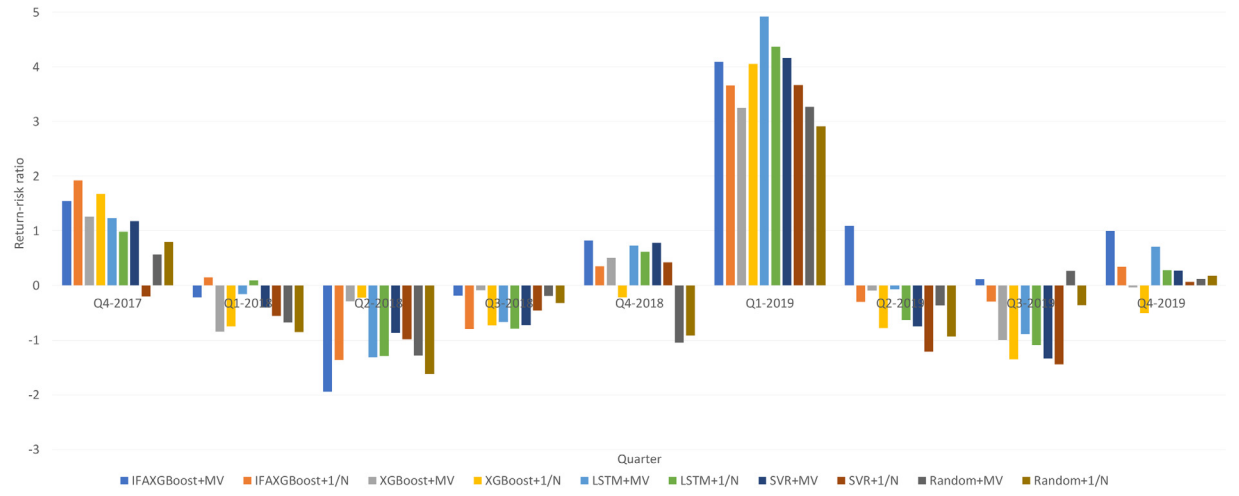

Fig. 12. Return–risk ratio of each quarter with transaction cost (0.5‰).

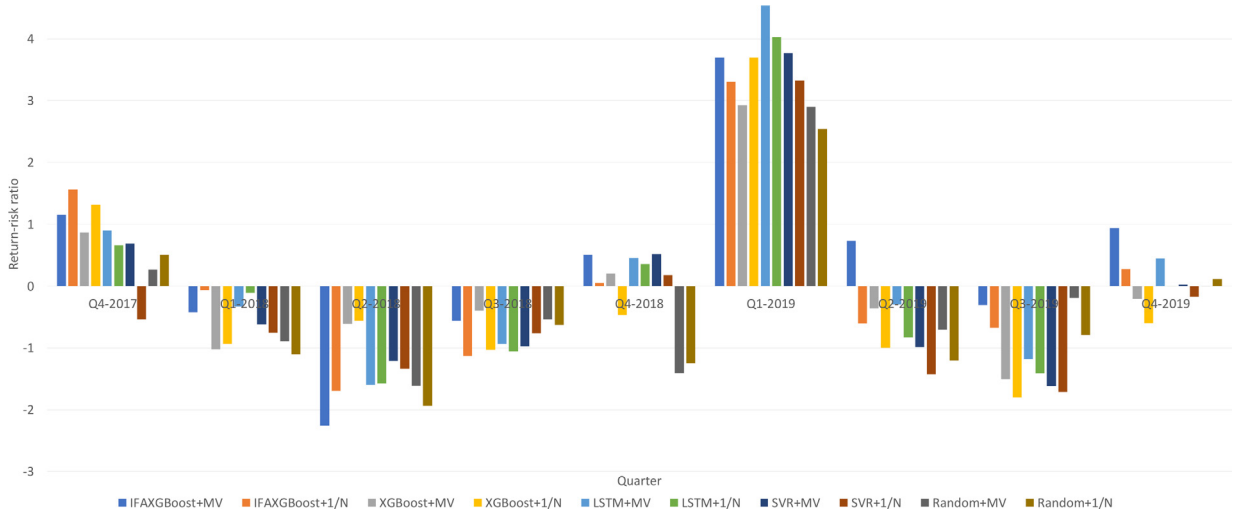

Fig. 13. Return–risk ratio of each quarter with transaction cost (1‰).

superior to Random+MV and Random+1/N in terms of return, risk, and risk-return ratio, which proves the necessity of incorporating stock prediction into portfolio selection; and (3) the models based on the MV can obtain better results than those based on the 1/N, which indicate that the MV model plays an essential role in portfolio selection. We also visualize the results in different

forms, such as box plot of daily return, line plot of cumulative return, and bar chart of return–risk ratio, which can intuitively reflect the superiority of the IFAXGBoost+MV.

In summary, the proposed IFAXGBoost+MV is superior to traditional methods (without stock prediction) and benchmarks in terms of returns, risks, and return–risk ratio. The reasons for this advantage include the following: (1) capturing the future characteristics of stock markets, (2) improving the accuracy of forecasts by the developed IFAXGBoost, and (3) introducing the MV model to improve the efficiency of asset allocation.

6.2. Future works

Although this research provides useful insights, there are some limitations to this study. The proposed study can further be improved and extended from the following aspects. First, only the Shanghai stock exchange asset data has been considered. However, due to different political environments and economic backgrounds, it is necessary to apply the IFAXGBoost+MV to other capital markets to examine its versatility and effectiveness. Besides, the proposed approach is only based on the MV for portfolio selection. However, multiple objectives such as higher moments (skewness and/or kurtosis), semi-variance, VaR, CVaR can be used for portfolio selection to better represent the real investment experience.

CRediT authorship contribution statement

Wei Chen: Conceptualization, Writing - review & editing, Supervision, Funding acquisition. **Haoyu Zhang:** Conceptualization, Methodology, Software, Writing - original draft. **Mukesh Kumar Mehlawat:** Methodology, Review & editing. **Lifen Jia:** Methodology, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Nos. 72071134, 71720107002), the Special Fund for Basic Scientific Research Operating Expenses of Beijing Municipal Colleges and Universities of Capital University of Economics and Business, Beijing, China (QNTD202002), the Project of High-level Teachers in Beijing Municipal Universities in the Period of 13th Five-year Plan, China (CIT&TCD20190338), the Humanity and Social Science Foundation of Ministry of Education of China (No. 19YJAZH005). The third author, Mukesh Kumar Mehlawat, acknowledges the support through MATRICS Scheme of DST-SERB, New Delhi, India.

References

[1] T. Bodnar, S. Mazur, Y. Okhrin, Bayesian estimation of the global minimum variance portfolio, European J. Oper. Res. 256 (1) (2017) 292–307.
[2] F. Yang, Z. Chen, J. Li, L. Tang, A novel hybrid stock selection method with stock prediction, Appl. Soft. Comput. 80 (2019) 820–831.
[3] M. Jensen, Some anomalous evidence regarding market efficiency, J. Financ. Econ. 6 (2–3) (1978) 95–101.
[4] S. Basak, S. Kar, S. Saha, L. Khaidem, S. Dey, Predicting the direction of stock market prices using tree-based classifiers, N. Am. J. Econ. Fianc. 47 (2018) 552–567.
[5] R. Engle, Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, Econometrica 50 (4) (1982) 987–1007.
[6] G. Box, G. Jenkins, Time series analysis: forecasting and control, J. Time 31 (4) (1976) 238–242.
[7] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, J. Econom. 31 (3) (1986) 307–327.
[8] R. Bisoi, P. Dash, A. Parida, Hybrid Variational Mode Decomposition and evolutionary robust kernel extreme learning machine for stock price and movement prediction on daily basis, Appl. Soft. Comput. 74 (2019) 652–678.
[9] Y. Liu, I. Yeh, Using mixture design and neural networks to build stock selection decision support systems, Neural Comput. Appl. 28 (3) (2017) 521–535.
[10] W.C. Hong, M.W. Li, J. Geng, Y. Zhang, Novel chaotic bat algorithm for forecasting complex motion of floating platforms, Appl. Math. Model. 72 (2019) 425–443.
[11] R. Chen, C.Y. Liang, W.C. Hong, D.X. Gu, Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm, Appl. Soft. Comput. 26 (2015) 435–443.
[12] S. Karasu, A. Altan, Recognition mode or solar radiation time series based on random forest with feature selection approach, in: 2019 11th International Conference on Electrical and Electronics Engineering, ELECO, 2019, pp. 8–11.
[13] F. Zhou, Q. Zhang, D. Sornette, L. Jiang, Cascading logistic regression onto gradient boosted decision trees forforecasting and trading stock indices, Appl. Soft. Comput 84 (2019) 105747.
[14] D. Gandhmal, K. Kumar, Systematic analysis and review of stock market prediction techniques, Compu. Sci. Rev. 34 (2019) 100190.
[15] O. Bustos, A. Pomares-Quimbaya, Stock market movement forecast: A Systematic review, Expert Syst. Appl. 156 (2020) 113464.
[16] W. Wang, W. Li, N. Zhang, K. Liu, Portfolio formation with preselection using deep learning from long-term financial data, Expert Syst. Appl. 143 (2020) 113042.
[17] O. Sagi, L. Rokach, Ensemble learning: A survey, Wiley Interdiscip. Rev.-Data Mining Knowl. Discov. 8 (4) (2018) e1249.
[18] X. Zhang, A. Li, R. Pan, Stock trend prediction based on a new status box method and AdaBoost probabilistic support vector machine, Appl. Soft. Comput. 49 (2016) 385–398.
[19] J. Nobre, R. Neves, Combining principal component analysis, discrete wavelet transform and xgboost to trade in the financial markets, Expert Syst. Appl. 125 (2019) 181–194.
[20] S. Dey, Y. Kumar, S. Saha, S. Basak, Forecasting to classification: Predicting the direction of stock market price using xtreme gradient boosting, Working paper, 2016.
[21] K. Song, F. Yan, T. Ding, L. Gao, S. Lu, A steel property optimization model based on the XGBoost algorithm and improved PSO, Comput. Mater. Sci. 174 (2020) 109472.
[22] S. Zhao, D. Zeng, W. Wang, X. Chen, Z. Zhang, F. Xu, X. Liu, Mutation grey wolf elite PSO balanced XGBoost for radar emitter individual identification based on measured signals, Measurement (2020) 107777.
[23] L. Li, X. Zhao, M. Tseng, R. Tan, Short-term wind power forecasting based on support vector machine with improved dragonfly algorithm, J. Clean Prod. 242 (2020) 118447.
[24] E. Hajizadeh, M. Mahootchi, A. Esfahanipour, M. Massahi, A new NN-PSO hybrid model for forecasting Euro/Dollar exchange rate volatility, Neural Comput. Appl. 31 (7) (2019) 2063–2071.
[25] Y.Z. Wang, Y.L. Nim, S. Lu, J.G. Wang, X.Y. Zhang, Remaining useful life prediction of lithium-ion batteries using support vector regression optimized by artificial bee colony, IEEE Trans. Veh. Technol. 68 (10) (2019) 9543–9553.
[26] S. Mirjalili, A. Gandomi, S. Mirjalili, S. Saremi, H. Faris, Salp swarm algorithm: a bio-inspired optimizer for engineering design problems, Adv. Eng. Softw. 114 (2019) 163–191.
[27] T. Xiong, Y. Bao, Z. Hu, Multiple-output support vector regression with a firefly algorithm for interval-valued stock price index forecasting, Knowl.-Based Syst. 55 (2014) 87–100.
[28] M. Akhavan-Amjadi, Fetal electrocardiogram modeling using hybrid evolutionary firefly algorithm and extreme learning machine, Multidimens. Syst. Signal Process. 31 (2020) 117–133.
[29] R.J. Kuo, P.S. Li, Taiwanese export trade forecasting using firefly algorithm based K-means algorithm and SVR with wavelet transform, Comput. Ind. Eng. 99 (2016) 153–161.
[30] I. Ibrahim, T. Khatib, A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm, Energy Conv. Manag. 138 (2017) 413–425.
[31] E.S. Chahnasir, Y. Zandi, M. Shariati, E. Dehghani, A. Toghroli, E.T. Mohamed, A. Shariati, M. Safa, K. Wakil, M. Khorami, Application of support vector machine with firefly algorithm for investigation of the factors affecting the shear strength of angle shear connectors, Smart. Struct. Syst. 22 (4) (2018) 413–424.
[32] A. Payne, G. Avendano-Franco, E. Bousquet, A. Romero, Firefly algorithm applied to noncollinear magnetic phase materials prediction, J. Chem. Theory Comput. 14 (8) (2018) 4455–4466.

[33] A.D. Mehr, V. Nourani, V.K. Khosrowshahi, M.A. Ghorbani, A hybrid support vector regression-firefly model for monthly rainfall forecasting, Int. J. Environ. Sci. Technol. 16 (1) (2019) 335–346.
[34] A. Kazem, E. Sharifi, F. Hussain, M. Saberi, O. Hussain, Support vector regression with chaos-based firefly algorithm for stock market price forecasting, Appl. Soft. Comput. 13 (2) (2013) 947–958.
[35] J. Zhang, Y. Teng, W. Chen, Support vector regression with modified firefly algorithm for stock price forecasting, Appl. Intell. 49 (2019) 1658–1674.
[36] H. Markowitz, Portfolio selection, J. Financ. 7 (1952) 77–91.
[37] R. Merton, Lifetime portfolio selection under uncertainty: The continuous-time case, Rev. Econ. Stat. 51 (3) (1969) 247–257.
[38] E.F. Fama, Multiperiod consumption-investment decisions, Amer. Econ. Rev. 60 (1) (1970) 163–174.
[39] W. Chen, D. Li, Y. Liu, A novel hybrid ICA-FA algorithm for multiperiod uncertain portfolio optimization model based on multiple criteria, IEEE Trans. Fuzzy Syst. 27 (5) (2019) 1023–1036.
[40] D. Roy, Safety-first and the holding of assets, Econometrica 20 (1952) 431–449.
[41] H. Markowitz, Portfolio Selection: Efficient Doversification of Investments, Wiley, 1959.
[42] H. Konno, H. Yamazaki, Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market, Manage. Sci. 37 (1991) 519–531.
[43] M. Speranza, Linear programming models for portfolio optimization, Finance 12 (1993) 107–123.
[44] W. Chen, Y. Wang, S. Lu, M. Mehlawat, A hybrid FA-SA algorithm for fuzzy portfolio selection with transaction costs, Ann. Oper. Res. 269 (1–2) (2018) 129–147.
[45] W. Chen, Y. Wang, P. Gupta, M. Mehlawat, A novel hybrid heuristic algorithm for a new uncertain mean-variance-skewness portfolio selection model with real constraints, Appl. Intell. 48 (9) (2018) 2996–3018.
[46] J. Zhou, X. Li, Mean-semi-entropy portfolio adjusting model with transaction costs, J. Data Inf. Manag. 2 (2020) 121–130.
[47] M. Akbay, C. Kalayci, O. Polat, A parallel variable neighborhood search algorithm with quadratic programming for cardinality constrained portfolio optimization, Knowl.-Based Syst. 198 (2020) 105944.
[48] R. Mansini, W. Ogryczak, M. Speranza, Twenty years of linear programming based portfolio optimization, European J. Oper. Res. 234 (2014) 518–535.
[49] M. Masmoudi, F. Abdelaziz, Portfolio selection problem: A review of deterministic and stochastic multiple objective programming models, Ann. Oper. Res. 267 (2018) 335–352.
[50] F. Paiva, R. Cardoso, G. Hanaoka, W. Duarte, Decision-making for financial trading: A fusion approach of machine learning and portfolio selection, Expert Syst. Appl. 115 (2019) 635–655.
[51] M. Thenmozhi, G. Chand, Forecasting stock returns based on information transmission across global markets using support vector machines, Neural Comput. Appl. 27 (4) (2016) 805–824.
[52] Y. Guo, S. Han, C. Shen, Y. Li, X. Yin, Y. Bai, An adaptive SVR for high-frequency stock price forecasting, IEEE Access 99 (6) (2018) 11397–11404.
[53] X. Li, Y. Sun, Stock intelligent investment strategy based on support vector machine parameter optimization algorithm, Neural Comput. Appl. 32 (6) (2020) 1765–1775.
[54] A. Jaafari, S. Termeh, D. Bui, Genetic and firefly metaheuristic algorithms for an optimized neuro-fuzzy prediction modeling of wildfire probability, J. Environ. Manag. 243 (2019) 358–369.
[55] T. Chen, C. Guestrin, XGBoost:A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[56] L. Lv, J. Zhao, The firefly algorithm with Gaussian disturbance and local search, J. Sign. Process. Syst. 90 (8–9) (2018) 1123–1131.
[57] A. Altan, S. Karasu, Recognition of COVID-19 disease from X-ray images by hybrid model consisting of 2D curvelet transform, chaotic salp swarm algorithm and deep learning technique, Chaos Solitons Fractals 140 (2020) 110071.
[58] D. Oliva, M. Elaziz, An improved brainstorm optimization using chaotic opposite-based learning with disruption operator for global optimization and feature selection, Soft Comput. 24 (18) (2020) 14051–14072.
[59] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of ICNN'95 - International Conference on Neural Networks, 1995, pp. 1942–1948.
[60] T. Chang, S. Yang, K. Chang, Portfolio optimization problems in different risk measures using genetic algorithm, Expert Syst. Appl. 36 (7) (2009) 10529–10537.
[61] B. Chen, J. Zhong, Y. Chen, A hybrid approach for portfolio selection with higher-order moments: Empirical evidence from Shanghai Stock Exchange, Expert Syst. Appl. 145 (2020) 113104.
[62] T. Fischer, C. Krauss, Deep learning with long short-term memory networks for financial market predictions, European J. Oper. Res. 270 (2) (2018) 654–669.
[63] Sayed, I. Gehad, Khoriba, Ghada, Haggag, H. Mohamed, A novel chaotic salp swarm algorithm for global optimization and feature selection, Appl. Intell. 48 (10) (2018) 3462–3481.
[64] E. Rashedi, N. Hossein, S. Saryazdi, GSA: A gravitational search algorithm, Inform. Sci. 179 (13) (2009) 2232–2248.
[65] X. Li, Z. Shao, J. Qian, An optimizing method based on autonomous animate: Fish swarm algorithm, Syst. Eng. Theory Pract. 22 (11) (2002) 32–38.
[66] M. Chen, G. Zeng, K. Lu, J. Wang, A two-layer nonlinear combination method for short-term wind speed prediction based on ELM, ENN, and LSTM, IEEE Internet Things J. 6 (4) (2019) 6997–7010.