Hybrid Robust Portfolio Selection Model Using Machine Learning-based Preselection

Tingting Hai*, Liangyu Min, *Member*, *IAENG*

***Abstract*—Robust portfolio optimization theory is an essential foundation for modern financial modeling, which is a well-studied but not fully conquered territory. Conservatism is one of the most discussed issues by numerous scholars. To obtain a robust portfolio model with satisfactory performance, we propose the hybrid robust mean-variance portfolio model constrained with different ellipsoidal uncertainty sets in this paper. Additionally, skewness is also considered in the objective function. Preselection is designed for picking out the high-quality risky assets, where two machine learning algorithms, Random Forest and Support Vector Machine, are involved. In the numerical experiments, the US 48 industry data set from Kenneth R. French is employed to verify the effectiveness of the proposed hybrid portfolio models. The comparative results between the proposed hybrid models and baseline portfolio models (equal-weighted model, mean-variance model, mean-variance-skewness model) show that the proposed hybrid robust mean-variance portfolios considering skewness with preselection beat the baseline strategies by a clear margin. Also, the actual effectiveness of skewness in the hybrid robust models is analyzed.**

***Index Terms*—Portfolio selection, Hybrid robust, Skewness, Machine learning**

I. Introduction

The seminal mean-variance (MV) portfolio selection model proposed by Markowitz [1] laid the essential foundation for the modern portfolio theory (MPT). However, some shortcomings of the MV portfolio selection model have been discussed by scholars [2], [3], [4]. To overcome the well-known parameter sensitivity of the classical MV model, robust programming is introduced into the portfolio formation. The pioneering research proposed by Ben-Tal & Nemirovski (1998) [5] considers building a robust convex optimization model even though the inputted data is uncertain to some extent. Goldfarb & Iyengar (2003) [6] constructed the robust portfolio selection model based on [5]. In their work, Fama-French factors [7] are used to estimate the covariance matrix, and the second-order cone (SOCP) form of robust portfolio model is given. Considering the worst-case scenario is a widely-used approach for solving robust optimization models, and the effectiveness of such method on out-of-sample data sets has been fully discussed and verified [6], [8], [9].

However, conservatism is an inevitable issue in robust optimization [10], [11], [12]. Lotfi et al (2017) [13] proposed to consider the best-case counterpart and the worst-case counterpart in portfolio formation. In their work, Value-at-risk (VaR) is the risk measure focused on, and the hybrid robust mean-VaR portfolio model is built. Inspired by [13], [14], this paper contributes to derive and construct the hybrid robust mean-variance portfolio models constrained with different ellipsoidal uncertainty sets. Skewness is also considered in investment decision-making by some scholars and practitioners [15], [16], [17], [18], [19], [20]. Because a portfolio with a higher skewness would result in satisfactory performance, which is more appealing to rational investors. Hence, following the existing researches, skewness is encompassed in the proposed hybrid robust mean-variance portfolio models to improve the performance.

Numerous studies have demonstrated that high-quality risky assets could bring about efficient and effective portfolio models [21], [22], [23]. Preselection is the procedure to pick out the eligible risky assets based on the customized rules. Accordingly, accurate forecasting information is important in preselection. Existing literature shows that some artificial intelligence techniques are beneficial to build the feasible preselection process. Wang et al (2019) [22] constructed the risky assets preselection with Long Short Term Memory (LSTM) networks. Chen et al (2021) [21] designed eXtreme Gradient Boosting (XGBoost) with an improved firefly algorithm to predict stock prices in preselection. To improve the performance of the proposed portfolio models, we develop a hybrid algorithm involving the forecasting results provided by Random Forest and Support Vector Machine regression.

The key innovations of this paper are as follows. Firstly, the hybrid robust mean-variance portfolio models constrained with different ellipsoidal uncertainty sets are derived and constructed, where skewness is also taken into account. Secondly, preselection combining the forecasting results provided by two machine learning algorithms are proposed, where a hybrid algorithm is developed to select the appropriate risky assets. Finally, comparative numerical experiments are implemented, in which detailed indicators of portfolio performance are presented. We also investigate the actual effectiveness of skewness in the hybrid robust portfolio models and provide the corresponding analysis for individual investors.

The rest of this paper is organized as follows. Confidence ellipsoids U_δ^1 and U_δ^2 are introduced in Section II. Section III presents the construction of the hybrid robust mean-variance portfolio models involving skewness. The preselection and the hybrid algorithm for selecting risky

Manuscript received June 29, 2021; revised Sep. 29, 2021. This work is supported by the National Natural Science Foundation of China (No. 71971132), the Graduate Innovation Fund in SUFE (Nos. CXJJ-2018-401, CXJJ-2019-325, CXJJ-2019-342), and the Fundamental Research Funds for the Central Universities (No. 2020110552).

Tingting Hai is a PhD candidate in the School of Finance, Shanghai University of Finance and Economics, Shanghai 200433, China (Corresponding author, e-mail: hai_tingting@163.com).

Liangyu Min is a PhD candidate in the School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China (e-mail: minux@163.sufe.edu.cn).

assets is illustrated in Section IV. Numerical experiments with detailed analysis are shown in Section V. Conclusions and related analysis are revealed in Section VI.

II. Confidence ellipsoids

In this section, two confidence ellipsoids [13], [14], U_δ^1 and U_δ^2, are introduced for the follow-up hybrid portfolio modeling. Assuming the risky assets returns follows a joint normal distribution, where the estimate of mean r can be obtained from a group of i.i.d samples of size S for the n risky assets.

A. Confidence ellipsoid U_δ^1

In U_δ^1, the covariance matrix of the n risky assets is assumed to be fixed, and we have:

$$\frac{S(S-n)}{(S-1)n}(r-\bar{r})'\Sigma^{-1}(r-\bar{r}) \sim \chi_n^2 \tag{1}$$

The associated confidence ellipsoid around the point estimator $\hat{r}$ is as follows:

$$U_\delta^1 = \{r \in \mathbb{R}^n \mid S(r-\hat{r})'\Sigma^{-1}(r-\hat{r}) \leq \delta^2\} \tag{2}$$

where $\chi_n^2(\delta^2) = \theta$ with $\theta \in (0,1)$ is the chosen confidence, and $\mathbb{P}(r \in U_\delta^1) = \theta$.

B. Confidence ellipsoid U_δ^2

In essence, U_δ^2 is an extension of the ellipsoidal uncertainty set U_δ^1. In U_δ^2, a joint uncertainty set for the pair (r, Σ) is considered, and the distributions for the independent sample estimators, $\hat{r}$ and $\hat{\Sigma}$ are as follows:

$$\begin{cases} \hat{r} \sim \mathcal{N}(r, \dfrac{\Sigma}{S}) \\ \hat{r} \sim \mathcal{W}(\dfrac{\Sigma}{S-1}, S-1) \end{cases} \tag{3}$$

where $\hat{r} = \frac{1}{S}\sum_{i=1}^{S} r_i, \hat{\Sigma} = \frac{1}{S-1}\sum_{i=1}^{S}(r_i - \hat{r})(r_i - \hat{r})'$, and $\mathcal{N}(\mu, \sigma^2)$ represents the Gaussian distribution, $\mathcal{W}(G, \nu)$ denotes the Wishart distribution with scale matrix G and degree of freedom ν. An appropriate joint ellipsoidal uncertainty set can be derived according to the procedure of Schottle & Werner (2009) [24] as follows:

$$\begin{aligned} U_\delta^2 = &\{(r,\Sigma) \in \mathbb{R}^n \times \mathbb{S}^n \mid S(r-\hat{r})\hat{\Sigma}^{-1}(r-\hat{r}) + \\ &\frac{S-1}{2}\|\hat{\Sigma}^{-1/2}(\Sigma - \hat{\Sigma})\hat{\Sigma}^{-1/2}\|_F^2 \leq \delta^2\} \end{aligned} \tag{4}$$

where $\|A\|_F^2 = tr(AA')$.

III. Hybrid robust mean-variance model with skewness

The hybrid robust portfolio selection models constrained with the two ellipsoidal uncertainty sets introduced in section II are proposed in this section. Existing literature [17], [20], [16], [15], [18], [19] points out that a portfolio with a larger skewness is more appealing to the investors. As a result, skewness is integrated into the corresponding portfolio objective functions to improve the overall performance. We assume that no short-selling is allowed in our proposed models, that is, the feasible set of the portfolio weight is $\mathbb{X} = \{x \in \mathbb{R}^n \mid x \geq 0, \|x\|_1 = 1\}$.

A. Ellipsoidal uncertainty set U_δ^1

The best-case counterpart of the robust mean-variance model constrained with U_δ^1 is as follows:

$$\min_{x\in\mathbb{X}} \min_{r\in U_\delta^1} -r'x + \lambda\|\Sigma^{1/2}x\| \tag{5}$$

and the worst-case counterpart is as follows:

$$\min_{x\in\mathbb{X}} \max_{r\in U_\delta^1} -r'x + \lambda\|\Sigma^{1/2}x\| \tag{6}$$

where λ is the trade-off parameter between return and risk. Following the procedure of Lotfi et al. [13], [14], the associated hybrid robust mean-variance models can be derived by introducing a trade-off parameter β, which describes the possible optimistic level of the potential market conditions.

$$\begin{aligned} &\min_{x\in\mathbb{X}} \Bigg[\beta\bigg(-\hat{r}'x - \frac{\delta}{\sqrt{S}}\|\hat{\Sigma}^{1/2}x\| + \lambda\|\hat{\Sigma}^{1/2}x\|\bigg) + \\ &(1-\beta)\bigg(-\hat{r}'x + \frac{\delta}{\sqrt{S}}\|\hat{\Sigma}^{1/2}x\| + \lambda\|\hat{\Sigma}^{1/2}x\|\bigg)\Bigg] \\ = &\min_{x\in\mathbb{X}} -\hat{r}'x + \lambda\|\hat{\Sigma}x\| + \bigg((1-2\beta)\frac{\delta}{\sqrt{S}}\bigg)\|\hat{\Sigma}^{1/2}x\| \end{aligned} \tag{7}$$

Hence, we propose the multi-objective portfolio model constrained with the U_δ^1 as follows:

$$\begin{cases} \min\limits_{x\in\mathbb{X}} & -\hat{r}'x + \lambda\|\hat{\Sigma}^{1/2}x\| + \bigg((1-2\beta)\dfrac{\delta}{\sqrt{S}}\bigg)\|\hat{\Sigma}x\| \\ \max\limits_{x\in\mathbb{X}} & Skew(x) = \mathbb{E}(x'(r-\bar{r})^3) = x'M_3(x\otimes x) \end{cases} \tag{8}$$

where M_3 is the co-skewness matrix, $\otimes$ denotes the kronecker product. To solve the problem (8), we can transform it into the following nonlinear programming using the linear weighted method.

$$\begin{aligned} \min_{x\in\mathbb{X}} &- \hat{r}'x + \lambda\|\hat{\Sigma}^{1/2}x\| + \bigg((1-2\beta)\frac{\delta}{\sqrt{S}}\bigg)\|\hat{\Sigma}^{1/2}x\| \\ &- \gamma x'M_3(x\otimes x) \end{aligned} \tag{9}$$

where $\gamma > 0$ is the risk preference parameter.

B. Ellipsoidal uncertainty set U_δ^2

In the joint ellipsoidal uncertainty set U_δ^2, the best-case counterpart is as follows:

$$\min_{x\in\mathbb{X}} \min_{(r,\Sigma)\in U_\delta^2} -r'x + \lambda\|\Sigma^{1/2}x\| \tag{10}$$

and the worst-case counterpart is as follows:

$$\min_{x\in\mathbb{X}} \max_{(r,\Sigma)\in U_\delta^2} -r'x + \lambda\|\Sigma^{1/2}x\| \tag{11}$$

Similar to the procedure in the case of U_δ^1, the hybrid robust mean-variance model constrained with U_δ^2 can be obtained as follows:

$$\begin{aligned} &\min_{x\in\mathbb{X}} \Bigg[\beta\bigg(\min_{(r,\Sigma)\in U_\delta^2} -r'x + \lambda\|\Sigma^{1/2}x\|\bigg) + \\ &(1-\beta)\bigg(\max_{(r,\Sigma)\in U_\delta^2} -r'x + \lambda\|\Sigma^{1/2}x\|\bigg)\Bigg] \\ = &\min_{x\in\mathbb{X}} -\hat{r}'x + \bigg(\beta \min_{k\in[\max(0,1-\frac{S-1}{2\delta^2}),1]} g_2(k) + \\ &(1-\beta)\max_{k\in[0,1]} g_1(k)\bigg)\|\hat{\Sigma}^{1/2}x\| \end{aligned} \tag{12}$$

where

$$g_1(k) = \delta\sqrt{\frac{k}{S}} + \lambda\sqrt{\left(1 + \delta\sqrt{\frac{2(1+k)}{S-1}}\right)}$$

$$g_2(k) = -\delta\sqrt{\frac{k}{S}} + \lambda\sqrt{\left(1 - \delta\sqrt{\frac{2(1-k)}{S-1}}\right)}$$

Accordingly, the multi-objective portfolio model constrained with the ellipsoidal uncertainty set U_δ^2 is as follows:

$$\begin{cases} \min\limits_{x\in\mathbb{X}} & -\hat{r}'x + \left(\beta \min\limits_{k\in[\max(0,1-\frac{S-1}{2\delta^2})],1} g_2(k) + \right. \\ & \left. (1-\beta)\max\limits_{k\in[0,1]} g_1(k)\right)\|\hat{\Sigma}^{1/2}x\| \\ \max\limits_{x\in\mathbb{X}} & Skew(x) = \mathbb{E}(x'(r-\bar{r})^3) = x'M_3(x\otimes x) \end{cases} \tag{13}$$

Also, the corresponding nonlinear programming form is as follows:

$$\min_{x\in\mathbb{X}} -\hat{r}'x + \left(\beta \min_{k\in[\max(0,1-\frac{S-1}{2\delta^2}),1]} g_2(k) + \right. \\ \left. (1-\beta)\max_{k\in[0,1]} g_1(k)\right)\|\hat{\Sigma}^{1/2}x\| - \gamma x'M_3(x\otimes x) \tag{14}$$

IV. Preselection by machine learning algorithms

Many scholars have demonstrated that the machine learning-based preselection could improve the performance of portfolio models [23], [22], [21]. Two machine learning algorithms within the scope of this work are introduced in this section. Firstly, the basic principle of Random Forest and Support Vector Machine are briefly presented and explained. Following that, we propose a hybrid algorithm to select risky assets based on the forecasting results provided by the two machine learning algorithms.

A. Random Forest

Random Forest (RF) [25] is one of the ensemble learning algorithms which has been widely used to handle classification or regression tasks. In RF, n base estimators are constructed to generate forecasting results independently. Bagging is used to effectively synthesize these outputs from the base estimators. Suppose that there are M features in the data set, then at most $k = \sqrt{M}$ features would be randomly selected by a base estimator to learn the potential data pattern. Information gain is a frequently-used criterion for the tree node splitting as follows:

$$Gain(D, k_j) = Entropy(D) - \sum_i \frac{|D_i|}{D} Entropy(D_i)$$

where D is the parent node, D_i is the children nodes after splitting, k_j is the feature to be split. Upon the leaves reach the threshold of defined impurity, Random Forest would obtain the final result $R(x)$ according to the following voting equation:

$$R(x) = \arg\max \sum_{i=1}^{n} \mathbf{I}(r_i(x) = R(x))$$

where $r_i(x)$ represents the output of the ith base estimator, and $\mathbf{I}(\cdot)$ is the indicator function.

B. SVM

Support Vector Machine (SVM) is a supervised learning algorithm orginating from the statistical learning theory [26]. Considering the possibility of sample scarcity, SLT substitutes the conventional expected risk minimization with empirical risk minimization (ERM). Based on the rule of ERM, however, the learning model would be over-fitted when the complexity of the model is high while the number of samples is limited [27]. Structure Risk Minimization (SRM) is used in SVM to avoid such a dilemma. SRM tries to approach to the true risk by empirical risk as well as an associated confidence interval. According to the rule of SRM, SVM could obtain the learning model with appropriate complexity even when the number of samples is limited [28], [29], [30].

Similar to random forest, SVM is also applicable to deal with classification and regression problems. In this paper, SVM regression (SVR) is used to give forecasting results for the follow-up hybrid algorithm. Assuming that we have samples $x_i \in \mathbb{R}^m, i = 1, 2, \ldots, n$ and labels $y \in \mathbb{R}^n$ for training. The goal of SVR is find $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that minimizing the regression error. Setting the upper bound of deviation from the true label is ϵ, the $\epsilon-$SVR [31], [32] solves the following optimization problem:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}w'w + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \\ s.t. \begin{cases} y_i - w'\phi(x_i) - b \le \epsilon + \xi_i \\ w'\phi(x_i) + b - y_i \le \epsilon + \xi_i^* \\ \xi, \xi^* \ge 0, i = 1, 2, \ldots, n \end{cases} \tag{15}$$

where n is the number of observations, m is the dimension of feature space, C is the penalty term, ξ and ξ^* represents the predictions lie above and below the ϵ tube, respectively, ϕ is the function to map the training vectors into a high-dimensional space. Considering the primal optimization problem (15) is hard to solve in some cases, the dual form of problem (15) is derived as follows:

$$\min_{\alpha,\alpha^*} \frac{1}{2}(\alpha-\alpha^*)'K(\alpha-\alpha^*) + \epsilon\mathbf{1}'(\alpha+\alpha^*) - y'(\alpha-\alpha^*) \\ s.t. \begin{cases} \mathbf{1}'(\alpha-\alpha^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le C, i = 1, 2, \ldots, n \end{cases} \tag{16}$$

where α and α^* are the corresponding dual variables, $\mathbf{1}$ is the vector of all ones, K is an $n \times n$ positive semi-definite matrix with $K_{ij} = \phi(x_i)'\phi(x_j)$.

C. Proposed hybrid algorithm for selecting assets

Preselection is a crucial step in investment decision-making when the individual investor has many assets to manage. Moreover, high-quality risky assets would be beneficial to the optimal portfolio construction. To this end, we propose a hybrid algorithm based on the forecasting results provided by RF and SVR in this section.

The fundamental features used in the machine learning algorithms are Fama-French five factors, which demonstrated to explain more than 70% of the associated market

TABLE I
FUNDAMENTAL FACTORS IN FAMA-FRENCH MODEL.

Factor	Formula	Details
Mkt-Rf	$r_M - r_f$	the excess return rate of market, market risk.
SMB	$r_S - r_B$	the return spread of small minus large stocks, size risk.
HML	$r_H - r_L$	the return spread of cheap minus expensive stocks, value risk.
RMW	$r_R - r_W$	the return spread of the most profitable firms minus the least profitable.
CMA	$r_C - r_A$	the return spread of firms that invest conservatively minus aggressively.

return [7], [33], [34], [35]. The effectiveness of the Fama-French factors are also illustrated in [36], [37], [38]. Table I summarizes the fundamental five factors in Fama-French model.

In the numerical experiments, Random Forest and SVR would give the forecasting returns based on Fama-French factors. In order to select high-quality risky assets according to the predictions, we design the following hybrid algorithm. Firstly, we build a heap Q to sort the forecasting results and define k to represent the number of risk assets to be selected. Secondly, we set a flexible number $K > k$, and the top-K results from RF and SVR respectively are selected by Q, the indexes of these selected risky assets are also recorded. Finally, the intersection of the two selected results sets are calculated, and the final risky assets set can be obtained according to the recorded indexes. Algorithm 1 illustrates the details of the hybrid strategy.

Algorithm 1 Preselection of risky assets.

```
Input: Data set containing Fama-French five factors $D$; Risky assets set $A$; Size of risky assets set $N$; Flexible number $K$; Size of selected risky assets $k$; Random Forest; SVR;
Output: Selected risky assets $A_{sel}$
 1: Split $D$ into two parts, $D_t$ for training and $D_p$ for prediction;
 2: Train RF and SVR based on $D_t$, then use the well-trained machine learning models to predict the returns based on $D_p$. The predicted results are $P_s$ and $P_r$, respectively.
 3: Build a heap $Q$ with size $K$, for sorting $P_s$ and $P_r$. The sorted results are $Q_s$ and $Q_r$, respectively. Record the indexes of the selected risky assets in a set $I$.
 4: Calculate the intersection of $Q_s$ and $Q_r$, and the obtained set is $R_{sel}$ with size of $k'$.
 5: while $R_{sel}$ is $\emptyset$ do
 6:     $K = \max(2 * K, N)$ and turn to the step 3.
 7: end while
 8: if $k' < k$ then
 9:     $K = K + \varepsilon$, where $\varepsilon$ is a small number.
10:     Turn to the step 3.
11: else if $k' > k$ then
12:     Drop the last values of $R_{sel}$ until $k' = k$.
13: else
14:     Based on the indexes set $I$ and risky assets set $A$, the selected risky assets $A_{sel}$ can be obtained by the operation of mapping.
15: end if
16: return $A_{sel}$;
```

V. NUMERICAL EXPERIMENTS

To verify the effectiveness of the proposed hybrid robust portfolio models, the numerical experiments are designed and implemented in this section. The hyper-parameters of the proposed hybrid robust portfolio models are as follows. $\lambda = 0.5, \beta = 0.3$, and $\gamma = 2$. More details of the hyper-parameters setting can be referred to [20]. Fig 1 presents the flowchart of the designed numerical experiments. Table II summarizes the portfolio models in the numerical experiments.

A. Data set

The US 48 industry portfolio daily data set from Kenneth R. French is employed in the empirical research, among which the data from June 1, 2015, to Feb. 28, 2018, is divided into the training set (total 693 observations), and the data from June 1, 2018, to May 31, 2019, is divided into the testing set (total 251 observations). In the training set, the preselection is carried out and the proposed hybrid robust portfolio models and baseline portfolios are constructed, whereas both the proposed models and benchmarks are verified in the testing set.

B. Preselection

As shown in Fig. 1, preselection is set to sort out the potential high-quality risky assets. To investigate the accuracy of the machine learning algorithms used in preselection, we define the following indicators as [40], [41], [21] for comparison:

$$\begin{aligned} \text{MSE} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ \text{MAE} &= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \end{aligned} \tag{17}$$

where MSE is short for mean square error, MAE is short for mean absolute error, y_i represents the true value, $\hat{y}_i$ represents the predicted value, n is the number of observations.

Fig. 2 shows the MAE and MSE of Random Forest and SVR, respectively. The X-axis represents the 48 industries in data set, the left Y-axis indicates the MAE while the right Y-axis is the MSE; the red bar is the MAE of RF, the green bar is the MAE of SVR, the blue line represents the MSE of RF, and the grey line represents the MSE of SVR. It can be observed that the accuracy of Random Forest is slightly higher than SVR, which again illustrates the advantage of the ensemble learning algorithm. Also, the rationality of the proposed hybrid method for selecting risky assets can be demonstrated in Fig. 2 to some extent.

Table III presents the descriptive statistics of the selected industries by Algorithm 1. Overall, the industry data is rather stable, which could avoid some extreme cases in portfolio validation. As a result, the performance on

TABLE II
SUMMARY OF THE PORTFOLIO MODELS IN NUMERICAL EXPERIMENTS.

Model	Expected return	Variance	Skewness	Hybrid robust	Uncertainty set	Detail
EWM	✗	✗	✗	✗	✗	Equal weighted portfolio model [39], baseline model
MV	✓	✓	✗	✗	✗	Conventional Markowitz model, baseline model
MVS	✓	✓	✓	✗	✗	Mean-variance-skewness model [15], baseline model
HMVu1	✓	✓	✗	✓	U_δ^1	Hybrid robust MV model constrained with U_δ^1, without preselection
HMVu2	✓	✓	✗	✓	U_δ^2	Hybrid robust MV model constrained with U_δ^2, without preselection
HMVSu1U	✓	✓	✗	✓	U_δ^1	Hybrid robust MV model constrained with U_δ^1, considering skewness, without preselection.
HMVSu2U	✓	✓	✗	✓	U_δ^2	Hybrid robust MV model constrained with U_δ^2, considering skewness, without preselection.
HMVSu1	✓	✓	✗	✓	U_δ^1	Hybrid robust MV model constrained with U_δ^1, considering skewness, with preselection.
HMVSu2	✓	✓	✗	✓	U_δ^2	Hybrid robust MV model constrained with U_δ^2, considering skewness, with preselection.

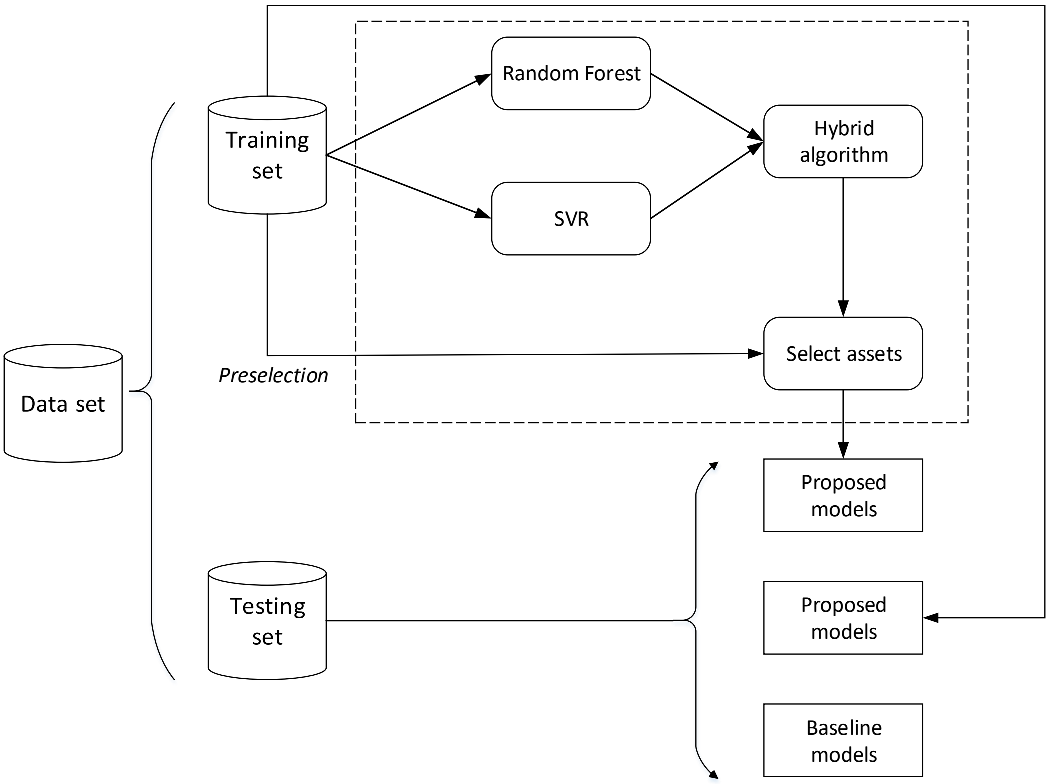

Fig. 1. Flowchart of numerical experiments.

the stable data set can demonstrate the universality of the proposed hybrid portfolio selection models.

C. Performance of portfolio models

To comprehensively evaluate the performance of the proposed portfolio models, some evaluation metrics are defined as follows:

(1). Return on Investment (ROI)

ROI is used to measure the efficiency and profitability of a portfolio model. The standard formula of ROI is as follows:

$$\mathrm{ROI} = \frac{\mathrm{FinalWealth} - \mathrm{InitialWealth}}{\mathrm{InitialWealth}} \times 100\% \qquad (18)$$

where FinalWealth represents the cumulative return obtained from the investment bought with InitialWealth.

(2). Annual Percentage Yield (APY)

APY is also an indicator evaluating the profitability of an investment, and can be calculated based on the following equation:

$$\mathrm{APY} = \sqrt[n]{1 + \mathrm{ROI}} - 1 \qquad (19)$$

TABLE III
DESCRIPTIVE STATISTICS OF THE SELECTED RISKY ASSETS.

No.	Industry	Min	25% Quantile	Median	Mean	75% Quantile	Max
1	BusSv	−0.0444	−0.0036	0.0009	0.0008	0.0060	0.0463
2	Soda	−0.0414	−0.0039	0.0005	0.0004	0.0053	0.0265
3	Beer	−0.0428	−0.0040	0.0007	0.0004	0.0052	0.0304
4	Smoke	−0.0470	−0.0044	0.0005	0.0005	0.0059	0.0422
5	Chips	−0.0438	−0.0050	0.0013	0.0007	0.0065	0.0510
6	LabEq	−0.0456	−0.0038	0.0010	0.0007	0.0064	0.0385
7	Paper	−0.0472	−0.0037	0.0009	0.0004	0.0052	0.0326
8	Rtail	−0.0390	−0.0039	0.0011	0.0006	0.0057	0.0399
9	MedEq	−0.0391	−0.0042	0.0012	0.0006	0.0060	0.0277
10	Mach	−0.0593	−0.0051	0.0007	0.0008	0.0072	0.0369
11	Aero	−0.0536	−0.0037	0.0012	0.0009	0.0065	0.0336
12	Guns	−0.0431	−0.0039	0.0009	0.0010	0.0059	0.0622
13	Gold	−0.1176	−0.0132	0.0012	0.0008	0.0147	0.1042
14	Util	−0.0393	−0.0045	0.0006	0.0002	0.0051	0.0289

* The acronyms of the industries can be referred to Kenneth R. French.

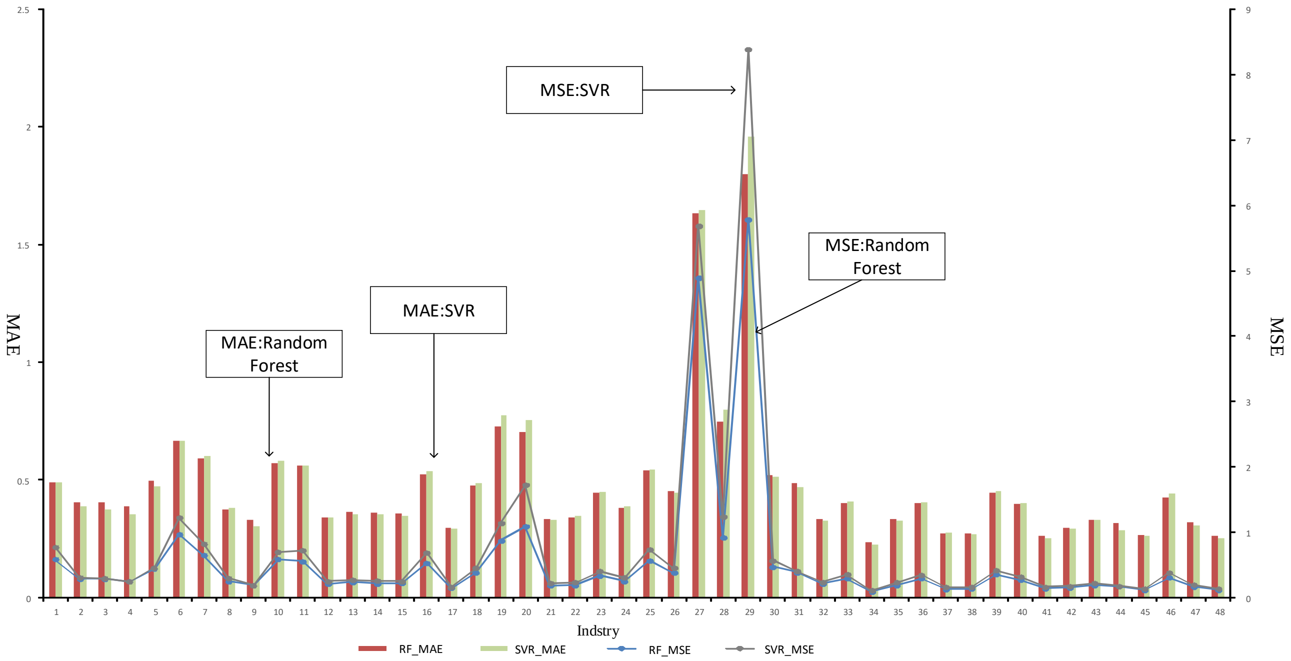

Fig. 2. MAE & MSE of RF and SVR

where n is the number of investment horizontal years.

(3). Sharpe ratio (SR)

SR [42] is a well-known risk-adjusted indicator that has been widely used in both academia and industry [43], [44], [45]. SR measures the excess return obtained while assuming a specific risk and can be computed as follows:

$$\mathrm{SR} = \frac{\mathrm{APY} - r_f}{\sigma_p} \tag{20}$$

where r_f is the annual risk-free rate, which is 3% in this work. σ_p is the annual standard deviation (STD) of the excess return.

(4). Maximum drawdown (MDD)

MDD is one of risk measures, which is equal to the upper bound decline from the peak to a through before a new peak is attained. Customarily, MDD is computed as a percentage of the peak value as follows:

$$\mathrm{MDD} = \max_{t\in[0,T]} \left\{ \frac{\max_{i\in[0,t]} \mathrm{ROI}_i - \mathrm{ROI}_t}{\max_{i\in[0,t]\mathrm{ROI}_i}} \right\} \tag{21}$$

Table IV reveals the performance of the proposed hybrid robust portfolio models and baseline strategies on the testing data set. With regard to ROI, HMVSu2 has the highest ROI of 0.1011, followed by HMVSu1 with 0.0969, HMVSu2U ranks third with 0.0701. All of the hybrid robust portfolio models have higher ROI than benchmarks (EWM, MV, MVS). HMVSu2 also achieves the highest APY of 0.1015, compared to 0.0973 for HMVSu1, 0.0704 for HMVSu2U, 0.0701 for HMVu2. It can be found that portfolios with preselection show better return characteristics, which is consistent with the rule of sifting risky assets.

In terms of risk, HMVSu2U and HMVu2 have the

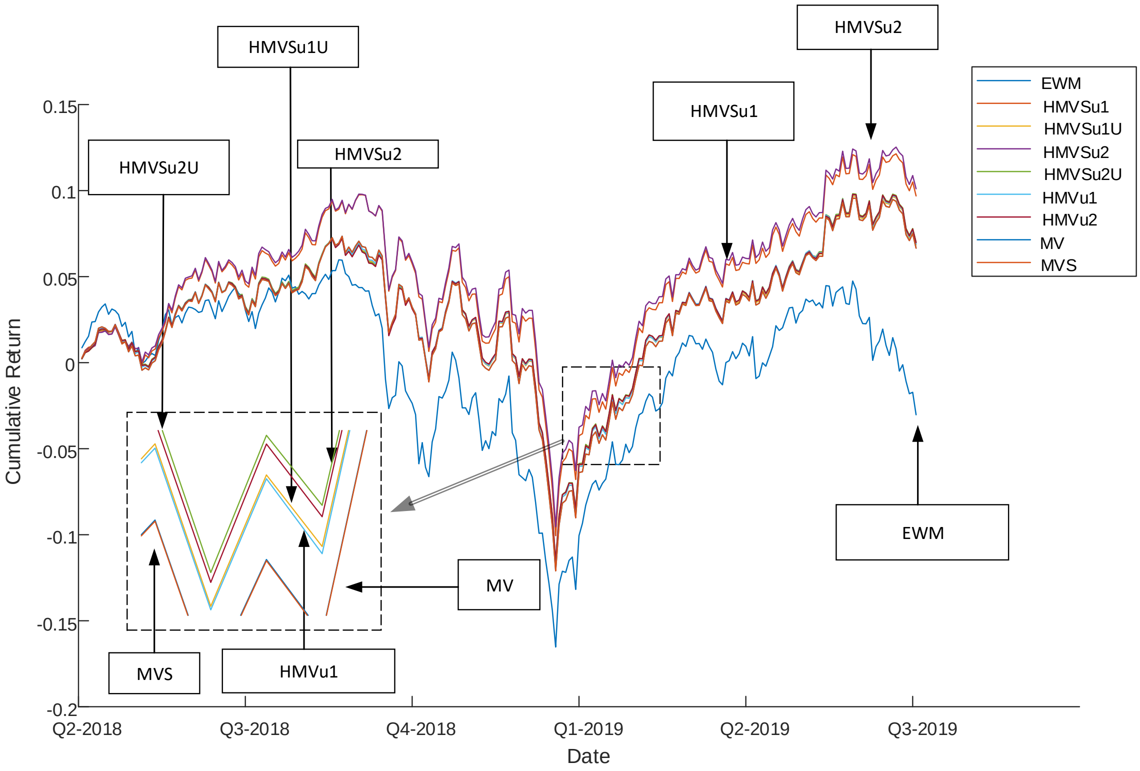

Fig. 3. Cumulative returns of portfolio models.

TABLE IV
PERFORMANCE OF THE PORTFOLIO MODELS.

	EWM	HMVSu1	HMVSu1U	HMVSu2	HMVSu2U	HMVu1	HMVu2	MV	MVS
ROI	-0.0303	0.0969	0.0685	**0.1011**	0.0701	0.0690	0.0698	0.0666	0.0665
APY	-0.0304	0.0973	0.0688	**0.1015**	0.0704	0.0693	0.0701	0.0669	0.0668
STD	0.1493	0.1289	0.1214	0.1272	**0.1208**	0.1219	**0.1208**	0.1238	0.1238
MDD	0.2124	0.1806	0.1767	0.1761	0.1752	0.1771	**0.1750**	0.1812	0.1812
SR	-0.4043	0.5223	0.3193	**0.5620**	0.3345	0.3220	0.3315	0.2978	0.2974
Skewness	**-0.0556**	-0.2645	-0.2142	-0.2675	-0.2165	-0.2154	-0.2138	-0.2205	-0.2203
VaR$(\%5)$	0.0174	0.0148	0.0126	0.0143	0.0125	0.0128	**0.0124**	0.0132	0.0132
$p-$Value*	—	0.0608	0.0967	**0.0574**	0.0944	0.0956	0.0963	0.0962	0.0963

* One-sided t-tests are implemented and the benchmark model is EWM.

lowest STD, 0.1208, HMVSu1U follows, 0.1214. The MDD HMVu2 also reaches the best level of 0.1750, HMVSu2U achieves the second best MDD of 0.1752, HMVSu2 follows with 0.1761. Due to the number of available risky assets being narrowed by the preselection process, portfolios with our designed preselection do not show as satisfying risk characteristics as the original hybrid portfolio models.

SR is an important indicator to evaluate portfolio models. In this regard, HMVSu2 has the highest SR of 0.5620, HMVSu1 ranks second with 0.5223. The two portfolio models' SR are significantly higher than others. Specifically, HMVSu2U has the third highest SR of 0.3345, HMVu2 follows, with 0.3315. The performance of SR demonstrates the effectiveness of the developed preselection process, by which risky assets with higher Sharpe ratios are pooled to form portfolio.

It can be observed from skewness that, HMVSu1U has higher skewness than HMVu1 (-0.2142 vs -0.2154), whereas HMVSu2U has lower skewness than HMVu2 (-0.2165 vs -0.2138), which means the actual effectiveness of skewness is related to the risk preference parameter for skewness, and this part is discussed in the next section.

VaR$(\%5)$ is presented to reflect the tail risk of different portfolio models. HMVu2 has the lowest VaR$(\%5)$ of 0.0124, HMVSu2U shows similar performance of 0.0125. HMVSu1U follows, with 0.0126. However, HMVSu2 and HMVSu1 shows higher VaR$(\%5)$, which are 0.0143 and 0.0148, respectively. It seems that the preselection process has positive influences on the portfolio VaR, but more

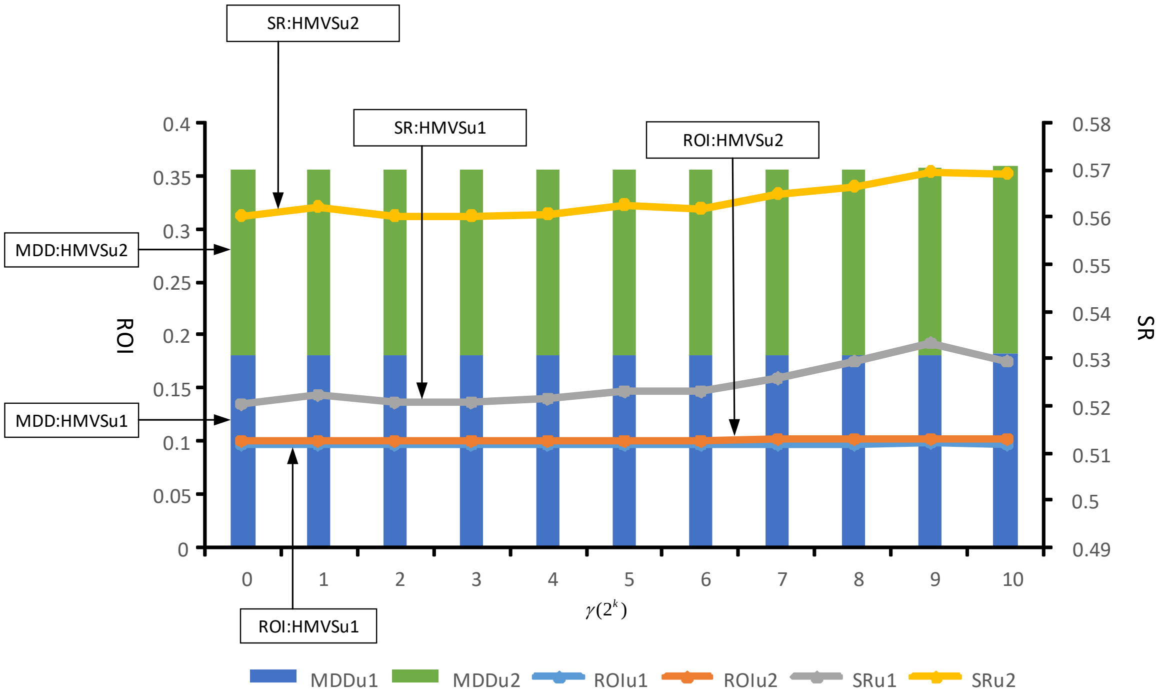

Fig. 4. Effect of the skewness in HMVSu1 and HMVSu2.

numerical experiments should be conducted to verify this conclusion.

One-sided t-tests are implemented to further show the superiority of the proposed hybrid portfolio models. The null hypothesis is the difference between the tested model and the benchmark is equal to zero, against the alternative that the difference is greater than zero. It can be found that HMVSu2 is the most significant portfolio model at the level 10%.

Fig. 3 visualizes the results in Table IV, which further presents the superiority of the proposed preselection process. Comparing the performance of HMVSu1 and HMVSu1U, as well as HMVSu2 and HMVSu2U, portfolio models with the customized preselection have better performance than those without this process, especially in terms of return. Models consider ellipsoidal uncertainty sets outperform the benchmarks regarding both return and risk. Specifically, portfolio models constrained with the ellipsoidal uncertainty U_δ^2 reveal more appealing performance than those in U_δ^1, which also illustrates the importance of an appropriate uncertainty set in robust modeling. However, the actual effect of skewness in the proposed hybrid robust portfolio models is not clear, detailed numerical experiments would be done in the next section for further analysis.

D. Analysis of the effect of skewness

Based on the numerical experiments above, the effect of skewness in the proposed hybrid robust portfolio models would be explored in this section. Specifically, the proposed portfolio models with better performance, HMVSu1 and HMVSu2, are chosen for the sensitive analysis.

Table V and Fig. 4 shows the effect of skewness on ROI, SR, MDD in HRMVSu1 and HMVSu2, respectively. The X-axis represents the coefficient of skewness (γ) in HMVSu1 and HMVSu2, where the logarithmic values based 2 are indicated, that is, the value of γ ranges from 2^0 to 2^{10}. The left Y-axis indicates the ROI of HMVSu1 and HMVS2, and the right Y-axis indicates the SR of HMVSu1 and HMVSu2. It can found that, SR is more sensitive to γ than ROI and MDD. When γ is less than 2^9, SR shows a general upward trend in both HMVSu1 and HMVSu2. However, ROI and MDD in the proposed hybrid robust portfolio models are not very sensitive to γ. The main reason for this result is that the hybrid robust portfolio models have taken the potential uncertainty of distribution of the returns into account, where the skewness of returns is also partly involved in the designed ellipsoidal uncertainty sets. In a nutshell, skewness is a useful objective when the portfolio model pursuing a better SR.

TABLE V
SENSITIVE ANALYSIS OF HMVSU1 & HMVSU2.

γ	ROI: HMVSu1	MDD: HMVSu1	SR: HMVSu1	ROI: HMVSu2	MDD: HMVSu2	SR: HMVSu2
2^0	0.0967	0.1806	0.5204	0.1008	0.1761	0.5600
2^1	0.0969	0.1806	0.5223	0.1011	0.1761	0.5620
2^2	0.0967	0.1806	0.5207	0.1008	0.1761	0.5602
2^3	0.0967	0.1806	0.5206	0.1008	0.1761	0.5600
2^4	0.0968	0.1806	0.5214	0.1009	0.1761	0.5605
2^5	0.0970	0.1806	0.5232	0.1011	0.1761	0.5624
2^6	0.0969	0.1805	0.5229	0.1010	0.1760	0.5619
2^7	0.0972	0.1806	0.5259	0.1013	0.1761	0.5648
2^8	0.0975	0.1807	0.5292	0.1014	0.1762	0.5666
2^9	0.0978	0.1810	0.5331	0.1016	0.1764	0.5696
2^{10}	0.0973	0.1823	0.5295	0.1014	0.1772	0.5693

VI. CONCLUSIONS & DISCUSSIONS

In this paper, we propose and construct the hybrid robust mean-variance portfolio models with skewness considered. Both the worst-case counterpart and the best-case counterpart are integrated into our models with a trade-off parameter β. From the comparative results provided by the designed numerical experiments, the proposed hybrid robust mean-variance portfolio models (HMVu1, HMVu2) outperform the conventional mean-variance and mean-variance-skewness portfolio models. Meanwhile, when the skewness is taken into account, the performance of hybrid robust portfolio models is further improved. Preselection plays a key role in improving model performance. Hybrid robust mean-variance portfolio models considering skewness outperform those without preselection by a clear margin. However, the only objective in the designed preselection is return, more objectives such volatility, tail risk, and risk-adjusted indicators can also be considered in the preselection, which would be implemented in our subsequential work.

In terms of the ellipsoidal uncertainty sets, the hybrid robust portfolio models constrained with U_δ^2 have better performance than those constrained with U_δ^1 in the experiments. As precedent scholars pointed out, an appropriate and feasible uncertainty set is vital for robust modeling. Some sophisticated algorithms should be developed for describing the uncertainty set.

Overall, the effectiveness of the hybrid robust portfolios with preselection has been demonstrated in this work, which also shows the feasibility of applying the artificial intelligence techniques to financial modeling.

ACKNOWLEDGMENT

Special thanks for the guidance from J.J. Gao, the associate professor in School of Information Management & Engineering, Shanghai University of Finance & Economics. Thanks also to Dewen Liu and Xiangxi Kong, Ph.D. students from College of Business, Shanghai University of Finance & Economics.

REFERENCES

[1] H. M. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, p. 77, 1952.

[2] M. J. Best and R. R. Grauer, "Sensitivity analysis for mean-variance portfolio problems," *Management Science*, vol. 37, no. 8, pp. 980–989, 1991.

[3] V. K. Chopra and W. T. Ziemba, "The effect of errors in means, variances, and covariances on optimal portfolio choice," *Journal of Portfolio Management*, vol. 19, no. 2, p. 6, 1993.

[4] D. Kuhn, P. Parpas, B. Rustem, and R. Fonseca, "Dynamic mean-variance portfolio analysis under model risk," *J. Comput. Finance*, vol. 12, no. 91115, p. 7, 2009.

[5] A. Ben-Tal and A. Nemirovski, "Robust convex optimization," *Mathematics of operations research*, vol. 23, no. 4, pp. 769–805, 1998.

[6] D. Goldfarb and G. Iyengar, "Robust portfolio selection problems," *Mathematics of operations research*, vol. 28, no. 1, pp. 1–38, 2003.

[7] E. F. Fama, "The behavior of stock-market prices," *The journal of Business*, vol. 38, no. 1, pp. 34–105, 1965.

[8] L. E. Ghaoui, M. Oks, and F. Oustry, "Worst-case value-at-risk and robust portfolio optimization: A conic programming approach," *Operations research*, vol. 51, no. 4, pp. 543–556, 2003.

[9] S. Zhu and M. Fukushima, "Worst-case conditional value-at-risk with application to robust portfolio management," *Operations research*, vol. 57, no. 5, pp. 1155–1168, 2009.

[10] A. Thiele, "A note on issues of over-conservatism in robust optimization with cost uncertainty," *Optimization*, vol. 59, no. 7, pp. 1033–1040, 2010.

[11] C. Shang, X. Huang, and F. You, "Data-driven robust optimization based on kernel learning," *Computers & Chemical Engineering*, vol. 106, pp. 464–479, 2017.

[12] E. Roos and D. den Hertog, "Reducing conservatism in robust optimization," *INFORMS Journal on Computing*, vol. 32, no. 4, pp. 1109–1127, 2020.

[13] S. Lotfi, M. Salahi, and F. Mehrdoust, "Adjusted robust mean-value-at-risk model: less conservative robust portfolios," *Optimization and Engineering*, vol. 18, no. 2, pp. 467–497, 2017.

[14] S. Lotfi and S. A. Zenios, "Robust var and cvar optimization under joint ambiguity in distributions, means, and covariances," *European Journal of Operational Research*, vol. 269, no. 2, pp. 556–576, 2018.

[15] H. Konno and K.-i. Suzuki, "A mean-variance-skewness portfolio optimization model," *Journal of the Operations Research Society of Japan*, vol. 38, no. 2, pp. 173–187, 1995.

[16] T. Joro and P. Na, "Portfolio performance evaluation in a mean–variance–skewness framework," *European Journal of Operational Research*, vol. 175, no. 1, pp. 446–461, 2006.

[17] W. Briec, K. Kerstens, and O. Jokung, "Mean-variance-skewness portfolio performance gauging: a general shortage function and dual approach," *Management science*, vol. 53, no. 1, pp. 135–149, 2007.

[18] X. Li, Z. Qin, and S. Kar, "Mean-variance-skewness model for portfolio selection with fuzzy returns," *European Journal of Operational Research*, vol. 202, no. 1, pp. 239–247, 2010.

[19] L. Yu, S. Wang, and K. K. Lai, "Neural network-based mean–variance–skewness model for portfolio selection," *Computers & Operations Research*, vol. 35, no. 1, pp. 34–46, 2008.

[20] B. Chen, J. Zhong, and Y. Chen, "A hybrid approach for portfolio selection with higher-order moments: Empirical evidence from shanghai stock exchange," *Expert Systems with Applications*, vol. 145, p. 113104, 2020.

[21] W. Chen, H. Zhang, M. K. Mehlawat, and L. Jia, "Mean–variance portfolio optimization using machine learning-based stock price prediction," *Applied Soft Computing*, vol. 100, p. 106943, 2021.

[22] W. Wang, W. Li, N. Zhang, and K. Liu, "Portfolio formation with preselection using deep learning from long-term financial data," *Expert Systems with Applications*, vol. 143, p. 113042, 2020.

[23] F. D. Paiva, R. T. N. Cardoso, G. P. Hanaoka, and W. M. Duarte, "Decision-making for financial trading: A fusion approach of machine learning and portfolio selection," *Expert Systems with Applications*, vol. 115, pp. 635–655, 2019.

[24] K. Schöttle and R. Werner, "Robustness properties of mean-variance portfolios," *Optimization*, vol. 58, no. 6, pp. 641–663, 2009.

[25] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[26] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[27] X. Zhang, "Introduction to statistical leanring theory and support vector machines," *Acta Automatica Sinica*, vol. 26, no. 1, pp. 32–42, 2000.

[28] L. Min, J. Dong, D. Liu, and X. Kong, "A black-litterman portfolio selection model with investor opinions generating from machine learning algorithms." *Engineering Letters*, vol. 29, no. 2, pp. 710–721, 2021.

[29] W. L. Al-Yaseen, "Improving intrusion detection system by developing feature selection model based on firefly algorithm and support vector machine," *IAENG Int. J. Comput. Sci*, vol. 46, no. 4, pp. 534–540, 2019.

[30] R. E. Caraka, R. C. Chen, S. A. Bakar, M. Tahmid, T. Toharudin, B. Pardamean, and S.-W. Huang, "Employing best input svr robust lost function with nature-inspired metaheuristics in wind speed energy forecasting," *IAENG Int. J. Comput. Sci*, vol. 47, no. 3, pp. 572–584, 2020.

[31] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[33] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of financial economics*, vol. 33, no. 1, pp. 3–56, 1993.

[34] ——, "Size, value, and momentum in international stock returns," *Journal of financial economics*, vol. 105, no. 3, pp. 457–472, 2012.

[35] Q. Lin, "Noisy prices and the fama–french five-factor asset pricing model in china," *Emerging Markets Review*, vol. 31, pp. 141–163, 2017.

[36] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *The Review of Financial Studies*, vol. 33, no. 5, pp. 2223–2273, 2020.

[37] O. Paliienko, S. Naumenkova, and S. Mishchenko, "An empirical investigation of the fama-french five-factor model," *Investment Management & Financial Innovations*, vol. 17, no. 1, p. 143, 2020.

[38] D. Horváth and Y.-L. Wang, "The examination of fama-french model during the covid-19," *Finance Research Letters*, p. 101848, 2020.

[39] V. DeMiguel, L. Garlappi, and R. Uppal, "Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy?" *The review of Financial studies*, vol. 22, no. 5, pp. 1915–1953, 2009.

[40] L.-L. Li, X. Zhao, M.-L. Tseng, and R. R. Tan, "Short-term wind power forecasting based on support vector machine with improved dragonfly algorithm," *Journal of Cleaner Production*, vol. 242, p. 118447, 2020.

[41] J. Zhang, Y.-F. Teng, and W. Chen, "Support vector regression with modified firefly algorithm for stock price forecasting," *Applied Intelligence*, vol. 49, no. 5, pp. 1658–1674, 2019.

[42] W. F. Sharpe, "The sharpe ratio," *Journal of portfolio management*, vol. 21, no. 1, pp. 49–58, 1994.

[43] D. Vukovic, Y. Vyklyuk, N. Matsiuk, and M. Maiti, "Neural network forecasting in prediction sharpe ratio: Evidence from eu debt market," *Physica A: Statistical Mechanics and its Applications*, vol. 542, p. 123331, 2020.

[44] C. D. Wang, Z. Chen, Y. Lian, and M. Chen, "Asset selection based on high frequency sharpe ratio," *Journal of Econometrics*, 2020.

[45] D. Chakrabarti, "Parameter-free robust optimization for the maximum-sharpe portfolio problem," *European Journal of Operational Research*, vol. 293, no. 1, pp. 388–399, 2021.