Zhang, Yue-Jun; Zhang, Han; Gupta, Rangan

Book

Forecasting the Artificial Intelligence index returns : a hybrid approach

University of Pretoria
Department of Economics Working Paper Series

Forecasting the Artificial Intelligence Index Returns: A Hybrid Approach

Yue-Jun Zhang
Hunan University
Han Zhang
Hunan University
Rangan Gupta
University of Pretoria

Working Paper: 2021-82
November 2021

Department of Economics
University of Pretoria
0002, Pretoria
South Africa
Tel: +27 12 420 2413

Forecasting the Artificial Intelligence Index Returns: A Hybrid Approach

Yue-Jun Zhang[*], Han Zhang[**], Rangan Gupta[***]

Abstract: Forecasting of the artificial intelligence index returns is of great significance for financial market stability and the development of artificial intelligence industry. To provide investors more reliable reference in terms of artificial intelligence index investment, this paper selects the Nasdaq CTA Artificial Intelligence and Robotics (AI) Index as the research target, and proposes novel hybrid methods to forecast the AI index returns by considering its nonlinear and time-varying characteristics. Specifically, this paper uses the ensemble empirical mode decomposition (EEMD) method to decompose the AI index returns, and combines the least square support vector machine approach together with the particle swarm optimization (PSO-LSSVM) method and the generalized autoregressive conditional heteroskedasticity (GARCH) model to construct novel hybrid forecasting methods. The empirical results indicate that: first, the decomposition and integration models usually produce superior forecasting accuracy than the single forecasting models, due to the complicated feature of the non-decomposed data. Second, the newly proposed hybrid forecasting method (i.e., the EEMD-PSO-LSSVM-GARCH model) which combines the advantage of traditional econometric models and machine learning techniques can yield the optimal forecasting performance for the AI index returns.

Keywords: AI index return forecasting; PSO-LSSVM model; GARCH model; Decomposition and integration model; Combination model

JEL codes: Q43; G15; E37

[*] Corresponding author. Business School, Hunan University, Changsha 410082, China; Center for Resource and Environmental Management, Hunan University, Changsha 410082, China. Email: zyjmis@126.com.

[**] Business School, Hunan University, Changsha 410082, China; Center for Resource and Environmental Management, Hunan University, Changsha 410082, China. Email: hanalms@163.com.

[***] Department of Economics, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa; Email: rangan.gupta@up.ac.za.

1. Introduction

Making accurate predictions of financial time series is one of the most challenging tasks for researchers and financial-market participants (Zhang and Wang, 2019; Zhang et al., 2020; Ghosh et al., 2021). In recent years, index investment has gained extensive attention around the world, and has impacted the traditional mode of asset-management business, while the compound annual growth rate (CAGR) of index funds has always maintained rapid growth. Especially with the development of artificial intelligence (AI) technology, in the wake of the so-called 4th Industrial Revolution, more enterprises have begun to join the artificial intelligence industry and the artificial intelligence technologies have dramatically reduced the global poverty (Gruetzemacher et al., 2021), and investors have also started to pay closer attention to the emerging field. According to the AI Index 2021 annual report, despite the pandemic, the year of 2020 still saw a 9.3% increase in the amount of private AI investment from 2019, a higher percentage increase than in 2019 (5.7%). Furthermore, the statistical data also show that the United States remains the leading destination for private investment, with over USD 23.6 billion in funding in 2020, followed by China (USD 9.9 billion) and the United Kingdom (USD 1.9 billion) (Zhang et al., 2021). It can be seen that the growth in the AI investment trend is here to stay, and how to seize this smart investment boom is an important question. In this regard, relevant investors need to first choose a reliable index reflecting the investment opportunities associated with the AI technology, which allows them to dynamically grasp the evolution of the returns for the entire range of this industry.

At present, the indexes related to artificial intelligence and robots mainly include the Nasdaq CTA Artificial Intelligence and Robotics Index (NQROBO Index), the Global Robotics and Automation Index (ROBO Index), and the Indxx Global X Robotics & Artificial Intelligence Index (IBOTZ Index). Among them, the Nasdaq CTA Artificial Intelligence and Robotics Index (hereafter referred to as AI index) is designed to track the performance of companies engaged in the artificial intelligence and robotics segment of the technology, industrial, medical and other economic sectors, and mainly includes the companies in artificial intelligence or robotics that are classified as either enablers, engagers or enhancers. Therefore, this index can comprehensively reflect the overall stock prices change and the associated development of the AI industry, and is clearly the most important among the three major indexes above. Meanwhile, based on its price data, we can further figure out that from December 19, 2017 to July 23, 2021, the cumulative return rate of AI index reached 84.84% and the annualized return rate was 33.92%. Besides, the movement in this index is also closely tied with other financial assets (Le et al., 2021; Tiwari et al., 2021). Therefore, it is of great significance to forecast the AI index returns accurately, so as to provide an important reference for investors to select suitable index funds and investment tools, and help them target the investing opportunities of the growing artificial intelligence and robotics industries by grasping its associated trends and risks.

However, the existing literature about AI index returns forecasting is relatively scarce, especially accounting for the nonlinear and time-varying characteristics of this

index. The financial time series forecasting models commonly used can be classified into traditional econometric models and machine learning methods, and both of them possess their own advantages and disadvantages in the process of forecasting. For example, the traditional econometric models are usually effective in capturing the linear and time-varying components, while they cannot fully capture the nonlinear components and have several requirements for data stability (Lin et al., 2011; Zhang et al., 2015). However, the machine learning methods are suitable for prediction of non-stationary nonlinear time series because of their flexible nonlinear function-fitting capabilities and less-restrictive assumptions that are imposed on the data, but their forecasting performance is easily affected by data size and parameter settings (Wang et al., 2005; Psaradellis and Sermpinis, 2016). In addition, the existing literature also points out that the single models, characterizing a specific feature of the data, usually cannot identify all states and correlations in complex time series (Khashei and Bijari, 2011), and are unable to extract their inherent dynamics, which consequently affects the forecasting accuracy. Given these limitations, hybrid models gradually emerged in the financial time series prediction literature (Zhang and Zhang, 2018; Li et al., 2021; Xiao et al., 2021). For instance, Yu et al. (2008) propose the "decomposition-integration" hybrid models, and the results show that the hybrid models always possess better forecasting ability. Bildirici and Ersin (2013) combine the multi-layer perceptron model with the new Smooth Transition Autoregressive model and GARCH model, which introduce the fractional integration and asymmetric property (LSTAR-LST-GARCH-MLP model), and prove that this hybrid framework

can capture volatility clustering, asymmetry and nonlinearity characteristics of petrol prices. Rapach et al. (2010) indicate that model combination can improve the prediction performance by synthesizing the features-capturing capability of individual models. Given this, relevant issues pertaining to the AI index returns involve which type of model has better predicting capability, and how to design a reliable predicting method that accurately explore the intrinsic structural characteristics of AI index returns.

Therefore, this paper attempts to combine the econometric models and machine learning methods to depict the linear and nonlinear characteristics of AI index return, and then develop the hybrid forecasting approach given the complexity of the data generating process of the AI index. Specifically, this paper first employs the ensemble empirical mode decomposition (EEMD) model to decompose the AI index return series into a series of intrinsic mode functions (IMFs) and the residual term. Second, the different models (namely, the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) and Particle Swarm Optimization (PSO) with Least Squares Support Vector Machine (LSSVM), i.e., PSO-LSSVM) are developed to forecast the IMFs and the residual term, respectively, with the sum of forecasted values for all components being the final forecasting results of the decomposition and integration models. Finally, this paper employs two methods to combine the econometric models and machine learning models, and then constructs two novel hybrid models that can capture the nonlinear and time-varying components of AI index returns simultaneously.

The contribution of this paper mainly involves two aspects: (1) Previous research has primarily focused on the common comprehensive indexes in the financial market (such as the S&P500 index; see Rapach and Zhou (2021) for a detailed discussion on the literature associated with international stock market forecasting), but the comprehensive index cannot reflect and predict the development of artificial intelligence industry. This paper focuses on the Nasdaq CTA Artificial Intelligence and Robotics Index, and conducts an in-depth analysis and forecasting of this index, so as to provide more information for investors about the development in the AI industry. (2) The previous studies usually employ the single forecasting model but could not systematically capture the inherent structural characteristics of overall index returns (Rapach and Zhou, 2013; Tiwari et al., 2016). Given this, our paper attempts to explore the appropriate forecasting models for the AI index return from multiple perspectives for the first time.

The empirical results imply that the hybrid model (i.e., EEMD-PSO-LSSVM-GARCH) driven by data characteristics can overcome the limitations of a single model, and effectively depict the time-varying and non-linear characteristics in the AI index returns and achieve superior forecasting performance for the AI index returns, which provides relevant investors more reliable reference in terms of portfolio selection and asset management.

The remainder of this paper is organized as follows: Section 2 briefly describes the models. The data set is presented in Section 3, and the empirical results are discussed in Section 4, with Section 5 concluding the paper.

2. Methods

2.1. The EEMD method

In order to decompose the complex original signal into components with different characteristics, and maintain the non-stationary and non-linear features of the original time series data, this paper selects the EEMD method (Wu and Huang, 2009) to decompose the AI index returns sequence. The main steps of decomposition are as follows:

(1) Add a white noise series $o^i(t)$ with a given amplitude (i.e., 0.1) to the AI Index returns series $x(t)$, and the new series $x^i(t)$ is as follows:

$$x^i(t)=x(t)+o^i(t) \tag{1}$$

(2) Decompose the time series with added white noise $x^i(t)$ into *n* IMFs $c_j^i(t)$ (j = 1, 2, . . . , n) and a residual term $r^i(t)$ by the EMD method, and the results are as follows:

$$x^i(t)=\sum_{j=1}^{n} c_j^i(t)+r^i(t) \tag{2}$$

where $c_j^i(t)$ is the *j*th IMF in the *i*th trial.

(3) Repeat steps (1) and (2) for *M* number of times with different white noise each time and obtain the corresponding IMF components of the decomposition;

(4) Calculate the average of corresponding IMFs of *M* trials as the final IMFs as follows:

$$c_j(t)=\frac{1}{M}c_j^i(t) \tag{3}$$

Once the EEMD completes, the original time series can be expressed as a linear

combination of IMFs and the residual term, as follows:

$$x(t) = \sum_{j=1}^{n} c_j(t) + r(t) \tag{4}$$

where $c_j(t)$ (t = 1, 2, . . . , T) is the *j*th IMF using the EEMD method at time *t*, $r(t)$ is the final residual term, and *n* is the number of IMFs.

2.2. The PSO–LSSVM method

2.2.1. The LSSVM method

In order to better describe the non-linear characteristics of the AI index returns, we single out the LSSVM model which is the typical method in machine learning (Suykens and Vandewalle, 1999). The main reason is that the LSSVM regression algorithm can obtain a global optimization by solving a set of linear equations, which allows the model to be faster than the SVM framework. The specific description of the model is as follows: Given a set of samples $\{y_t, x_t\}_{t=1}^{T}$, $\mathbf{x}_t$ is the input vector and $\mathbf{y}_t$ is the output variable. Then the decision function can be defined as:

$$y(x) = \mathbf{w}^T \Gamma(x) + c_{bias} \tag{5}$$

where $\Gamma(x)$ denotes the nonlinear function that maps the input space to a high dimension feature space, *w* represents the weight vector, and c_{bias} is the bias term.

The objective function of the LSSVM model is:

$$\begin{aligned} &\min\left(\frac{1}{2}\|\mathbf{w}\|^2 + \frac{c_{reg}}{2}\sum_{t=1}^{T^{(\mathrm{tr})}}\sigma_t^2\right) \\ &s.t.\ \ y_t = \mathbf{w}^T\Gamma(x_t) + \zeta_t^{(tr)} + c_{bias},\ t = 1,2,\cdots,T^{(tr)} \end{aligned} \tag{6}$$

where c_{reg} is the regularization constant, and σ_t denotes the training error.

According to the Kuhn-Tucker conditions (Kuhn and Tucker, 1950), the final outcome of the LSSVM method can be described as:

$$y(x) = \sum_{t=1}^{T^{(tr)}} \tilde{\lambda}_t K(x, x_t) + c_{bias} \tag{7}$$

where $K(x, x_t)$ indicates the kernel function. We apply the radial basis function (RBF), which is commonly used in nonlinear regression problems (Keerthi and Lin, 2003). The RBF with a width of ω can be shown as:

$$K(x, x_t) = \exp(-0.5\|x - x_t\|^2 / \omega^2) \tag{8}$$

When using the LSSVM method with the RBF kernel function, the parameters ω and c_{reg} should be estimated and optimized.

2.2.2. The PSO method

The PSO method is an evolutionary computational technique, which is based on the simulation of flocking and swarming behaviors of birds and insects (Eberhart and Kennedy, 1995). Compared to other evolutionary computational methods, it can efficiently find the optimal or near optimal solutions to the problem under consideration. The PSO method uses a set of particles, representing potential solutions to the problem. Then each particle moves towards the optimal position, which can be found out by adjusting the direction of its previously best position, and its best global position.

We can define each particle as a potential solution to the problem in a *d*-dimensional search space. Let $U_i = (u_{i1}, u_{i2}, \cdots, u_{id})$ be the current position of particle *i*, $V_i = (v_{i1}, v_{i2}, \cdots, v_{id})$ be the current velocity, $P_i = (p_{i1}, p_{i2}, \cdots, p_{id})$ be the previous position, and $P_g = (p_{g1}, p_{g2}, \cdots, p_{gd})$ be the best position among all

particles. Then the best position of particle *i* can be computed using the following equations:

$$v_{id}^{k+1} = wv_{id}^{k} + c_1 r_1 [p_{id} - u_{id}^{k}] + c_2 r_2 [p_{gd} - u_{id}^{k}] \quad (9)$$

$$u_{id}^{k+1} = u_{id}^{k} + v_{id} \quad (10)$$

where v_i^k and u_i^k is the current velocity and position of particle *i*, respectively; w is the inertia weight; c_1 and c_2 are two positive constants called acceleration coefficients, and r_1 and r_2 are two independently uniformly distributed random variables with the range [0, 1].

2.2.3. The PSO-LSSVM method

Due to the parameters ω and c_{reg} having great influence on the forecasting accuracy, we employ the PSO method to obtain the optimal parameters (Eberhart and Kennedy, 1995), and the main steps of the PSO-LSSVM approach can be described as follows:

Step 1. Take the parameters (ω, c_{reg}) as swarms and initialize a population of particles with random positions and velocities;

Step 2. Evaluate the fitness of each particle. We use the following fitness function: $\text{Fitness} = [\frac{1}{N}\sum_{i=1}^{20}(\hat{y}_i - y_i^2)]^{1/2}$, where y_i and $\hat{y}_i$ represent the actual and forecasted AI index returns, respectively;

Step 3. Update the previous and global best fitness according to the fitness evaluation results;

Step 4. Update the velocity and position values for each particle until the stop conditions are satisfied (i.e., the number of iterations reaches the maximum 100 or the

optimal parameters satisfy the accuracy requirement, i.e., the value of fitness is less than 0.001). The velocity for each particle is calculated based on Eq. (9), and each particle moves to its next position according to Eq. (10).

2.3. The GARCH model

To capture the time-varying character of AI index returns movement, we employ the GARCH model proposed by Bollerslev (1986) which is the most commonly used econometric model in analyzing the returns volatility of financial markets. The model is defined as follows:

$$\begin{aligned} r_t &= \delta r_{t-1} + \gamma r_{t-1} + u_t \\ u_t &= \varepsilon_t \sqrt{h_t} \\ h_t &= \alpha_0 + \alpha u_{t-1}^2 + \beta h_{t-1} \end{aligned} \tag{11}$$

where u_t represents residual series, and h_t is conditional variance. When $t = 1, \ldots, n$, $\varepsilon_t \sim \mathrm{N}(0, 1)$; and the model should satisfy $\alpha_0 > 0$, $\alpha \geq 0$, $\beta \geq 0$ and $\alpha + \beta < 1$.

2.4. The hybrid method for forecasting AI index returns

The hybrid method, which has capability to model both nonlinearity and time variations, can be considered to be a good strategy for AI index returns forecasting. Under this circumstance, we attempt to construct a hybrid model based on the decomposition and integration, and model combination methods, and the procedures can be described as follows:

(1) The EEMD method is used to decompose the original AI index return series to obtain the IMFs and the residual term;

(2) We normalize the decomposed IMFs components and residual term, and appropriately select training samples and test samples. Then, the single models above (i.e., the GARCH and PSO-LSSVM models) are used to forecast the IMF components and residual term respectively;

(3) The forecasting results of each IMF component and residual term are superimposed to obtain the final forecasting results of the decomposition and integration models (i.e., the EEMD-GARCH and EEMD-PSO-LSSVM models);

(4) The following two methods are used to obtain the hybrid predictions:

a) The GARCH model is built to predict high-frequency IMFs with time-varying characteristics, while the PSO-LSSVM model is built to predict low-frequency IMFs and residual terms with nonlinear characteristics. Next, the final forecasting results of the new hybrid model is obtained by superimposing the forecasts above, i.e., EEMD-PSO-LSSVM-GARCH(A) model;

b) We combine the forecasting results of EEMD-GARCH and EEMD-PSO-LSSVM in Step (3) by mean combination approach, and the new hybrid model i.e., EEMD-PSO-LSSVM-GARCH(B),[①] is used to obtain the final forecasting results.

2.5. The evaluation criteria for forecasting performance

According to Hansen and Lunde (2005), we apply two widely used statistical loss functions, i.e., Mean Square Error (MSE) and Mean Absolute Error (MAE), to

[①] We use the mean combination approach to obtain the EEMD-PSO-LSSVM-GARCH(B) forecasts. This is because some research points out that the simple mean forecast combination cannot be outperformed by other complicated forecast combination methods (Rapach et al., 2010; Claeskens et al., 2016).

evaluate the out-of-sample forecasting performance for AI index returns, which are defined as Eqs. (12)-(13):

$$MSE = \frac{1}{T-N}\sum_{t=N+1}^{T}(\hat{h}_t - h_t)^2 \tag{12}$$

$$MAE = \frac{1}{T-N}\sum_{t=N+1}^{T}|\hat{h}_t - h_t| \tag{13}$$

where h_t represents the actual return whereas $\hat{h}_t$ represents the forecasted return; T and N respectively stand for the number of full-sample and in-sample observations, while $T-N$ is the out-of-sample observations.

3. Data description

This paper, following Huynh et al., (2020), chooses the daily AI+Robo index price data from NASDAQ market as the research focus, with the data obtained from Bloomberg.[②] The full sample ranges from 12/19/2017-07/26/2021, and the specific sample periods for training- and testing-samples are 12/19/2017-10/13/2020 and 10/14/2020-07/26/2021, respectively. The AI index returns are calculated as: $r_t=100\times[log(p_t)-log(p_{t-1})]$, where p_t indicates the AI index price at time *t*. The daily AI index log-returns are shown in Figure 1.

[②] Further information about AI index is available at: https://indexes.nasdaqomx.com/Index/Overview/NQROBO.

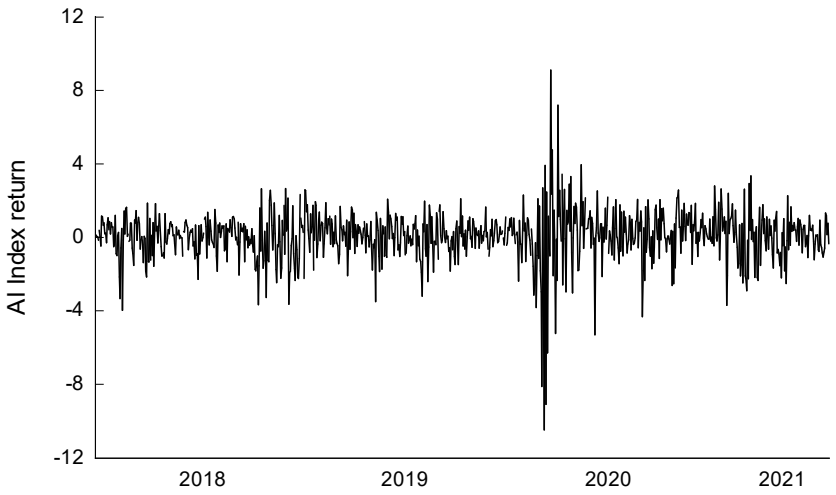

Figure 1. The AI index log-returns

Table 1 presents the descriptive statistics of the AI index returns. It can be observed that the AI index returns series has negative skewness and positive excess kurtosis, suggesting the presence of a leptokurtic and fat-tailed distribution. Moreover, the Jarque-Bera test results indicate that the null hypothesis of a normal distribution is rejected at the 1% significance level. The Ljung-Box Q-statistics for the squared returns also reject the null hypothesis of no autocorrelation up to the 10th order at the 1% significance level, which indicate the existence of autocorrelation in the AI index returns volatility. Additionally, Table 1 also presents the results of unit root tests. Specifically, the results of the Augmented Dickey-Fuller (ADF; Dickey and Fuller, 1981) test and Phillips-Perron (PP, Phillips and Perron (1988)) test reject the null hypothesis of a unit root at the 1% significance level, indicating that the AI index returns are stationary over the sample period.

Table1. Descriptive Statistics of the log-returns of the AI index

	AI index		AI index
Mean	0.0672	$Q(10)$	91.354 (0.0000)
S.D.	1.3775	$Q^2(10)$	619.02 (0.0000)
Skew	-0.9783	ADF	-8.8190 (0.0000)
Kurtosis	13.2993	PP	-29.2280 (0.0000)
J-B	4180.96 (0.0000)		

Note: *p*-values are reported in parentheses. S.D. represents the standard deviation. J-B is Jarque-Bera test statsitic with the null hypothesis of normal distribution. $Q(10)$ and $Q^2(10)$ denote the Ljung-Box Q-statistics of the returns and squared returns series for up to 10th order serial autocorrelation. ADF and PP are the statistics of the augmented Dickey-Fuller and Phillips-Perron unit root tests, respectively, based on lags determined by the Akaike Information Criterion (AIC).

4. Results and discussions

4.1. The EEMD decomposition results

Based on the discussion of the methods above, we obtain the EEMD decomposition result of AI index returns in Figure 2. First, the original AI index returns series is decomposed by the EEMD method into eight independent intrinsic mode functions and one residual term, which are defined as sub-series in the following section. As seen from Figure 2, the IMFs obtained by the EEMD algorithm are irregular, which are caused by the nonlinear and noise components of the AI index returns. In addition, the frequency of eight IMF components and the residual term is arranged from high to low, which shows the diversity in terms of frequency, and multi-scale characteristics of the AI index returns. Specifically, the average period of IMF1-IMF5 is relatively short, which is the high-frequency component of the original AI index returns series and reflect the impact of short-term irregular events in the artificial intelligence industry, and the GARCH model is used to forecast these

sub-series; the average period of IMF6-IMF8 is relatively long, and indicates the impact of major events in the field of the artificial intelligence, and the PSO-LSSVM model is applied for forecasting these sub-series. Moreover, the residual term declines slowly since September 2019, which reflects the impact of economic fundamentals on the Artificial intelligence industry, and indicates that the AI index returns have tended to decline since September 2019.

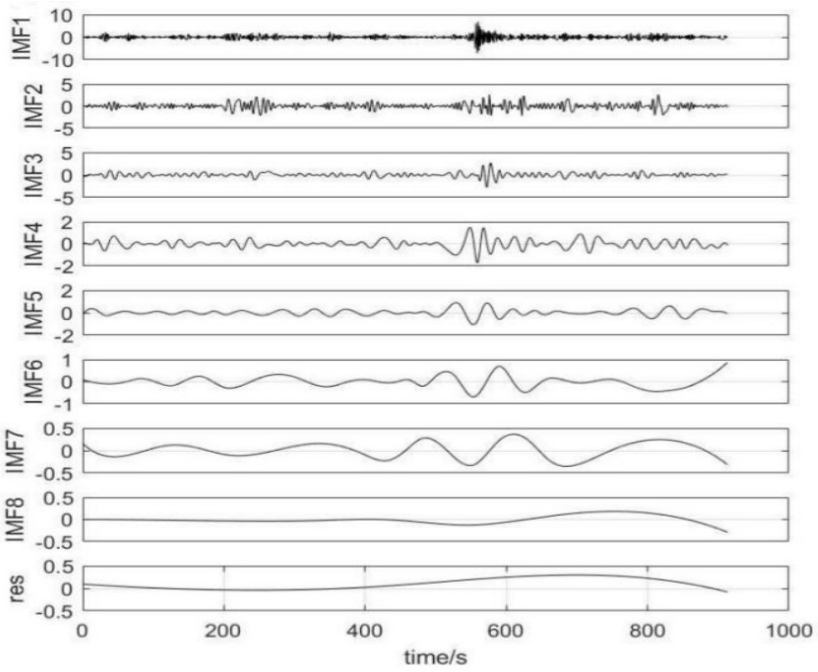

Figure 2. The EEMD decomposition of the log-returns of the AI index

4.2. Forecasting results of AI index returns

In order to find the optimal forecasting model for AI index returns based on the randomness, periodicity and trend of this series, we examine the forecasting performance of all competitive models. First, we consider the single forecasting model without data decomposition method, i.e., traditional econometric model (GARCH) and the machine learning framework (PSO-LSSVM). Second, we employ

the GARCH and the PSO-LSSVM models to forecast all the sub-series respectively, and obtain the forecasts from integrated-decomposed models (i.e., the EEMD-GARCH and EEMD-PSO-LSSVM models). Finally, we use two methods to combine the EEMD-GARCH and EEMD-PSO-LSSVM models, and derive the forecasts from the new hybrid models (i.e., EEMD-PSO-LSSVM-GARCH (A) and (B) models). The forecasting results are reported in Table 2. From this table, we can identify the following findings:

(1) As seen from Table 2, the values of the MSE and MAE indicate that the forecasting ability of PSO-LSSVM model is better than the GARCH model. The PSO-LSSVM model can better capture the non-linear characteristics of the AI index returns, and has the pros of non-linear mapping, self-learning and self-organization, which is observed to outperform the advantage of the GARCH model in its ability to capture the time-varying and volatility clustering characteristics of the AI index returns.

(2) Compared with the single model, the decomposition integration models usually perform better in their forecasting ability of the AI index returns. Specifically, as shown in Table 2, the values of MSE and MAE always indicate that the forecasting results of EEMD-GARCH and EEMD-PSO-LSSVM model are significantly better than the corresponding models that do not apply the EEMD algorithm. This result shows that the single model is greatly affected by the characteristics of the data itself such that their prediction ability is weaker. But the EEMD method can account for the periodicity, randomness and trend characteristics of the AI index, and decompose the

original sequence into simple modes effectively, so that we can obtain stable IMFs components and residual term and improve the accuracy of forecasting.

(3) The new hybrid models according to the methods above always perform the best. As seen from Table 2, the values of the two loss functions involving the two hybrid models are significantly lower than the single model, indicating that the hybrid models can consider the linear and non-linear characters of the AI index returns, and combine the advantages of the PSO-LSSVM model and GARCH model, thereby obtaining superior forecasting performance compared to the single model. Specifically, the values of MSE and MAE of the hybrid models are significantly reduced compared to other models, and the mean combination model (i.e., the B model) usually performs best among all the models considered.

Table 2. 1-day ahead forecasting results for the daily log-returns of the AI index

	MSE	MAE
GARCH	1.3760	0.8885
PSO-LSSVM	1.3530	0.8807
EEMD-GARCH	0.6745	0.6277
EEMD-PSO-LSSVM	0.6738	0.6375
EEMD-PSO-LSSVM-GARCH(A)	0.6433	0.6089
EEMD-PSO-LSSVM-GARCH(B)	**0.6423**	**0.6087**

Note: The numbers in the table refer to the values of the two loss functions. Bold numbers indicate that the corresponding models have the lowest forecasting losses.

4.3 Robustness checks

There may be some uncertainties affecting the central results above; therefore, two kinds of robustness checks are conducted in terms of data frequency and sample periods.

To determine whether our findings are robust to the frequency of the data, we use

weekly data to re-estimate the models, with the specific periods of training- and testing-samples being 12/19/2017-08/23/2020 and 08/24/2020-07/26/2021, respectively, given the full-sample of 12/19/2017-07/26/2021.

As seen from Table 3, the values of the MSE and the MAE of the EEMD-GARCH and EEMD-PSO-LSSVM are lower than the GARCH and PSO-LSSVM model without the EEMD method. It shows that the EEMD method can effectively decompose the AI index return series with the noise, so that we can obtain more accurate data for the following prediction process. Hence, as with daily data, the decomposition-integrated forecasting models are better than the single models at a weekly frequency too. Further, the hybrid models also yield superior forecasting performance than other models. Specifically, the MSE and MAE values of the hybrid models (A and B models) are significantly lower, and the mean combination model (B model) performs best among all models.

Table 3. 1-week ahead forecasting results for weekly log-returns of the AI index

	MSE	MAE
GARCH	5.6166	1.8659
PSO-LSSVM	5.8967	1.9826
EEMD-GARCH	2.7728	1.2973
EEMD-PSO-LSSVM	2.7412	1.3188
EEMD-PSO-LSSVM-GARCH(A)	2.7085	1.2966
EEMD-PSO-LSSVM-GARCH(B)	**2.6995**	**1.2869**

Note: The numbers in the table refer to the values of the two loss functions. Bold numbers indicate that the corresponding models have the lowest forecasting losses.

In summary, under the new data frequency, the forecasting performance of new hybrid models is significantly superior to the single model, as with daily data. In particular, compared to the EEMD-PSO-LSSVM-GARCH(A) model, the forecasting

accuracy of EEMD-PSO-LSSVM-GARCH(B) model that considers mean-combination is better, indicating its suitability for the AI index returns forecasting. In short, the new hybrid methods can describe the non-linearity and non-stationary characteristics of the AI returns series more comprehensively, and combine the advantages of different single models and obtain forecasts that contains the important predictive information contained in each model, as defined by their specific characteristics. Overall, our results are robust across high- and low frequencies of data.

Next, to determine whether different sample periods can affect our findings, we select a new sample period of 07/26/2018-07/26/2021 to re-estimate the models, and the corresponding in-sample and out-of-sample periods are chosen to be 07/26/2018-11/25/2020 and 11/26/2020-07/26/2021, respectively. The results of the 1-day ahead forecasting under this new set-up are presented in Table 4. By comparing the results from the two loss functions, we find that the forecasting results of the decomposition-integration models are superior to the single model. Besides, compared to other models, the two new hybrid models continue to achieve better forecasting performance. In summary, the central results obtained above are also robust to different sample periods.

Table 4. 1-day ahead forecasting results for daily log-returns of the AI index under alternative sample periods

	MSE	MAE
GARCH	1.1642	0.8452
PSO-LSSVM	1.2654	0.8807
EEMD-GARCH	0.6614	0.6239
EEMD-PSO-LSSVM	0.6674	0.6146

	MSE	MAE
EEMD-PSO-LSSVM-GARCH(A)	0.5855	0.5935
EEMD-PSO-LSSVM-GARCH(B)	**0.5592**	**0.5902**

Note: The numbers in the table refer to the values of the two loss functions. Bold numbers indicate that the corresponding models have the lowest forecasting losses.

5. Conclusions and future work

In order to forecast the AI index returns accurately, and judge which type of model can better predict the AI index return, this paper, for the first time, attempts to combine machine learning techniques with traditional econometric models based on "decomposition-integration" and "model combination" methods, so as to deeply explore the intrinsic structural characteristics of the returns associated with the AI index. Specifically, the EEMD method is used for decomposition and integration, and the basic single models in this paper include the PSO-LSSVM and GARCH models. Some main conclusions are drawn as follows:

First of all, the EEMD decomposition and integration method significantly improves the forecasting performance of single models of the AI index returns. This is mainly because the EEMD method can obtain more stable and simple mode, and fully consider the periodicity, randomness and trend characteristics of the AI index returns, thereby obtaining more accurate forecasting results driven by features of the data. In addition, the result is valid no matter for the PSO-LSSVM model or the GARCH model, which further proves the applicability of the decomposition and integration method to the AI index returns.

Second, regardless of whether we use daily or weekly data and different sample

periods, the forecasting performances of the GARCH model and the PSO-LSSVM model are not significantly different, and the final hybrid model (i.e., EEMD-PSO-LSSVM-GARCH) that combines these frameworks can significantly improve the forecasting performance of single models. This result shows that the traditional econometric model is suitable for describing the time-varying characteristics in the AI index return, while the machine learning model can better capture the nonlinear characteristics. And the final hybrid model can effectively combine their advantages, thereby capturing the time-varying and non-linear characteristics of the data simultaneously and obtaining superior forecasting results than single models.

The conclusions above have clear implications for financial-market participants. Specifically, the relevant investors can utilize the EEMD-PSO-LSSVM-GARCH model to capture and mine more data characteristics of the AI index returns and make more accurate forecasting decisions, which can provide important reference for them to target investment opportunities in the artificial intelligence industry. In addition, the conclusions above are also conducive to the healthy development of financial markets, especially the artificial intelligence industry, involving financial market risk management, option pricing, and asset allocation.

In the future, there is still much interesting work to be explored regarding the artificial intelligence industry. In particular, we can further explore the influence-factors for the artificial intelligence index, and analyze the characteristics of the artificial intelligence industry in further detail so as to construct an accurate

explanatory variables-based forecasting framework, which in turn can help investors further grasp the investment opportunities in the artificial intelligence industry.

Acknowledgments

We gratefully acknowledge the financial support from National Natural Science Foundation of China (no. 71774051) and Science and Technology Innovation Program of Hunan Province (no. 2020RC4016).

References

Bildirici, M., Ersin, Ö.Ö., 2013. Forecasting oil prices: smooth transition and neural network augmented GARCH family models. J. Pet. Sci. Eng. 109, 230–240.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econom. 31, 307-327.

Claeskens, G., Magnus, J.R., Vasnev, A.L., Wang, W., 2016. The forecast combination puzzle: A simple theoretical explanation. Int. J. Forecast. 32, 754-762.

Dickey, D.A., Fuller, W.A., 1981. Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica. 49, 1057-1072.

Eberhart, R.C., Kennedy, J.A., 1995. A new optimizer using particle swarm theory. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan, pp. 39-43.

Ghosh, P., Neufeld, A., Sahoo, J.K., 2021. Forecasting directional movements of stock prices for intraday trading using LSTM and random forests. Financ. Res. Lett. 3.

Gruetzemacher, R., Dorner, F.E., Bernaola-Alvarez, N., et al. 2021. Forecasting AI progress: A research agenda. Technol. Forecast. Soc. Chang. 170, 120909.

Hansen, P.R., Lunde, A., 2005. A forecast comparison of volatility models: does anything beat a GARCH(1,1)? J. Appl. Econom. 20(7), 873-889.

Huynh, T.L.D., Hille, E., Nasir, M.A., 2020. Diversification in the age of the 4th industrial revolution: the role of artificial intelligence, green bonds and cryptocurrencies. Technol. Forecast. Soc. Chang. 159, 120188.

Keerthi, S. S., Lin, C.J., 2003. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput. 15(7), 1667-1689.

Khashei, M., Bijarai, M., 2011. A new hybrid methodology for nonlinear time series forecasting model. Mod. Simul. Eng. 15, 1-5.

Kuhn, H.W., Tucker, A.W., 1950. Nonlinear programming. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. 2, 481-492.

Le, TN-L, Abakah, E.J.A., Tiwari, A.K., 2021. Time and frequency domain connectedness and spill-over among fintech, green bonds and cryptocurrencies in the age of the fourth industrial revolution. Technol. Forecast. Soc. Chang. 162, 120382.

Li, Y.Z., Jiang, S.R., Li, X.R., Wang, S.Y., 2021. The role of news sentiment in oil futures returns and volatility forecasting: Data-decomposition based deep learning approach. Energy Econ. 95, 105140.

Lin, K.P., Pai, P.F., Yang, S.L., 2011. Forecasting concentrations of air pollution by logarithm support vector regression with immune algorithms. Appl. Math. Comput.

217(12), 5318-5327.

Phillips, P.C.B., Perron, P., 1988. Testing for a Unit Root in Time Series Regression. Biometrika. 75, 335-346.

Psaradellis, I., Sermpinis, G., 2016. Modelling and trading the U.S. implied volatility indices. Evidence from the VIX, VXN and VXD indices. Int. J. Forecast. 32(4), 1268-1283.

Rapach, D.E., Strauss, J.K., Zhou, G.F., 2010. Out□of□sample equity premium prediction: Combination forecasts and links to the real economy. Rev. Financ. Stud. 23, 821-862.

Rapach, D.E., Zhou, G., 2021. Asset pricing: Time-series predictability. The Oxford Research Encyclopedia of Economics and Finance.

Rapach, D.E., Zhou, G., 2013. Forecasting stock returns. Handbook of Economic Forecasting. 2(Part A), 328-383.

Suykens, J.A.K., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural. Process. Lett. 9(3), 293-300.

Tiwari, A.K., Abakah, E.J.A., Le, TN-L., Leyva-de la Hiz, D.I., 2021. Markov-switching dependence between artificial intelligence and carbon price: The role of policy uncertainty in the era of the 4th industrial revolution and the effect of COVID-19 pandemic. Technol. Forecast. Soc. Chang. 163, 120434.

Tiwari, A.K., Dar, A.B., Bhanja, N., Gupta, R., 2016. A Historical Analysis of the US Stock Price Index Using Empirical Mode Decomposition over 1791-2015. Economics. 10, 1-15.

Wang, S.Y., Yu, L., Lai, K.K., 2005. Crude oil price forecasting with TEI@I methodology. J. Syst. Sci. Complex. 18(2), 145-166.

Wu, Z., Huang, N.E., 2009. Ensemble empirical mode decomposition: A noise—assisted data analysis method. Adv. Adapt. Data. 11, 1-41.

Xiao, Y.J., Wang, X.K., Wang, J.Q., et al., 2021. An adaptive decomposition and ensemble model for short-term air pollutant concentration forecast using ICEEMDAN-ICA. Technol. Forecast. Soc. Chang. 166, 120655.

Yu, L., Wang, S.Y., Lai, K.K., 2008. Forecasting crude oil price with an EMD—based neural network ensemble learning paradigm. Energy Econ. 30(5), 2623-2635.

Zhang, D., Mishra, S., Brynjolfsson, E., et al., 2021. The AI index 2021 annual report. 2021.

Zhang, J.L., Zhang, Y.J., Zhang, L. 2015. A novel hybrid method for crude oil price forecasting. Energy Econ. 49, 649-659.

Zhang, Y.J., Zhang, J.L., 2018. Volatility forecasting of crude oil market: A new hybrid method. J. Forecast. 37, 781-789.

Zhang, Y.J., Wang, J.L., 2019. Do high frequency stock market data help forecast crude oil prices? Evidence from the MIDAS models. Energy Econ. 78, 192-201.

Zhang, Y.J., Chu, G., Sheng, D.H., 2020. The role of investor attention in predicting stock prices: the long short-term memory networks perspective. Financ. Res. Lett. 38(2), 101484.