Adaptive Hybrid Optimized Support Vector Regression with Lasso Feature Selection for Short-term Load Forecasting

Jinxing Che, Huafeng Xian, Yuhua Zhang

***Abstract*—Accurate short-term load forecasting (STLF) is of positive significance to the effective management of power companies and the stable operation of society. In spite of many studies conducted in this field, there are few to consider the inherent disadvantages of an individual module, which results in sub-optimal forecasting accuracy. Therefore, by integrating data preprocessing module and optimization module into support vector regression (SVR) forecasting module, this paper successfully presents a novel model (AHO-Lasso-SVR). The data preprocessing module, which is comprised of feature construction and Lasso feature selection, is used to construct and select meaningful features. An adaptive hybrid optimization (AHO) algorithm is proposed by introducing two strategies on the basis of standard particle swarm optimization (PSO). The AHO algorithm inevitably increases the computational complexity of model learning, thus, this paper proposes a subsampling technology to improve the optimization efficiency of the algorithm based on the sparsity of SVR. The proposed model is used to forecast the load at 48 points in the next day. To verify the properties of the proposed model, power load data from New South Wales, Australia are adopted as a case study. The results reveal that our model positively exceeds all comparison models in terms of forecasting accuracy and stability.**

***Index Terms*—Short-term load forecasting, support vector regression, Lasso feature selection, adaptive hybrid optimization, subsampling technology**

I. Introduction

Due to the rise of artificial intelligence (AI), and the advent of the information age, STLF has attracted great attention. The deep integration of big data will produce unprecedented social value and commercial value for various industries, especially for the power industry. For power companies, the accuracy of STLF immediately influences the power distribution schemes and thus it is crucial to their economic returns. In practice, accurate forecasting can not only avoid the waste of electricity but also reduce the probability of power failure in small areas and contribute to social functioning [1]. However, due to the nonlinear and stochastic patterns of load consumption behavior, accurate load forecasting is challenging. To date, a mass of STLF methods has been developed from statistical methods (e.g., linear regression [2], ARIMA [3], ARMA [4]) to artificial intelligence methods (e.g., ANN [5], CNN [6], SVR [7]) to adapt to these patterns.

In spite of many studies have sprung up in this field, there are few to consider the inherent disadvantages of an individual module, which results in sub-optimal forecasting accuracy. To address the issue, some scholars have begun to tend to integrate multiple modules, which is an integrated framework, to process and predict power load data [8]–[11]. Data preprocessing play a vital role in the integrated framework, such as feature selection [28], mode decomposition [29], [30], wavelet transform [31], [32], etc. Similar to linear regression, the Lasso regression can also be used for prediction [33]. Since Lasso regression can compress the coefficient of features and make some regression coefficients become zero, researchers prefer to use the Lasso regression for feature selection in most cases [34]. In machine learning, when dealing with multi-dimensional data, dimensionality reduction is the first consideration to try to solve the problem with as few and representative features as possible. In that sense, using Lasso regression for feature selection is also an effective dimensionality reduction method.

Among the forecasting models, SVR has achieved excellent experimental performance because of its superior generalization ability. However, the learning complexity of SVR problems is $O(N^3)$ (N is the sample size of the training dataset) [27], and the learning complexity of SVR problems is $K*O(N^3)$ (K is the number of parameters setting search). As SVR parameters are difficult to be derived from a mathematical formula, most researchers rely on swarm intelligent (SI) algorithms to optimize SVR parameters. At present, the popular SI algorithms include PSO [12], [13], differential evolution (DE) [14], ant colony optimization (ACO) [9], simulated annealing (SA) [15] and genetic algorithm (GA) [16], etc. All these algorithms are random search algorithms, and they search the most possible solution randomly in the solution space. As a result, there is a risk of falling into the local optimum. How to avoid falling into the local optimum effectively is a key point as well as a difficult point for SI algorithm research. To enhance the performance of SI algorithms, scholars have addressed mass improvement measures [17], [18]. Qin et al. [19] propose an

Manuscript received April 9, 2021; revised September 9, 2021. The work was supported by the National Natural Science Foundation of China (Grant No. 71971105) and the National Statistical Science Research Project (Grant No. 2020LZ03) and "Thousand talents plan" for high level talents in Jiangxi Province (Grant No. jxsq2019201064). The work was also supported by the Humanities and Social Sciences Projects in Jiangxi Universities (Grant No. JC20128).

Jinxing Che is an associate professor of College of Science, Nanchang Institute of Technology, Nanchang 330099, Jiangxi, China (e-mail: jinxingche@163.com).

Huafeng Xian is a postgraduate student of School of Information Engineering, Nanchang Institute of Technology, Nanchang 330099, Jiangxi, China (corresponding author, e-mail: xianhuafeng950914@163.com).

Yuhua Zhang is a lecturer of College of Science, Nanchang Institute of Technology, Nanchang 330099, Jiangxi, China (e-mail: 1984zhangyuhua@163.com).

adaptive inertia weighted particle swarm optimization (AIWPSO), but they do not make any improvement on the search strategy. Similar PSO algorithms include FPSO [20] and QPSO [21]. Chen et al. [22] improve the sparrow search algorithm, mainly using the levy flight strategy to eliminate the convergence of the origin and reduce the jump of the optimal position.

The improved algorithm can avoid falling into the local optimal value. However, as the performance of a single algorithm is not satisfied, more scholars tend to use hybrid optimization. Specifically, two or more optimization algorithms are combined to achieve better results. Wang et al. [23] propose a GSA-PSO hybrid optimization algorithm, which utilizes a gravity search algorithm (GSA) to optimize PSO's particles. Mafarja et al. [24] propose a WOASA algorithm that uses the SA algorithm to enhance the optimal solution found by the whale optimization algorithm (WOA). Assad et al. [25] propose HS-SA by combining harmony search algorithm and SA algorithm to overcome the sensitivity of HS to the initial solution. Yu et al. [26] propose an improved DE algorithm with the levy flights mechanism and the SA algorithm to solve the contraction stagnation and early convergence problems. However, the above studies either improve the algorithm itself only or combine other algorithms simply, and few studies consider both aspects. To this end, an adaptive hybrid optimization (AHO) algorithm is proposed to take both aspects into account. Firstly, the SA algorithm is combined to increase the population diversity of the optimal particle neighborhood. Secondly, an adaptive inertia weight is introduced to balance the contradiction between global search and local search.

How to reduce the value of *N* on the basis of maintaining the uniform distribution of original samples provides us another promising direction to design an efficient optimization algorithm. Moreover, over-learning will lead to the risk of over-fitting. To reduce the learning complexity, this paper proposes a subsampling technology based on the sparsity of SVR to reduce the number of training samples, and thus improve the learning efficiency of the AHO algorithm. Since the complexity of model learning is reduced from the two dimensions of feature selection and subsampling technology, the learning time of the proposed integrated model can be effectively reduced. In this regard, a novel integrated model based on Lasso feature selection, subsampling technology, AHO algorithm and SVR forecasting model is proposed, and this model is applied to the STLF problem. The main contributions of this paper are as follows:

1） Lasso feature selection (LFS) is employed to select key features from candidate feature sets, which effectively solves the problem of feature redundancy and multicollinearity.

2） Machine learning algorithm SVR is adopted as STLF model, which is allowed through learning to forecast the power load at 48 points in the future day.

3） An adaptive hybrid optimization algorithm is proposed, which not only improves the standard PSO algorithm but also combines the SA algorithm.

4） To reduce the learning complexity, subsampling technology based on the sparsity of SVR is proposed to improve the efficiency of model learning.

The rest of the paper is arranged as follows: Section II demonstrates the proposed integrated model. Section III discusses the study results of a real case. Finally, Section IV concludes the paper.

II. Proposed Integrated Model

In this paper, an integrated model based on Lasso feature selection, subsampling technology, AHO algorithm and SVR is proposed. Its structure is shown in Fig. 1. This model consists of three modules: (i) data preprocessing module based on feature construction and Lasso feature selection; (ii) optimization module based on subsampling technology and AHO algorithm; (iii) SVR forecasting module.

Generally speaking, the original data provided by the power companies is a scalar power load sequence, and we can convert the power load series to the supervised learning problem of input-output mode based on phase-space reconstruction theory. In other words, we need to construct features that are related to load behavior, so it is necessary to detect features that influence load behavior. These features generally include historical load features and time features. However, there may be problems such as redundant features and multicollinearity, so that it is not reasonable to input all candidate features into the forecasting model based on SVR. Therefore, the constructed candidate features are firstly transported to the LFS stage. The data preprocessing module is used as the input of the SVR forecasting model. The subsampling technology and AHO algorithm optimizes hyperparameters by minimizing the 3-fold cross-validation error of SVR model, which is one of the main innovations of this study. The main technologies of the proposed integration model are analyzed as follow.

A. Data Preprocessing Module

Given a time series of power load $\{x(i): i = 1, \ldots, n\}$. To extract more useful information from time series, phase space reconstruction (PSR) theory proposed by Packard [35] provides a basis for transforming one-dimensional nonlinear data into multi-dimensional phase space. The most common method of PSR is to use Takens' delay embedding theorem [36], but Takens hasn't given the specific expression of the embedding dimension d and delay τ. Currently, autocorrelation function (ACF) can be used to estimate the multi-dimensional phase space d and τ [37]. Let x be the half-hour load series of New South Wales, and perform ACF analysis on x to detect historical load features. Autocorrelation coefficient measures the degree of correlation between two different periods of the same series. Figuratively speaking, it measures the influence of one's past behavior on one's present. The autocorrelation coefficient is calculated as follows.

$$\rho(\tau)=\frac{\sum_{i=1}^{n-\tau}(x_i-\mu)(x_{i+\tau}-\mu)}{\sum_{i=1}^{n}(x_i-\mu)^2} \tag{1}$$

where, τ is the delay time, n is the number of elements in x, and μ is the mean of x.

Nomenclature

ARIMA	auto-regressive integrated moving average	λ	non-negative regular parameter
ARMA	auto-regressive moving-average	p	proportion of stratified sampling
RNN	recurrent neural network	*popsize*	population size
LSTM	long short-term memory	*max_iter*	maximum number of iterations
ANN	artificial neural network	T	temperature
CNN	convolutional neural network	β	cooling coefficient
BPNN	backpropagation neural network	δ	stop temperature
RF	random forest	α	fine-tuning coefficient
XGBoost	extreme gradient boosting	C	penalty factor
QPSO	quantum-behavior particle swarm optimization	γ	kernel function width
FPSO	flexible particle swarm optimization	ε	pipe width
MAE	mean absolute error	ω	Inertia weight
RMSE	root of mean square error	τ	delay time
MAPE	mean absolute percentage error	N	sample size of training dataset
TIC	theil inequality coefficient	STD	standard deviation of the forecasting error

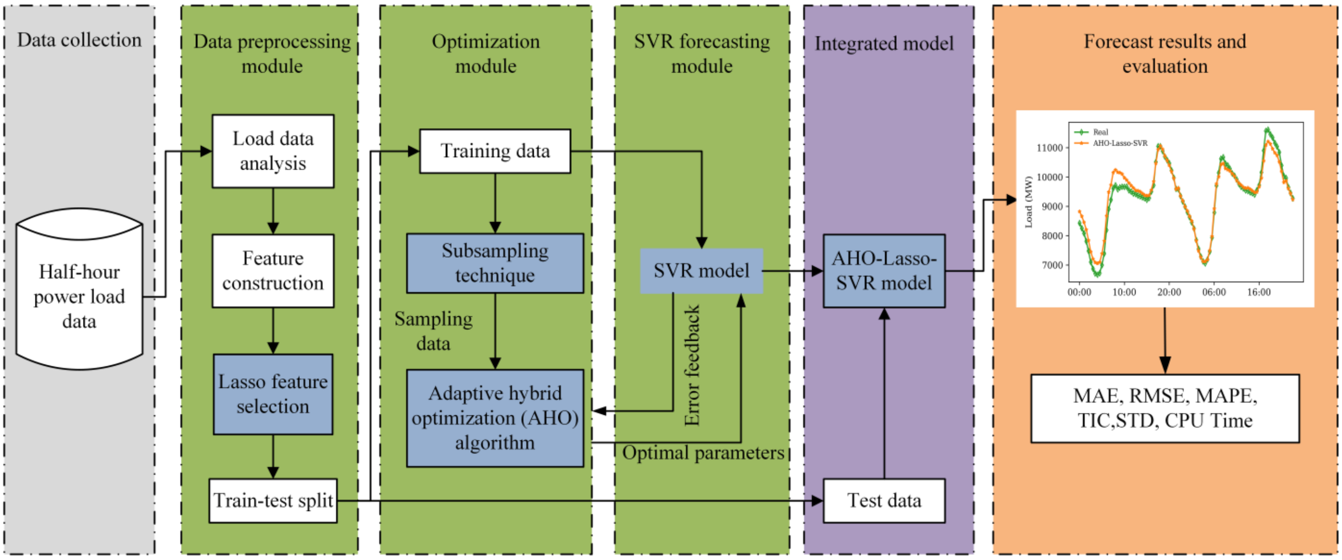

Fig. 1. Schematic diagram and main process of the proposed integration model.

In machine learning, when dealing with multi-dimensional data, dimensionality reduction is the first consideration to try to solve the problem with as few and representative features as possible. In that sense, using Lasso regression method for feature selection is an effective dimensionality reduction method [38]. It is a shrinkage estimation method with the idea of reducing the feature set (descending order). Lasso method can compress the coefficient of features and make some regression coefficients become zero, which can be widely used in model improvement and selection. Through the selection of penalty function, the purpose of feature selection is achieved by using the idea and method of Lasso. Model selection is essentially a process of seeking model sparse expression, which can be accomplished by optimizing a function problem of “loss” + “penalty”. The Lasso parameter estimation definition is shown in the formula.

$$\hat{\beta}(Lasso) = \arg\min_{\beta}{}^{2} \left\| y - \sum_{j=1}^{p} x_i \beta_i \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_i| \tag{2}$$

where λ is a non-negative regular parameter. The larger λ, the stronger penalty will be imposed on linear models with more features, resulting in a model with fewer features.

B. SVR Forecasting Module

The function of SVR is

$$f(x) = \omega\psi(x) + b \tag{3}$$

where **ω** is a weight vector, b is a constant. The goal expression is defined as

$$R(C) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n} \mathrm{L}(f(x_i) - y_i) \tag{4}$$

$$\mathrm{L}(z) = \begin{cases} 0 & \text{if } |z| \le \varepsilon \\ |z| - \varepsilon & \text{otherwise} \end{cases} \tag{5}$$

$\mathrm{L}(z)$ is called the ε-insensitive loss, and C is the penalty parameter. Therefore, introducing two slack-factor ξ and ξ^*, the following expression can be obtained.

$$\min_{\omega,b,\xi_i,\xi_i^*} \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$
$$\text{s.t.} \begin{cases} f(x_i) - y_i \le \varepsilon + \xi_i \\ y_i - f(x_i) \le \varepsilon + \xi_i^* \\ \xi_i \ge 0, \xi_i^* \ge 0, i = 1,2,\ldots,n \end{cases} \tag{6}$$

The above optimization problem is solved by introducing Lagrange multiplier.

$$\begin{aligned} L = &\frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\beta_i\xi_i - \sum_{i=1}^{n}\beta_i^*\xi_i^* \\ &+\sum_{i=1}^{n}a_i(f(x_i) - y_i - \varepsilon - \xi_i) \\ &+\sum_{i=1}^{n}a_i^*(y_i - f(x_i) - \varepsilon - \xi_i^*) \end{aligned} \tag{7}$$

Our optimization objective satisfies KKT conditions, that is to say, we can transform our optimization problem into an equivalent dual problem by means of Lagrange dual. The solution process is as follows: first, find the minimum values of the optimization function L for ω,b,ξ,ξ^*, then the maximum values of the optimization function L for Lagrange multiplier $\alpha,\alpha^*,\beta,\beta^*$. The above process needs to meet the KKT conditions. Finally, the solution of SVR can be obtained as

$$\begin{aligned} f(x) &= \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)(\boldsymbol{\psi}(\boldsymbol{x}_i),\boldsymbol{\psi}(\boldsymbol{x}_j)) + b \\ &= \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)K(x_i,x_j) + b \end{aligned} \tag{8}$$

where, $K(x_i,x_j)$ is the kernel function. In this study, RBF is used, and its function is as follows.

$$K(x_i,x_j) = \exp(-\gamma\|x_i - x_j\|^2) \tag{9}$$

In summary, there are three undetermined parameters: penalty factor C, pipe width ε, and kernel function width γ.

C. Optimization Module

This part demonstrates the optimization module based on subsampling technology and the AHO algorithm in detail. This module aims to efficiently maximize forecasting accuracy. Maximizing forecasting accuracy can be achieved by minimizing error. Therefore, we use the MAE of 3-CV as the objective function (fitness function) of the optimization module and its mathematical expression is as follows:

$$\underset{C,\gamma,\varepsilon}{argmin}\, MAE(C,\gamma,\varepsilon) \tag{10}$$

Firstly, the existing PSO algorithm and SA algorithm are discussed. Then, subsampling technology and the AHO algorithm are discussed.

Particle Swarm Optimization (PSO)

By simulating the predation process of a flock of birds, each bird is regarded as a particle, which is the feasible solution of the problem we need to solve. Throughout the population, birds constantly update their position and velocity according to the condition as they search for food. This condition is to update the position at the next moment based on the individual's historical optimal *pbest* and the global optimal *gbest* of whole population. The update of particle position is based on the cognition of particle itself and the information sharing of population. If its fitness decreases, it means that the particle is moving in a good direction; On the contrary, the particle is deviating from the optimal solution. The particle updates its velocity v and position x by using the following formula.

$$\begin{aligned} v_{i+1} = &\omega \times v_i + c_1 \times \text{rand}() \times (pbest_i - x_i) \\ &+c_2 \times \text{rand}() \times (gbest_i - x_i) \end{aligned} \tag{11}$$

$$x_i = x_i + v_{i+1} \tag{12}$$

where, ω is the inertia weight; c_1 and c_2 are the acceleration constant; rand() is a uniformly distributed random number on the interval [0,1].

However, PSO algorithm often falls into the local optima and cannot jump out when facing the problem with multiple local optima, although it has a fast convergence speed. As a result, ideal solution accuracy cannot be obtained.

Simulated Annealing (SA)

Inspired by the heating-cooling process of metals, Kirkpatrick et al. proposed SA algorithm [39], which adopts Metropolis acceptance criteria [40] to achieve global optimum. The Metropolis acceptance criteria describes the annealing process using a simple mathematical model as follows.

Assuming that the particle's fitness under state i is $f(i)$, then the particle follows the following rule from state i to state j at temperature T:

- If $f(j) < f(i)$, accept j as the current state;
- If $f(j) >= f(i)$, accept j as the current state with a certain probability, and the probability of acceptance is $p = \exp[-(f(j) - f(i))/KT]$, where, K is the Boltzmann constant in physics, and K=1 can be set in practical problems.

Subsampling Technology

The SI algorithm inevitably increases the computational complexity of the model, therefore, a subsampling technology is used to improve the optimization efficiency of the algorithm. According to the sparsity of SVR, its solution is only determined by the observation samples of $\alpha_i^* - \alpha_i \ne 0$. These observation samples are called support vectors. Moreover, the number of support vectors is far less than the number of training samples. Aiming at the principle that the solution of SVR is only determined by these support vectors, we propose a sampling technology to reduce the number of training samples and improve the optimization efficiency of our algorithm. Specifically, the training set D is divided into several sub-regions according to a specified rule, and p% samples are randomly sampled from these sub-regions [41], as shown in Fig. 2. The subsampling technology reduces the influence of the variability of each data block, and ensures that the sampled samples are representative enough. The subsampling technology performs better in the case that the problem is limited by data sparsity [42].

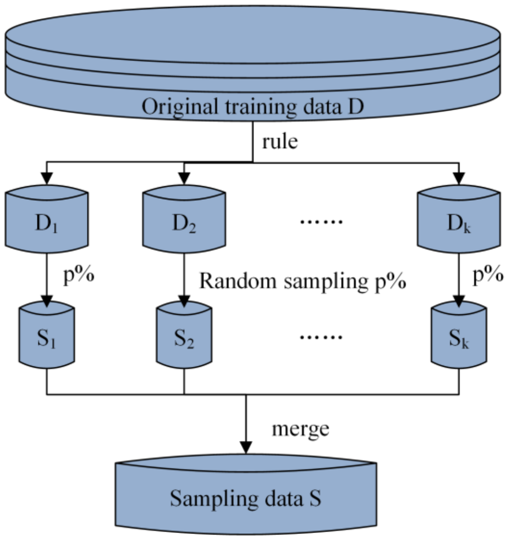

Fig. 2. Schematic diagram of the subsampling technique.

Adaptive Hybrid Optimization (AHO)

Our goal is to develop an algorithm with good global search ability and learning speed. So, we make two improvements on the basis of PSO algorithm.

First, adaptive inertia weight is introduced. Inspired by the individual search ability (ISA) defined in the literature [19], we define the particle global search ability (PGSA). In order to associate PGSA with inertia weight ω, we convert PGSA into inertia weight ω through sigmoid function. A large PGSA indicates that the particle has strong global searching ability, so the weight ω should be reduced to speed up the convergence. A small PDSA indicates that the particle has strong local search ability, which means that the weight ω should be increased to make the particle jump out of the local optimal neighborhood. The calculation formulas for PGSA and adaptive inertia weight ω are as follows.

$$PGSA_{id} = \frac{|x_{id} - pbest_{id}|}{|x_{id} - gbest_{id}| + \varepsilon} \quad (13)$$

$$\omega_{id} = 1 - 1/(1 + \exp(-PGSA_{id})) \quad (14)$$

where, x_{id} is the position of the i-th particle in the d-dimension, and ω_{id} is the inertia weight of the i-th particle in the d-dimension. ε is a positive number close to 0.

Secondly, the SA algorithm is introduced. To effectively increase the population diversity of optimal solution neighborhood, the SA algorithm is used to fine-tune the current global optimal particle, so that the search moves toward the direction where the optimal solution may exist, so as to find an optimal solution as soon as possible. The specific operation is as follows: In each iteration, the SA algorithm is used to fine-tune the current global optimal particle; accept a good position unconditionally, but accept a bad position with a certain probability. The velocity and position of other particles are updated according to Eq. (11) and (12), and ω is updated according to the adaptive weight of Eq. (14). The current global optimal particle is fine-tuned as follows: suppose $x_{\max d}$ is the upper bound of the d-dimensional variable in the solution space, and $x_{\min d}$ is the lower bound of the d-dimensional variable in the solution space.

$$x'_{id} = x_{id} + r_{id} \quad (15)$$

$$r_{id} = uniform(-\alpha(x_{\max d} - x_{\min d}), \alpha(x_{\max d} - x_{\min d})) \quad (16)$$

The *uniform(x,y)* function is to randomly generate a real number in the interval [x, y]. Define α as the fine-tuning coefficient. By introducing the idea of the SA algorithm, the global optimal particle that might have stalled is reactivated. Because the fine-tuning is performed in the neighborhood of the global optimal particle, it is more targeted than a random particle in the whole solution space, so the probability of finding the global optimal is greater.

Fig. 3 shows the implementation process of the AHO algorithm. The specific process of AHO algorithm is as follows:

step 1**:** subsampling of p% samples from the original data as the training data of fitness function;

step 2**:** initialize the velocity and position of each particle;

step 3**:** calculate the fitness function value of each particle;

step 4**:** update individual best position *pbest* and global best position *gbest*;

step 5**:** update the velocity and position of each particle in the following two cases:

1) if the particle is the global optimal particle, then we use SA algorithm to fine-tune the position of this particle, velocity v=0;
2) if the particle is not the global optimal particle, then the velocity and position are calculated by Eq. (11) and (12), and the inertia weight ω is given by Eq. (14).

step 6**:** if the end condition (usually the default precision or number of iterations) is not reached, return to *step 3*.

step 7**:** start the next iteration calculation; Otherwise, take the current *gbest* as the optimal solution.

In addition, the AHO algorithm also makes a "speed limit" processing for the particle's velocity. We limit the maximum velocity of particles to 20% of the variation range of variables per dimension to maintain the exploration and development ability of the algorithm.

III. Case Study

This section discusses the data source, data preprocessing results and forecasting results in detail. According to the forecasting results, the performance of proposed integrated model is verified by comparing with several models. The main parameters of the proposed integrated model are shown in Table I.

A. *Description of Dataset*

In the experiment, half-hour load data (48 sampling points per day) from New South Wales, Australia, in May 2007 is employed to verify the performance of the proposed model. Among them, the model is trained using data from May 1 to 29, and then the load of last two days is predicted. Specifically, the model is used to forecast the 48 points on May 30 according to the historical load (seven points) of the same time point in the previous seven days and the time features of predicted points. Then, we start rolling forecast, that is, after getting the actual load on May 30, predict the 48 points of the next day.

TABLE I
MAIN PARAMETERS OF ALL METHODS

Method	Parameter	Value
Lasso feature selection	λ	10
Subsampling technique	*p*	0.6
	rule	*'dow'*
AHO algorithm	*popsize*	10
	max_iter	20
	T	1
	β	0.5
	δ	0.001
	α	0.6
SVR	*C*	1268.09885405
	γ	0.0054497
	ε	0.75000663

B. *Results of Data Preprocessing*

Power Load Data Analysis

By analyzing the load data, it is found that there are two trends, obvious periodic volatility and horizontality. The specific load curve is shown in Fig. 4. On the one hand, it shows periodicity, with the change of week and day, the load has its own internal development pattern in each week; the load also has a greater similarity in the 5 workdays from Monday to Friday, but the load on weekend is usually lower than that on workday. On the other hand, it shows horizontality. The load trend between weeks shows horizontal volatility that has not obvious changing trend. In addition, the change of power load is also affected by many uncertain factors, such as the weather, people's activities and other factors.

The ACF is shown in Fig. 5. Only the recent load value has a significant impact on the current value, and the farther the past value is, the smaller the impact has on the current value. As the delay number τ increases, the autocorrelation coefficient of the load data decays slowly. The current load has a relatively higher correlation coefficient with the load from the same time one day ago than the same time seven days ago. Since there are 48 sampling points in a day, the value of τ is 48, 96, 144, 192, 240, 288, 366. In this way, the seven-day historical load can be used as historical load features.

Feature Construction

The original data is a scalar time series. According to the PSR, the influence degree of past values on present values can be detected by means of ACF analysis. Take the past values of τ=48, 96, 144, 192, 240, 288, 366 as historical load features. According to the above analysis of the load curve, it is found that the load curve has obvious periodicity. In other words, electricity consumption is related to time. Therefore, we can also extract the relevant time features from the time series and take them as the input. The constructed features are listed in Table II, and these features are preliminarily selected as candidate features.

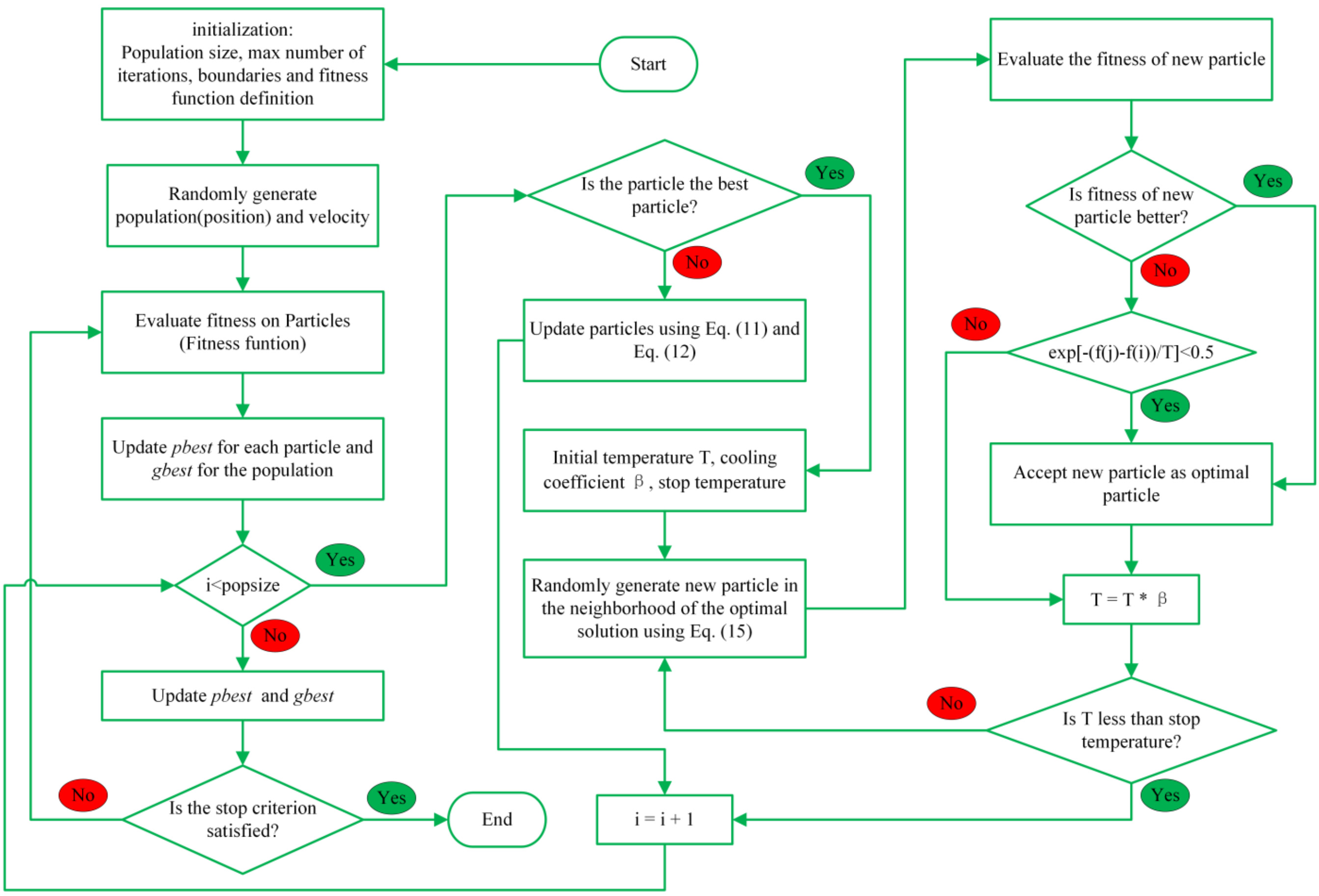

Fig. 3. The flow chart of our proposed AHO algorithm.

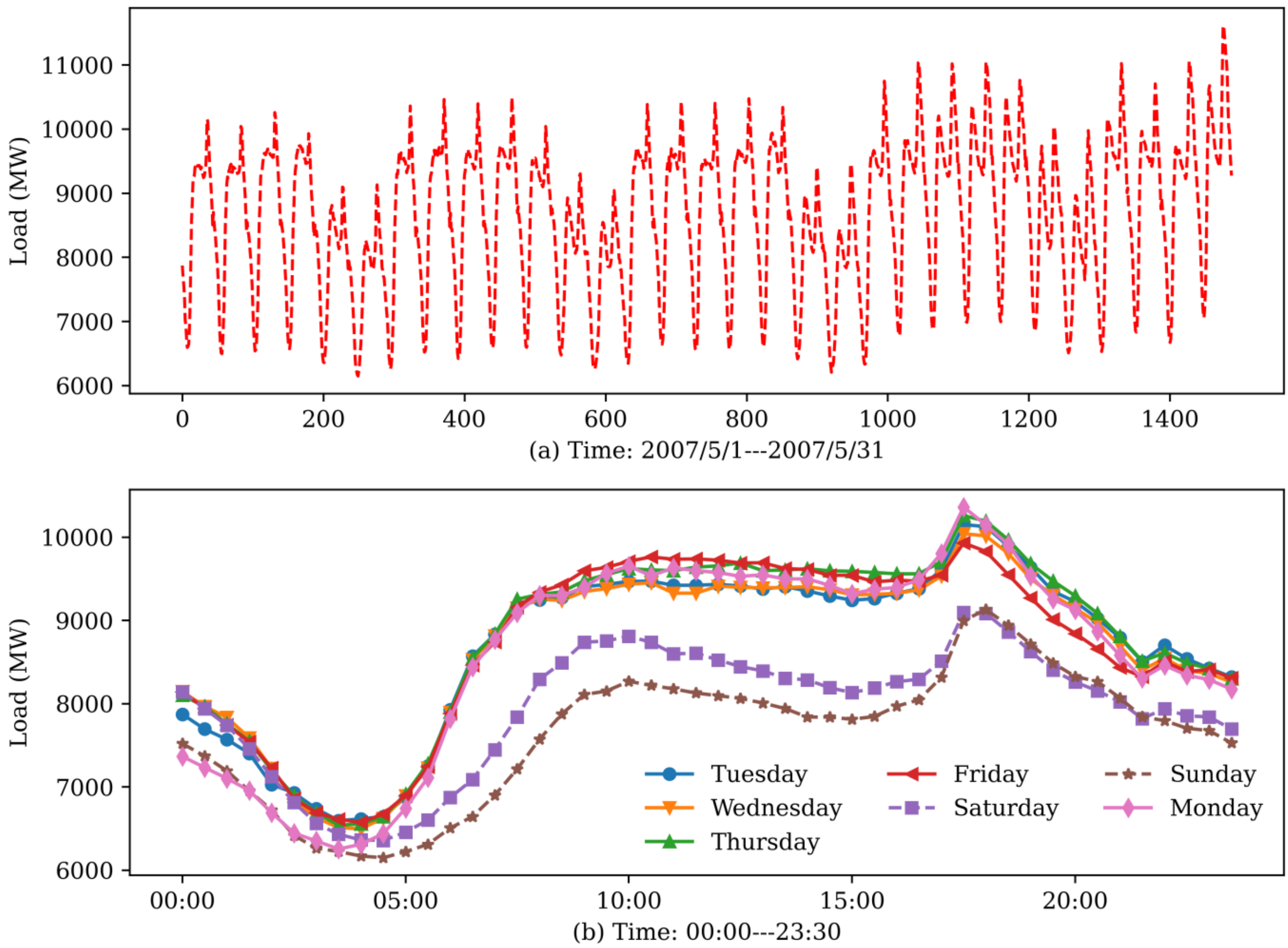

Fig. 4. Load curve for New South Wales; (a) Load curve in May 2007; (b) Daily load curve in the first week of May.

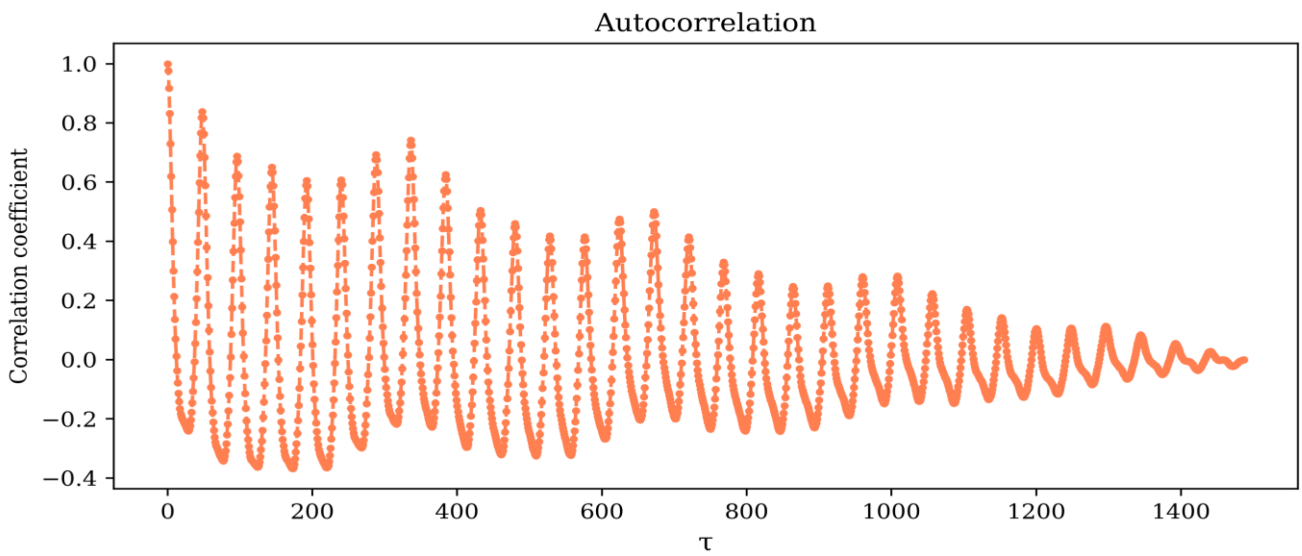

Fig. 5. Autocorrelation diagram of load data.

Lasso Feature Selection (LFS)

Actually, inputting irrelevant features that may complicate the model do not respond positively to the forecasting model [43], therefore, it is necessary to use Lasso regression method to select key features. Table III lists the coefficients for each feature. We regard features with a coefficient of 0 as non-critical features, and removing non-critical features can reduce model complexity and improve learning efficiency. Thus, the final features used in the model are *dow, dom, minute, night, t_m48, t_m96, t_m144, t_m192, t_m240, t_m288, t_m336.*

Data Standardization and Evaluation Metrics

Because there are large numerical differences among the inputs, the following data standardization can eliminate the impact of numerical differences.

$$x^* = \frac{x-\mu}{\sigma} \tag{17}$$

where, σ is standard deviation of x, and μ is the mean of x.

In this study, six evaluation metrics are employed: MAE, RMSE, MAPE and TIC are used to evaluate the forecasting accuracy; STD is used to evaluate the forecasting stability; CPU Time is used to estimate the running time of the optimization algorithms.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|\hat{y}_t - y_t\right| \tag{18}$$

$$RMSE = \sqrt{\frac{1}{n}\times\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2} \tag{19}$$

$$MAPE = \frac{1}{n}\times\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|\times 100\% \tag{20}$$

$$TIC = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2} \Big/ \left(\sqrt{\frac{1}{n}\sum_{t=1}^{n}y_t^2} + \sqrt{\frac{1}{n}\sum_{t=1}^{n}\hat{y}_t^2}\right) \tag{21}$$

$$STD = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(e_t - \bar{e})^2} \tag{22}$$

where, y_t is the actual load value at point t, and $\hat{y}_t$ is the forecasting load value at point t. $e_t = y_t - \hat{y}_t$, $\bar{e} = \frac{1}{n}\sum_{t=1}^{n} e_t$.

C. *Experimental Results*

Based on the Lasso feature selection results, two experiments are designed to compare the proposed model with other forecasting models.

Experiment I: Comparison with classic individual models

Experiment I is designed to compare the prediction performance of the proposed model with several classical models, such as the statistical model (ARIMA [44]), the neural network models (BPNN [45], LSTM [46], RNN [47]) and the ensemble models (RF [48], XGBoost [49]). The specific parameters of the comparison models are shown in Table IV. The parameters determined by the models are obtained through experiments and literature analysis. The predicted results are shown in Table V and Fig. 8. The detailed description of predicted results is as follows:

The best forecasting accuracy is achieved by our proposed model, that is, the smallest MAE, RMSE, MAPE, and TIC values are 181.185, 246.832, 2.006%, and 0.013, respectively. In contrast, the comparative individual models have worse forecasting accuracy, with a MAPE value of 10.517%, 2.33%, 3.373%, 3.077%, 3.162% and 3.76%, respectively. The standard deviation of the proposed model is also the best with an STD value of 237.144.

Because the ARIMA modeling merely relies on historical load data, and its prediction error will accumulate with the increase of the prediction, the ARIMA model shows the poorest prediction performance in terms of accuracy and stability. The MAPE value of ARIMA is 8.511% higher than that of our proposed model, and the standard deviation of ARIMA is 1120.525, while the standard deviation of our proposed model is 237.144.

Remark 1. The evaluation metric values of the proposed model are obviously lower than that of the classical individual model. Therefore, from the comparison results, we can conclude that the prediction performance of the proposed AHO-Lasso-SVR model is evidently superior to classic individual models in power load forecasting.

Experiment II: Comparison Between AHO Algorithm and Other Optimization Algorithms

Experiment II is designed to verify the performance of the AHO algorithm. The proposed model (AHO-Lasso-SVR) is compared with the comparison models, which are based on PSO and its variants (QPSO [19], AIWPSO [17], FPSO [18]) as well as SA algorithms. The comparison models are denoted as PSO-Lasso-SVR, SA-Lasso-SVR, QPSO-Lasso-SVR, AIWPSO-Lasso-SVR and FPSO-Lasso-SVR, respectively. Some general parameters of these algorithms such as C is set to [0.001,1500], γ is set to [0.001,1], and ε is set to [0, 1]. The predicted results are shown in Table VI and Fig. 9, and the detailed description of predicted results is as follows:

The proposed model achieves the best values of the MAE, RMSE, MAPE, TIC and STD, where the MAPE value is 2.006%. The smaller MAE, RMSE, MAPE and TIC indicate that the proposed model with the AHO algorithm achieves excellent prediction accuracy. The smaller STD means that the proposed model based on the AHO algorithm has the best prediction stability.

TABLE II
CONSTRUCTED CANDIDATE FEATURES

variable	description	variable	description
dow	day of the week (integer 0-6)	*night*	is night (0,1)
dom	day of the month (integer 1-31)	*t_m48*	load value from 1 day earlier
hhod	half-hour of the day (integer 0-47)	*t_m96*	load value from 2 day earlier
hour	hour of the day (integer 0-23)	*t_m144*	load value from 3 day earlier
minute	minute of the day (integer 0-1339)	*t_m192*	load value from 4 day earlier
weekend	is weekend (0,1)	*t_m240*	load value from 5 day earlier
tdpom	ten-day period of the month (1,2,3)	*t_m288*	load value from 6 day earlier
hom	half of the month (1,2)	*t_m336*	load value from 7 day earlier

TABLE III
COEFFICIENT TABLE

variable	*dow*	*dom*	*hhod*	*hour*	*minute*	*weekend*	*tdpom*	*hom*
coefficient	-35.79552	9.9388	0	0	-0.00038	0	0	0
variable	*night*	*t_m48*	*t_m96*	*t_m144*	*t_m192*	*t_m240*	*t_m288*	*t_m336*
coefficient	35.93089	0.20639	-0.00904	0.07845	-0.03883	-0.032	0.03917	0.77491

TABLE IV
EXPERIMENTAL PARAMETER SETTINGS IN COMPARISON MODELS

Model	Parameter	Value
BPNN	Epochs	50
	Learning_rate	0.01
	Hidden layer	[50,50]
LSTM	Epochs	50
	Learning_rate	0.01
	Hidden layer	[100,50]
RNN	Epochs	50
	Learning_rate	0.01
	Hidden layer	[10,10]
ARIMA(p,d,q)	Autoregressive term (p)	7
	Moving average number (q)	9
	Difference times (d)	2
RF	n_estimators	200
	max_depth	4
XGBoost	n_estimators	200
	learning_rate	0.8
	gamma	5
	max_depth	4

Fig. 6 shows the performance evaluation of the proposed model and the comparison models in terms of running time. The running time can reflect the computing efficiency of an algorithm. Since the AHO algorithm inevitably increases the computational complexity of model learning, we propose a subsampling technology to improve the learning efficiency of the proposed model. The running time of the proposed model based on subsampling technology and AHO algorithm is 38.999 seconds, which is faster than SA, standard PSO and its variants.

To further explore the performance of proposed model, relative error is introduced to analyze the model. Fig. 7 shows the box plot of relative error for six models. It can be intuitively seen from the box plot that the relative errors of most samples of proposed model are less than 0.03, and the median is the lowest, which indicates that the proposed model is more stable. Although the relative errors of most samples of PSO-Lasso-SVR, SA-Lasso-SVR and FPSO-Lasso-SVR are also less than 0.03, they contain abnormal points. QPSO-Lasso-SVR and AIWPSO-Lasso-SVR have the worst stability. Therefore, based on the analysis of the above results, the advantages of proposed model are proved in terms of accuracy and stability.

Remark 2. Results show an excellent performance of the AHO algorithm and it is better than that of PSO, SA, QPSO, AIWPSO and FPSO in terms of forecasting accuracy, stability and running time.

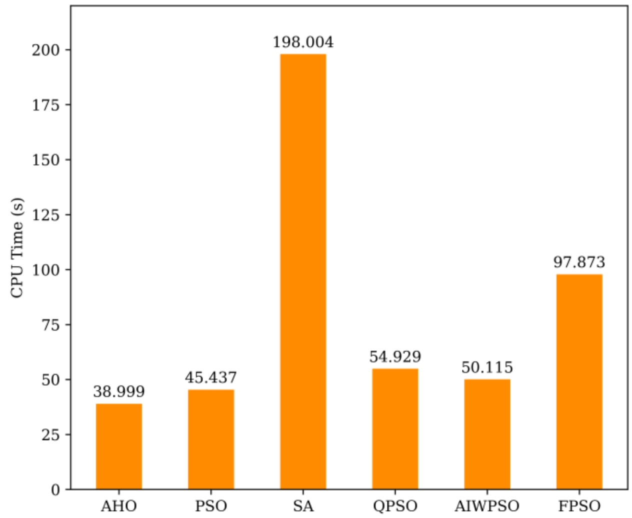

Fig. 6. Running time of AHO algorithm and other optimization algorithms.

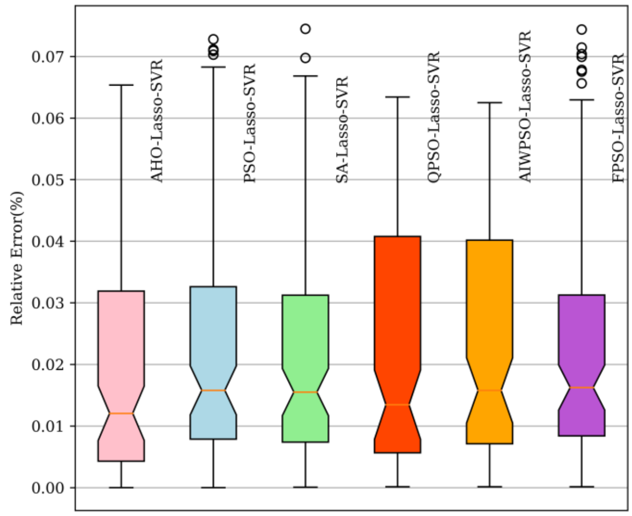

Fig. 7. Relative error box plot of the proposed model and the competitive models.

IV. CONCLUSION

Accurate STLF is of positive significance to the effective management of power companies and the stable operation of the society. In this paper, a novel integrated model includes three modules: (i) data preprocessing module based on feature construction and Lasso feature selection; (ii) optimization module based on subsampling technology and AHO algorithm; (iii) SVR forecasting module. The proposed model is evaluated on 30-minute power load data for New South Wales, Australia. The experiments show the following conclusions: (1) In Experiment I, the proposed model is compared with classical individual models (statistical model, neural network model and ensemble model), the lowest MAPE value is obtained from the proposed model, with a value of 2.006%. Thus, we can infer that the proposed model is obviously superior to the classical individual models in terms of power load forecasting. (2) Compared to the models based on other optimization algorithms (PSO, SA, QPSO, AIWPSO and FPSO), the proposed model (based on AHO algorithm) can achieve higher prediction accuracy and more stable prediction performance.

In conclusion, our proposed model is feasible, and it is superior to the classical individual models (ARIMA, BPNN, LSTM, RNN, RF and XGBoost) and other comparison models in terms of forecasting accuracy and stability.

TABLE V
PERFORMANCE COMPARISON BETWEEN THE PROPOSED MODEL AND THE CLASSICAL INDIVIDUAL MODELS

Metric	Proposed	Statistical model	Neural network model			Ensemble model	
		ARIMA	BPNN	LSTM	RNN	RF	XGBoost
MAE	**181.185**	963.829	218.312	307.305	286.629	299.836	362.452
RMSE	**246.832**	1170.757	257.512	368.964	323.108	374.762	481.894
MAPE (%)	**2.006**	10.517	2.33	3.373	3.077	3.162	3.76
TIC	**0.013**	0.064	0.014	0.019	0.017	0.02	0.026
STD	**237.144**	1120.525	242.985	345.805	324.797	375.047	469.807

TABLE VI
PERFORMANCE COMPARISON BETWEEN THE PROPOSED MODEL AND OTHER COMPETITIVE MODELS

Model	MAE	RMSE	MAPE (%)	TIC	STD	CPU Time (s)
Proposed	**181.185**	**246.832**	**2.006**	**0.013**	**237.144**	**38.999**
PSO-Lasso-SVR	207.418	271.116	2.344	0.014	245.246	45.437
SA-Lasso-SVR	201.274	264.890	2.265	0.014	238.48	198.004
QPSO-Lasso-SVR	205.505	271.824	2.227	0.014	271.879	54.929
AIWPSO-Lasso-SVR	208.525	266.240	2.280	0.014	266.144	50.115
FPSO-Lasso-SVR	209.100	272.278	2.365	0.014	242.265	97.873

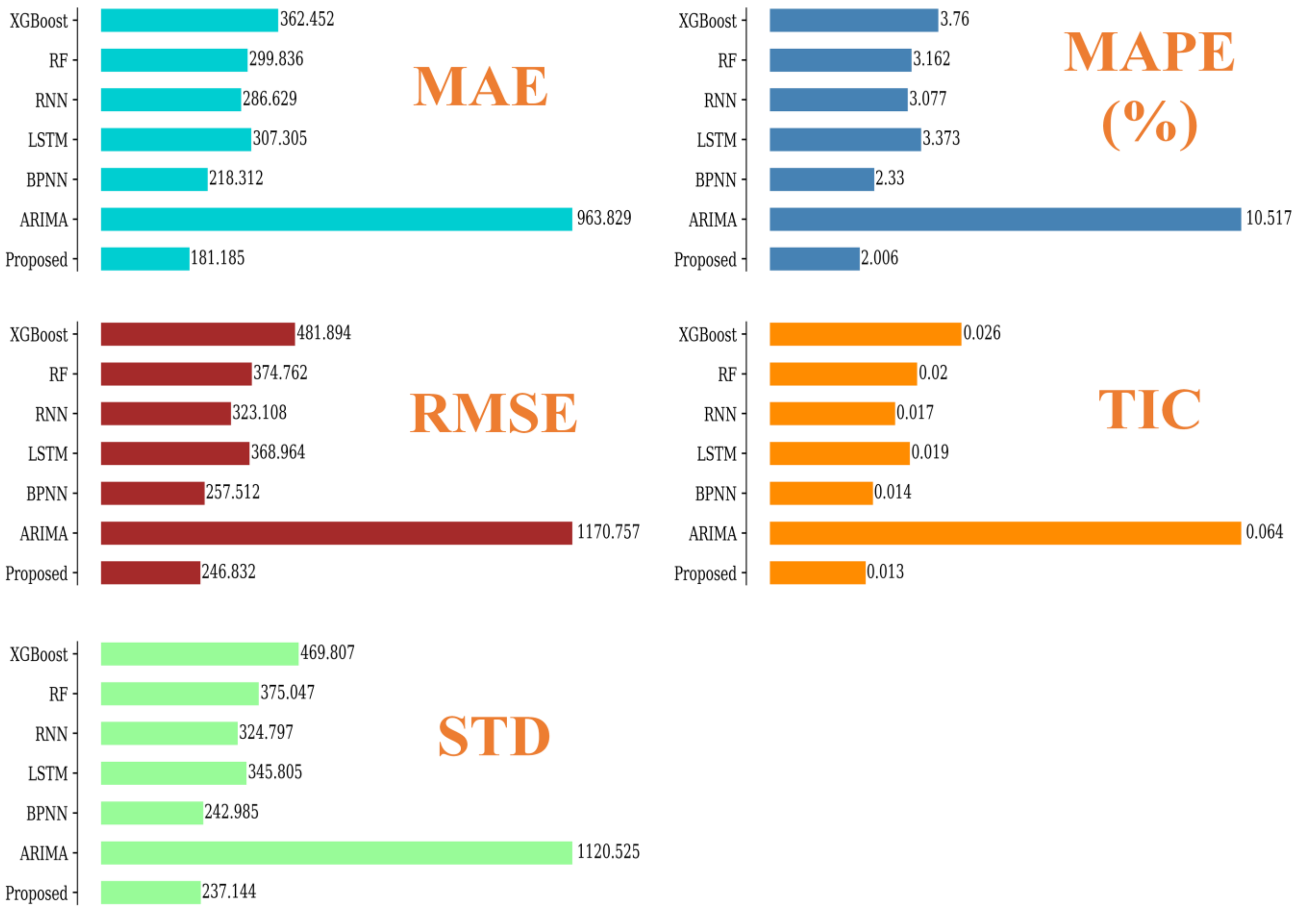

Fig. 8. Comparison results of our proposed model with the classical individual models.

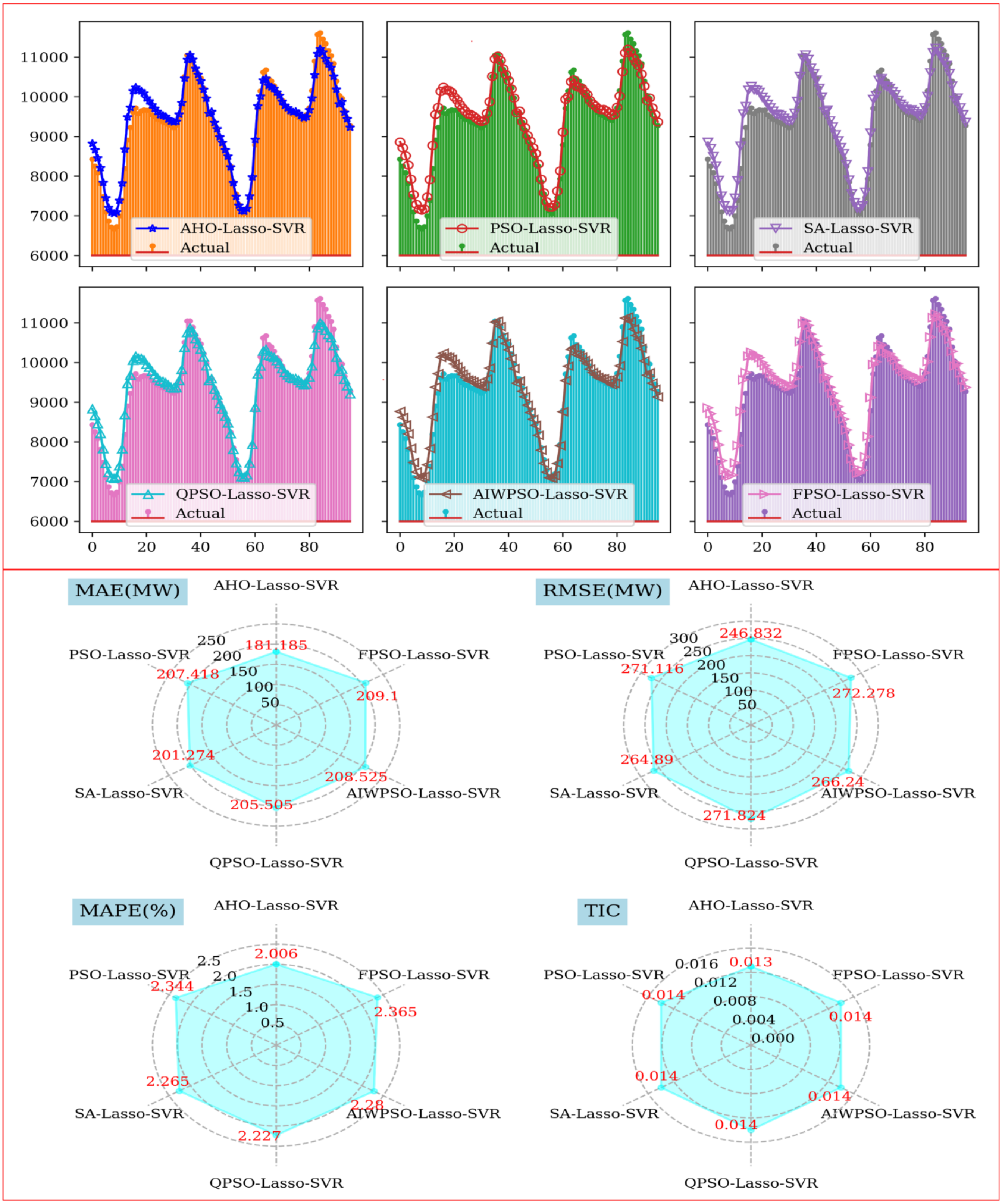

Fig. 9. Comparison results of our proposed model with other competitive models.

REFERENCES

[1] Miaomiao Wang, Qingwen Luo, Lulu Kuang, and Xiaoxi Zhu, "Optimized Rolling Grey Model for Electricity Consumption and Power Generation Prediction of China," IAENG International Journal of Applied Mathematics, vol. 49, no.4, pp577-587, 2019.

[2] Goia, C. May, G. Fusai, Functional clustering and linear regression for peak load forecasting, International Journal of Forecasting, vol. 26, no.4, pp700-711, 2010.

[3] W. Li, Z.G. Zhang, Based on Time Sequence of ARIMA Model in the Application of Short-Term Electricity Load Forecasting, in: 2009 International Conference on Research Challenges in Computer Science, Shanghai, pp11-14, 2009.

[4] K.G. Boroojeni, M.H. Amini, S. Bahrami, S.S. Iyengar, A.I. Sarwat, O. Karabasoglu, A novel multi-time-scale modeling for electric power demand forecasting: From short-term to medium-term horizon, Electric Power Systems Research, vol. 142, pp58-73, 2017.

[5] H. Dagdougui, F. Bagheri, H. Le, L. Dessaint, Neural network model for short-term and very-short-term load forecasting in district buildings, Energy and Buildings, vol. 203, pp109408, 2019.

[6] M. Cai, M. Pipattanasomporn, S. Rahman, Day-ahead building-level load forecasts using deep learning vs. traditional time-series techniques, Applied Energy, vol. 236, pp1078-1088, 2019.

[7] Y.L. Yang, J.X. Che, C.Z. Deng, L. Li, Sequential grid approach based support vector regression for short-term electric load forecasting, Applied Energy, vol. 238, pp1010-1021, 2019.

[8] M.S. Al-Musaylh, R.C. Deo, Y. Li, J.F. Adamowski, Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting, Applied Energy, vol. 217, pp422-439, 2018.

[9] R. Prasad, M. Ali, P. Kwan, H. Khan, Designing a multi-stage multivariate empirical mode decomposition coupled with ant colony optimization and random forest model to forecast monthly solar radiation, Applied Energy, vol. 236, pp778-792, 2019.

[10] G. Hafeez, K.S. Alimgeer, I. Khan, Electric load forecasting based on deep learning and optimized by heuristic algorithm in smart grid, Applied Energy, vol. 269, 2020.

[11] Z. Cao, C. Wan, Z. Zhang, F. Li, Y. Song, Hybrid Ensemble Deep Learning for Deterministic and Probabilistic Low-Voltage Load Forecasting, IEEE Transactions on Power Systems, vol. 35, no.3, pp1881-1897, 2020.

[12] K.O. Akande, T.O. Owolabi, S.O. Olatunji, A. AbdulRaheem, A hybrid particle swarm optimization and support vector regression model for modelling permeability prediction of hydrocarbon reservoir, Journal of Petroleum Science and Engineering, vol. 150, pp43-53, 2017.

[13] Wen-Chi Ho, Jau-Ming Su, and Chien-Yen Chang, "Maximal Market Potential of Feeder Bus Route Design Using Particle Swarm Optimization," IAENG International Journal of Applied Mathematics, vol. 49, no.4, pp470-477, 2019.

[14] F. Zhang, C. Deb, S.E. Lee, J. Yang, K.W. Shah, Time series forecasting for building energy consumption using weighted Support Vector Regression with differential evolution optimization technique, Energy and Buildings, vol. 126, pp94-103, 2016.

[15] S.M. Mousavi, E.S. Mostafavi, P. Jiao, Next generation prediction model for daily solar radiation on horizontal surface using a hybrid neural network and simulated annealing method, Energy Conversion and Management, vol. 153, pp671-682, 2017.

[16] S. Moreno-Carbonell, E.F. Sánchez-Úbeda, A. Muñoz, Rethinking weather station selection for electric load forecasting using genetic algorithms, International Journal of Forecasting, vol. 36, no.2, pp695-712, 2020.

[17] X. Liu, Y. Gu, S. He, Z. Xu, Z. Zhang, A robust reliability prediction method using Weighted Least Square Support Vector Machine equipped with Chaos Modified Particle Swarm Optimization and Online Correcting Strategy, Applied Soft Computing, vol. 85, 2019.

[18] Masoumi, S. Ghassem-zadeh, S.H. Hosseini, B.Z. Ghavidel, Application of neural network and weighted improved PSO for uncertainty modeling and optimal allocating of renewable energies along with battery energy storage, Applied Soft Computing, vol. 88, 2020.

[19] Q. Zheng, Y. Fan, S. Zhewen, W. Yu, Adaptive Inertia Weight Particle Swarm Optimization, Artificial Intelligence and Soft Computing – ICAISC 2006, vol. 4029, pp450-459, 2006.

[20] S.M. Ebrahimi, E. Salahshour, M. Malekzadeh, G. Francisco, Parameters identification of PV solar cells and modules using flexible particle swarm optimization algorithm, Energy, vol. 179, pp358-372, 2019.

[21] W. Song, C. Cattani, C.-H. Chi, Multifractional Brownian motion and quantum-behaved particle swarm optimization for short term power load forecasting: An integrated approach, Energy, vol. 194, 2020.

[22] Gonggui Chen, Bangrui Tang, Zhizhong Zhang, Xianjun Zeng, and Shuaiyong Li, "Short-Term Wind Speed Forecasting Based on Singular Spectrum Analysis, Fuzzy C-Means Clustering and Improved SSABP," Engineering Letters, vol. 29, no.2, pp351-364, 2021.

[23] R. Wang, J. Wang, Y. Xu, A novel combined model based on hybrid optimization algorithm for electrical load forecasting, Applied Soft Computing, vol. 82, 2019.

[24] M.M. Mafarja, S. Mirjalili, Hybrid Whale Optimization Algorithm with simulated annealing for feature selection, Neurocomputing, vol. 260, pp302-312, 2017.

[25] Assad, K. Deep, A Hybrid Harmony search and Simulated Annealing algorithm for continuous optimization, Information Sciences, vol. 450, pp246-266, 2018.

[26] Ziqiao Yu, Xiaoxia Zhang, and Xin Shen, "An Improved Differential Evolutionary Algorithm Based on Simulated Annealing and Levy Flights Mechanism," Engineering Letters, vol. 29, no.2, pp697-703, 2021.

[27] J. Che, Y. Yang, L. Li, Y. Li, S. Zhu, A modified support vector regression: Integrated selection of training subset and model, Applied Soft Computing, vol. 53, pp308-322, 2017.

[28] Koprinska, M. Rana, V.G. Agelidis, Correlation and instance based feature selection for electricity load forecasting, Knowledge-Based Systems, vol. 82, pp29-40, 2015.

[29] J. Bedi, D. Toshniwal, Energy load time-series forecast using decomposition and autoencoder integrated memory network, Applied Soft Computing, vol. 93, 2020.

[30] C. Li, Y. Tao, W. Ao, S. Yang, Y. Bai, Improving forecasting accuracy of daily enterprise electricity consumption using a random forest based on ensemble empirical mode decomposition, Energy, vol. 165, pp1220-1227, 2018.

[31] Z. Yang, L. Ce, L. Lian, Electricity price forecasting by a hybrid model, combining wavelet transform, ARMA and kernel-based extreme learning machine methods, Applied Energy, vol. 190, pp291-305, 2017.

[32] M. Imani, H. Ghassemian, Residential load forecasting using wavelet and collaborative representation transforms, Applied Energy, vol. 253, 2019.

[33] N. Tang, S. Mao, Y. Wang, R.M. Nelms, Solar Power Generation Forecasting With a LASSO-Based Approach, IEEE Internet of Things Journal, vol. 5, no.2, pp1090-1099, 2018.

[34] Brusaferri, L. Fagiano, M. Matteucci, A. Vitali, Day ahead electricity price forecast by NARX model with LASSO based features selection, in: 2019 IEEE 17th International Conference on Industrial Informatics(INDIN), Helsinki, Finland, pp1051-1056, 2019.

[35] N.H. Packard, J.P. Crutchfield, J.D. Farmer, R.S. Shaw, Geometry from a Time Series, Physical Review Letters, vol. 45, no.9, pp712-716, 1980.

[36] F. Takens, Detecting strange attractors in turbulence, In: Dynamical Systems and Turbulence, Warwick 1980. Springer Berlin Heidelberg, vol. 891, pp366-381, 1981.

[37] Z. Liang, Z. Xiao, J. Wang, L. Sun, B. Li, Y. Hu, Y. Wu, An improved chaos similarity model for hydrological forecasting, Journal of Hydrology, vol. 577, 2019.

[38] T. Robert, Regression shrinkage and selection via the Lasso, Journal of Royal Statistical Society. Series B(Methodological), vol. 58, no.1, pp267-288, 1996.

[39] S. Kirkpatrick, C.D.J. Gelatt, M.P. Vecchi, Optimization by Simulated Annealing, Science, vol. 220, no.4598, pp671-680, 1983.

[40] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of State Calculations by Fast Computing Machines, The Journal of Chemical Physics, vol. 21, no.6, pp1087-1092, 1953.

[41] E. Tipton, L. Hedges, M. Vaden-Kiernan, G. Borman, K. Sullivan, S. Caverly, Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling, Journal of Research on Educational Effectiveness, vol. 7, no.1, pp114-135, 2014.

[42] Z. Fu, J. Cheng, M. Yang, J. Batista, Y. Jiang, Wastewater discharge quality prediction using stratified sampling and wavelet de-noising ANFIS model, Computers & Electrical Engineering, vol. 85, 2020.

[43] H. Jiang, Model forecasting based on two-stage feature selection procedure using orthogonal greedy algorithm, Applied Soft Computing, vol. 63, pp110-123, 2018.

[44] Teresita L. Todelo, and Chris Jordan G. Aliac, "Predictability of Earthquake Occurrence Using Auto Regressive Integrated Moving Average (ARIMA) Model," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2019, 13-15 March, 2019,

Hong Kong, pp252-256.

[45] Y. Liu, D. Li, H. Pei, K. Liu, Y. Li, L. Yang, Short-term load prediction method for power distributing method based on back-propagation neural network, in: 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA), IEEE, Siem Reap, Cambodia, pp881-886, 2017.

[46] Narayan, K.W. Hipel, Long short term memory networks for short-term electric load forecasting, in: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, Banff, AB, Canada, pp2573-2578, 2017.

[47] B. Zhang, J.-L. Wu, P.-C. Chang, A multiple time series-based recurrent neural network for short-term load forecasting, Soft Computing, vol. 22, no.12, pp4099-4112, 2017.

[48] Nan Jiang, Fei Fu, Hua Zuo, Xiuping Zheng, and Qinghe Zheng, "A Municipal PM2.5 Forecasting Method Based on Random Forest and WRF Model," Engineering Letters, vol. 28, no.2, pp312-321, 2020.

[49] Y. Liu, H. Luo, B. Zhao, X. Zhao, Z. Han, Short-Term Power Load Forecasting Based on Clustering and XGBoost Method, in: 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, pp536-539, 2018.

Jinxing Che received his B.S. degree from Jiujiang University in 2007 and his M. S. degree in Applied Mathematics from Lanzhou University in 2010, as well as his Ph.D. degree in Probability Theory and Mathematical Statistics from Xidian University, China in 2019. He is currently an associate professor and Master's Supervisor in College of Science, Nanchang Institute of Technology, China. His main research interest is data analysis theory and application, hydrological information processing as well as prediction theory and method.

Huafeng Xian received his B.E. degree from Dongguan University of Technology in 2019. He is currently a master student in College of Information Engineering, Nanchang Institute of Technology, China. His main research interest is data analysis theory and application，as well as intelligent optimization algorithms.

Yuhua Zhang received his Bachelor of Science degree in Information and Computing Science from Hunan University in 2006 and his PhD degree in Computational Mathematics from Xiamen University in 2016. He is now a lecturer in the College of Science, Nanchang Institute of Technology, China. His main research interests include corporate finance and asset pricing, computer aided geometric design.