

RESEARCH

Open Access



# Using improved support vector regression to predict the transmitted energy consumption data by distributed wireless sensor network

Ni Guo<sup>1</sup>, Weifeng Gui<sup>2\*</sup>, Wei Chen<sup>3,4,5</sup>, Xin Tian<sup>6</sup>, Weiguo Qiu<sup>6</sup>, Zijian Tian<sup>1</sup> and Xiangyang Zhang<sup>1\*</sup>

\* Correspondence: [guiweifeng@hpu.edu.cn](mailto:guiweifeng@hpu.edu.cn); [zhangxyls1208@163.com](mailto:zhangxyls1208@163.com)

<sup>2</sup>School of Physics and Electronic Information Engineering, Henan Polytechnic University, Shiji Road 2001, Jiaozuo City 454003, Henan, China

<sup>1</sup>School of Mechanical Electronic and Information Engineering, China University of Mining and Technology (Beijing), No.11, Xueyuan Road, Haidian District, Beijing 100083, China  
Full list of author information is available at the end of the article

## Abstract

Massive energy consumption data of buildings was generated with the development of information technology, and the real-time energy consumption data was transmitted to energy consumption monitoring system by the distributed wireless sensor network (WSN). Accurately predicting the energy consumption is of importance for energy manager to make advisable decision and achieve the energy conservation. In recent years, considerable attention has been gained on predicting energy use of buildings in China. More and more predictive models appeared in recent years, but it is still a hard work to construct an accurate model to predict the energy consumption due to the complexity of the influencing factors. In this paper, 40 weather factors were considered into the research as input variables, and the electricity of supermarket which was acquired by the energy monitoring system was taken as the target variable. With the aim to seek the optimal subset, three feature selection (FS) algorithms were involved in the study, respectively: stepwise, least angle regression (Lars), and Boruta algorithms. In addition, three machine learning methods that include random forest (RF) regression, gradient boosting regression (GBR), and support vector regression (SVR) algorithms were utilized in this paper and combined with three feature selection (FS) algorithms, totally are nine hybrid models aimed to explore an improved model to get a higher prediction performance. The results indicate that the FS algorithm Boruta has relatively better performance because it could work well both on RF and SVR algorithms, the machine learning method SVR could get higher accuracy on small dataset compared with the RF and GBR algorithms, and the hybrid model called SVR-Boruta was chosen to be the proposed model in this paper. What is more, four evaluate indicators were selected to verify the model performance respectively are the mean absolute error (MAE), the mean squared error (MSE), the root mean squared error (RMSE), and the  $R$ -squared ( $R^2$ ), and the experiment results further verified the superiority of the recommended methodology.

**Keywords:** Distributed wireless sensor network, Energy consumption, Machine learning, Feature selection, Climate factors

## 1 Introduction

In recent years, the internet of things (IOT) was popularly applied to the smart city controls with the development of the information technology.

In IOT systems, the energy consumption monitoring system plays an important role in smart buildings, and it was always used to collect the real-time energy data and facilitate the energy control of the buildings. Actually, the real-time energy data was transmitted to energy consumption monitoring system by the wireless sensor network (WSN) [1, 2]. The wireless sensor is mainly composed of many intelligent distributed wireless sensor nodes, each of which has the function of sending message [3, 4]. In energy management system, accurately predict the energy use could assist energy manager make advisable strategies and schedule resource reasonably, this is the motivation of conducting this research, and the research data was acquired by the energy monitoring system. Another important reason for this study is the excessive energy consumption which caused large emissions of greenhouse gas has made huge impact on climate [5–7]. It was well known that more and more frequent and intense extreme weather events have occurred in recent decades which have profound effects on human society as well as on ecosystems [8–10]. The extreme weather events can affect people's life seriously in many obviously ways such as storms, floods, droughts, and heat waves. Many researchers have studied the attributions of extreme weather and showed that the human influence on the climate is obviously [11, 12]. Scott et al. studied that human influence may likely (probability > 90%) doubled the probability of a record warm summer [13] and pointed out that decreasing extreme weather events requires by reduction of greenhouse gas emissions which may resulting from the excessive use of energy consumption [14–16]. It is investigated that electricity dominated total energy consumption [17], and the building energy consumption accounts a large portion of the energy [18].

Supermarkets as one of the most energy-intensive type of commercial buildings have been studied by many researchers. However, the energy consuming situations are diverse in different climate areas. Many related works have been carried out in other countries. Kolokotroni et al. [19] have analyzed the supermarket energy consumption in the UK. Behfar et al. [20] have conducted to investigate supermarket equipment characteristics and operating faults in the USA. Braun et al. [21] have estimated the impact of climate change and local operational procedures on the energy use in several supermarkets throughout Great Britain. Many authors have also reported the building energy use in China. Lam et al. [22] have developed multiple regression model for office buildings in five major climates in China using 12 key building design variables. Authors Wang et al. [23] have been using different regression models to analyze the significant factors which affect the rural building heating energy in China. Although increasing investigations for the building energy use has been performed in China, very little work has been conducted in northwest region in China. Therefore, this paper focuses on analyzing the impacts of climate factors on the supermarket electrical energy use in northwest region of China using the transmitted energy consumption data by distributed wireless sensor network.

The following section has depicted the literature review and related works. In Section 3, the background of the research was presented, including the case information, data description, data pre-processing, and the involved methodologies. Section 4 has

demonstrated the development of hybrid models and the parameter optimization process. The results and discussion of the experiment were presented in Section 5. The last part of this paper concluded this research and summaries the limitation of this study.

## 2 Related works

It is a challenging work to precisely predicted the monthly electrical consumption owing to multiple influencing factors. Feature selection (FS) methods were effective ways to improve the prediction accuracy. Venkatesh and Anuradha [24] have illustrated that the FS methods are always applied to address the problem of dimensionality reduction which not only reduces the burden of the data but also avoid overfitting of the model. Three FS methods were presented in this paper, respectively: stepwise, Lars and Boruta. Among the stepwise algorithm was frequently employed to select useful subsets of variables and order the importance of variables [25]. The Lars algorithm was also applied to extract higher quality subset for model learning [26–28]. What is more, Boruta algorithm performed well in features filtering [29], and the details were depicted in literature [30]. In addition, the extracted variables will be taken as an optimal subset and used for model learning. In our work, power consumption of the equipment will be involved due to the energy consumption data that will be collected for a long time, and the data sampling frequency was once a month. The dataset used in this paper is the 5-year electricity consumption of supermarkets, and the collected dataset was still insufficient though it took 5 years.

With respect to the prediction model, machine learning methods were widely used in this field [31–33]. It was confirmed that the boosting and ensemble algorithms have superior performance in prediction field [34], but large dataset was required for those model learning, and it was not match to this work due to the collected dataset was insufficient. It is worth noting that the dataset in this work is high dimensional which was composed of electricity consumption and 40 climate factors, and SVR was recommended to perform the small-scale and high dimensional dataset. Mapping the original input variables to high dimensional feature space was an effective way to explore the nonlinear relationship between climate factors and building energy consumption, and this is in line with the model features of support vector machine (SVR). In addition, many researchers have verified the superiority of SVR in dealing with the small-scale dataset which with high dimensions. Ma et al. [35] have presented that the SVR could forecast the building energy consumption with good accuracy. Chen et al. have illustrated that the support vector regression (SVR) has a strong non-linear capability and could offer a higher degree of prediction accuracy and stability in short-term load forecasting [36]. Guo et al. have showed that the SVM is an effective method for problems with small number of samples, and the modified support vector machine has outperforms existing methods [37]. To this end, SVR was taken as the best option to model the dataset in this paper, and it was also combined with FS methods to improve the predictive performance of the model; the basic principles of SVR was detailed in Section 3. Another challenge work was the parameter tuning of the model. The optimal parameter will be searched in a certain range which was determined based on the prior knowledge of statistics, and the results were presented in Section 4. In this paper, another two machine learning methods were also used to compare with SVR,

respectively, and these are random forest (RF) regression and gradient boosting regression (GBR). RF algorithms was utilized in literatures [38–41], and the results have revealed the good performance of RF. GBR was also a widespread machine learning in predicting research field, as stated in [29, 42], and the GBR was performed well in energy prediction. The prediction performance of nine hybrid models was compared by measures of MAE, MSE, RMSE, and  $R$ -squared ( $R^2$ ), and the results exhibited that SVR-Boruta has higher performance with values of accuracy 90.585%.

### 3 Materials and prepare work

The transmitted energy consumption data by distributed wireless sensor network was regarding a supermarket building; it is essential to make a site visit investigation so as to obtain more detailed information about the supermarket, and the information of the supermarket was illustrated in the following part. What is more, data cleaning was also needed before conducting model learning, variable description, and the pre-processing of this work that was introduced in this section.

#### 3.1 Case information

The supermarket was opened in 2009 which is located in Shuozhou city, Shanxi province in the northwest region of China. Site visits were conducted to the selected supermarket, and it has three floors which with the total area of 1970 m<sup>2</sup> and sales area of 1598 m<sup>2</sup>; the detailed survey information is displayed in Table 1. The general lightings and a large number of spotlights are cross distributed on the second floor for fresh food sales and the third floor for commodity sales, and about 422 spotlights will work during the peak period of passenger flow that occurs at 8:00 to 11:00 am and 15:00 to 19:00 pm, and at the same time, the general lightings will be closed, other electrical equipment including cash registers and weighing instruments. The energy consumption of selected supermarket mainly composed by lightings, refrigeration and freezers, HVAC (heating, ventilation and air conditioning), and other electrical devices. What is more, the supermarket belongs to the temperate continental climate where the cold weather is longer than the warm weather, and the indoor temperature of the supermarket is greatly affected by outdoor climate, especially the HVAC and the refrigeration systems; this is also the main motivation of us to conducted this study.

**Table 1** Supermarket data

Latitude	39.52
Longitude	112.82
Total area (m <sup>2</sup> )	1970
Sales area (m <sup>2</sup> )	1598
Opening time	8:00–20:00
Weekly trading hours (h)	70
Total length of ref. display cabinet (m)	33.75
Total length of freezers (m)	30
Installed lighting (kW)	33.94

### 3.2 Data description and pre-processing

Great efforts had been made to prepare the required research data. The energy data was acquired from the building energy consumption monitoring systems, and the input variables were obtained from the weather forecast website (i.e., a website which covered various weather information) from January 2014 to March 2019 regarding the supermarket location. This generated data contains 40 types of weather parameters due to this paper that focuses on researching the performance of climate factors on the electrical usage, and the dataset was divided into 12 categories which was detailed in Table 2.

It is essential to conduct the data pre-processing before deploying the research. The max-min normalization method was utilized to deal with the values of the dataset. All the values were scaled into interval [0–1] by the line transformation which was expressed by the following formula:

$$x' = \frac{x - \min}{\max - \min} \tag{1}$$

where  $x$  denotes the original value of the variables, and  $x'$  is the normalized data through the formula above.

What is more, due to the max visibility (vis-max) and the min precipitation (pre-min) that were always constant and irrelevant with the electrical energy changes, they would not be considered into the further analysis. The following section will introduce the involved methodologies in this paper.

**Table 2** The categories of the input variables

Categories	Climate factors	Abbreviation	Value range	Categories	Climate factors	Abbreviation	Value range
Max temperature (°C)	Maximum	Max-T-max	5~36	Humidity (%)	Maximum	Hum-max	30~98
	Average	Max-T-avg	– 2~29		Average	Hum-avg	12~63
	Minimum	Max-T-min	– 16~21		Minimum	Hum-min	4~22
Avg temperature (°C)	Maximum	Avg-T-max	– 1~30	Sea level pressure (hPa)	Maximum	SLP-max	1011~1052
	Average	Avg-T-avg	– 7~25		Average	SLP-avg	1004~1035
	Minimum	Avg-T-min	– 17~20		Minimum	SLP-min	992~1023
Min temperature (°C)	Maximum	Min-T-max	– 6~27	Visibility (km)	Maximum	Vis-max	30
	Average	Min-T-avg	– 11~21		Average	Vis-avg	8~25
	Minimum	Min-T-min	– 18~18		Minimum	Vis-min	0.1~7
Dew point (°C)	Maximum	DP-max	– 14~23	Wind (km/h)	Maximum	W-max	14~40
	Average	DP-avg	– 23~18		Average	W-avg	5~10
	Minimum	DP-min	– 44~11		Minimum	W-min	0
Precipitation (mm)	Maximum	Pre-max	0~49	Growing degree days (base 50) (days)	Maximum	GDD-max	0~35
	Average	Pre-avg	0~4.50		Average	GDD-avg	0~27
	Minimum	Pre-min	0~143.6		Minimum	GDD-min	0~18
	Sum	Pre-sum	–		Sum	GDD-sum	0~839
Heating degree days (base 65) (days)	Maximum	HDD-max	0~63	Cooling degree days (base 65) (days)	Maximum	CDD-max	0~20
	Average	HDD-avg	0~45		Average	CDD-avg	0~12
	Minimum	HDD-min	0~35		Minimum	CDD-min	0~3
	Sum	HDD-sum	0~1391		Sum	CDD-sum	0~376

### 3.3 Methodologies

Three machine learning methods were involved in this work to perform the dataset: the methods respectively are random forest (RF), gradient boosting regression (GBR), and support vector regression (SVR), and they were all performed well in predicting the energy consumption; however, the SVR was superior to capture the mapping relationship between input variables and target variable in small-scale dataset, and SVR was recommended in our work to perform the dataset. RF and GBR were also introduced to compare with the performance of the SVR. The theories of three machine learning methods were detailed in the following.

#### 3.3.1 Random forest (RF)

The random forest (RF) is a popular ensemble algorithm, and it was started by Ho et al. [43] and was developed in literatures [44, 45] by Breiman. The random forest is integrated by decision tree, and RF has the advantage of overcoming the likely drawbacks of the single decision tree.

#### 3.3.2 Gradient boosting regression (GBR)

The gradient boosting regression (GBR) produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees [42]. More details on the main mathematical principles of the gradient boosting regression algorithm were given in [46].

#### 3.3.3 Support vector regression (SVR)

The support vector regression (SVR) is the most common application form of SVMs which was proposed in 1997 by Vapnik et al. [47]. Smola et al. have overviewed the development of support vector regression (SVR) and illustrated the basic idea of support vector regression (SVR) in 2003 [48]. Additionally, Basak et al. have also made an attempt to review the existing theory, methods, recent developments, and scopes of SVR in 2007 [49]. The core idea of SVR is mapping the original input variables to high dimensional feature space and to find the nonlinear relationship between input variables and target variables. Assuming the training data was  $(x_1, y_1) \dots (x_n, y_n)$ ,  $n$  is the number of training dataset. The regression function of SVR is briefly expressed as follows:

$$f(x) = w \cdot \phi(x) + b \quad (2)$$

where  $f(x)$  denotes the forecasting values,  $x$  is the input variable,  $w$  is the weight coefficient,  $b$  is the deviation value, and  $\phi(x)$  is the high dimensional feature space. The  $w$  and  $b$  were estimated by a regularized risk function, and the expression was presented as follows:

$$\frac{1}{2} \|w\|^2 + C \frac{1}{n} \sum_{i=1}^n L_\varepsilon(y_i, f(x_i)) \quad (3)$$

where  $\|w\|^2$  is a regularized term,  $C$  is the penalty parameter,  $C \frac{1}{n} \sum_{i=1}^n L_\varepsilon(y_i, f(x_i))$  is the empirical error and is measured by the  $\varepsilon$ -insensitive loss function, and the expression was as follows:

$$L_\varepsilon(y_i, f(x_i)) = \begin{cases} |y_i - f(x_i)| - \varepsilon, & |y_i - f(x_i)| \geq \varepsilon \\ 0, & \text{others} \end{cases} \quad (4)$$

where  $\varepsilon$  is the insensitive loss coefficient, and it represents a  $\varepsilon$  tube; if the predicted value is within the tube, the loss is zero, while if it is outside the tube, the loss is the magnitude of the difference between the predicted value and the radius  $\varepsilon$  of the tube. To estimate  $w$  and  $b$ , the above equation is transformed into the primal objective function, and the expression was given as follows:

$$\min \frac{1}{2} \|w\|^2 + C \frac{1}{n} \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (5)$$

where the  $\xi_i$  and  $\xi_i^*$  are the slack variables, the problem of Eq. (5) can be solved in its dual formulation as presented in Eq. (6), and  $\alpha_i$  and  $\alpha_i^*$  are the Lagrange multipliers.

$$\max w(\alpha_i, \alpha_i^*) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) y_i - \frac{1}{2} \sum_{i,j=1}^n (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) \cdot K(x_i, x_j) - \sum_{i=1}^n (\alpha_i + \alpha_i^*) \varepsilon \quad (6)$$

Equation (2) can be rewritten as Eq. (7), where  $K(x_i, x_j)$  was represented as the kernel function.

$$f(x) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) \cdot K(x_i, x_j) + b \quad (7)$$

The radial basis function (RBF) was used as the kernel function in this paper, and Eq. (2) can be further denoted as follows:

$$f(x) = \sum_{i=1}^n \left[ (\alpha_i - \alpha_i^*) \times \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \right] + b \quad (8)$$

Where  $\sigma$  represented the width parameter of RBF, and  $i$  and  $j$  are different samples. The RBF was the widely used function in kernel functions, and it could map the original input variables to high dimensional feature space and facilitate to find the non-linear relationship between climate factors and building energy consumption.

#### 4 Model development

It is well known that overlarge or too few input variables could make a decrease of accuracy in model prediction. This problem was studied for a long time, and many algorithms were developed to conduct feature selection (FS). We select three widely used feature selection (FS) methods to extract the optimal size feature subset, respectively: stepwise, Lars, and Boruta methods. These FS methods will be introduced in the following part, and they would be further combined with three involved machine learning methods to improve the performance of the model.

##### 4.1 Feature selection

The feature selection technologies have been explored by many people for a long time thanks to its well performance for improving the accuracy of the energy forecasting [24]. In particular, the feature selection could well estimate the importance of characteristic features, relevance, dependencies, weighting, and ranking [50]. We have applied three FS methods to conduct feature selection, respectively, stepwise, least angle regression (Lars), and Boruta algorithms, and the performance of Lars algorithm for the

feature filtering was detailed in literatures [26, 27]. Also, the characteristics of the stepwise method will be found in literature [25]. The Boruta algorithm was also employed in this paper to implement the feature filtering and attempted to extract all relevant variables of the database [30]. Candanedo et al. have successfully discarded two random variables which are deliberately added into the existing dataset, and the results have proved the validity of the Boruta algorithm [29]. The Boruta algorithm was implemented by a R package Boruta, and the feature selection process of Boruta methods in this research was exhibited in Fig. 1. Red and green boxplots respectively represent the rejected and confirmed attributes [30]. Actually, the Boruta algorithm is a wrapper built around the random forest classification algorithm [51], and the random forest runs during the Boruta can be stopt by the argument maxRuns. The parameter maxRuns was set to 1000 in this research. The ranking of the variable importance regarding the three feature methods above was exhibited in Fig. 2.

The prediction accuracy varies greatly with the number of variables. Hence, it is essential to seek the optimal size of the feature subset. Fig. 3 has present the prediction performance of various feature subsets with different size. It is worth noting that the prediction accuracy of the hybrid model varies greatly when the features number changing. The highest accuracy of each model will be selected, and the corresponding feature size will be taken as the optimal size of the model; the final selection results were specified in Table 3.

### 4.2 Parameter optimization

Parameter optimization is a vital step in model processing. In order to obtain the optimal prediction performance, a lot of time was paid to optimize the parameters of each model, and the optimal parameter was found in a certain range which was determined based on the prior knowledge of statistics; the results were presented in Table 4. In the RF algorithm, the “n\_estimators” and “max\_depth” were adjusted with 10 and 7 values. In the GBR algorithm, the “max\_depth,” “learning\_rate,” and “n\_estimators” were adjusted with values 7, 0.19, and 40. In the SVR algorithm, the kernel function has three types, respectively, and these are linear, polynomial, and radial basis function

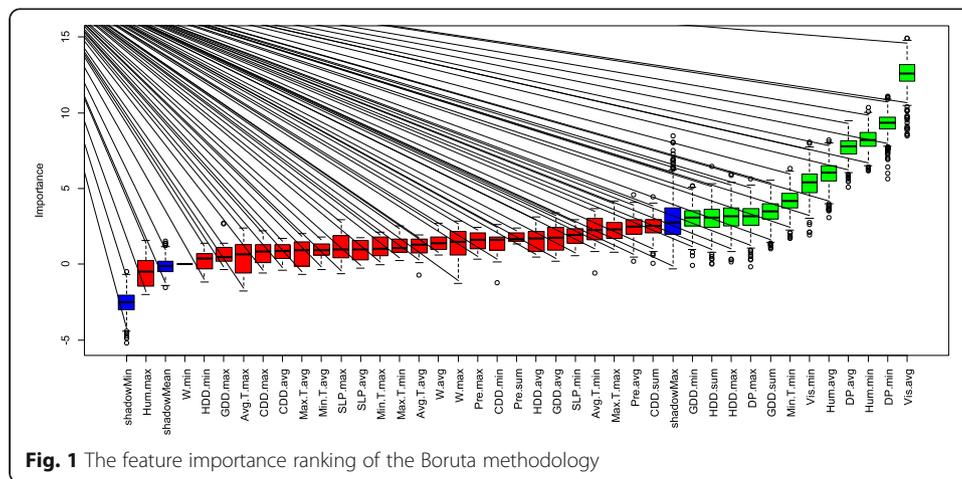
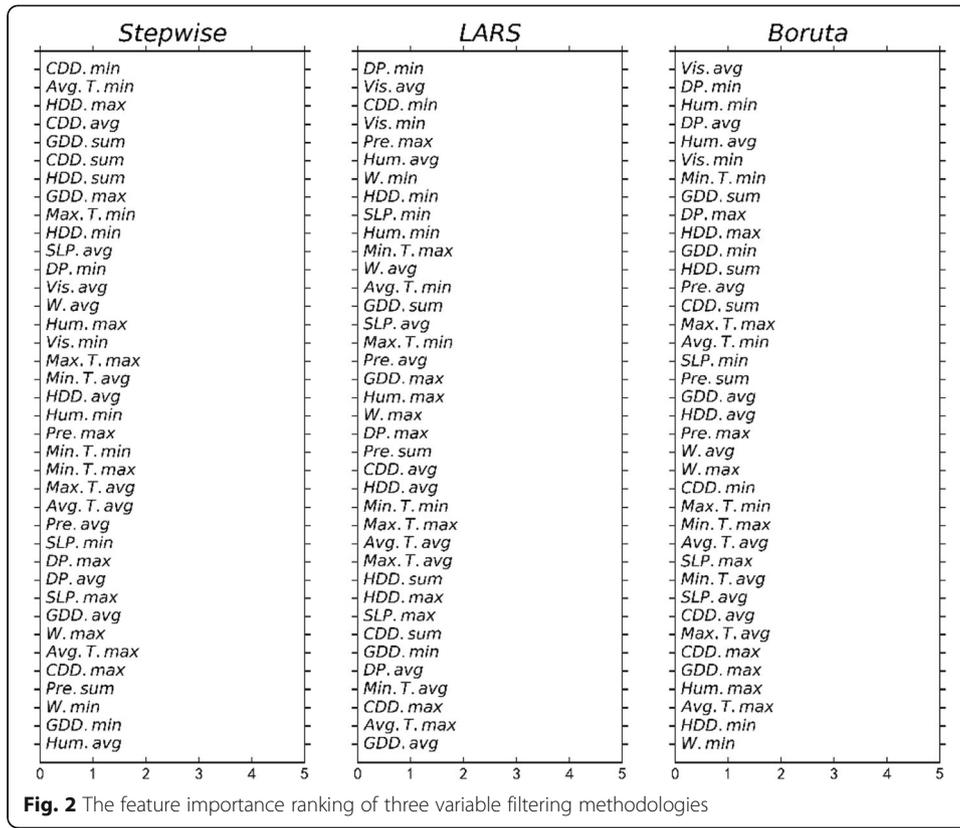


Fig. 1 The feature importance ranking of the Boruta methodology



(RBF); the kernel function adopted in this paper was RBF. In addition, the penalty parameter “C” and insensitive loss coefficient “ε” of the SVR model were adjusted with values 150 and 0.08, the parameter “γ” was set to 0.1, and the remaining arguments are default.

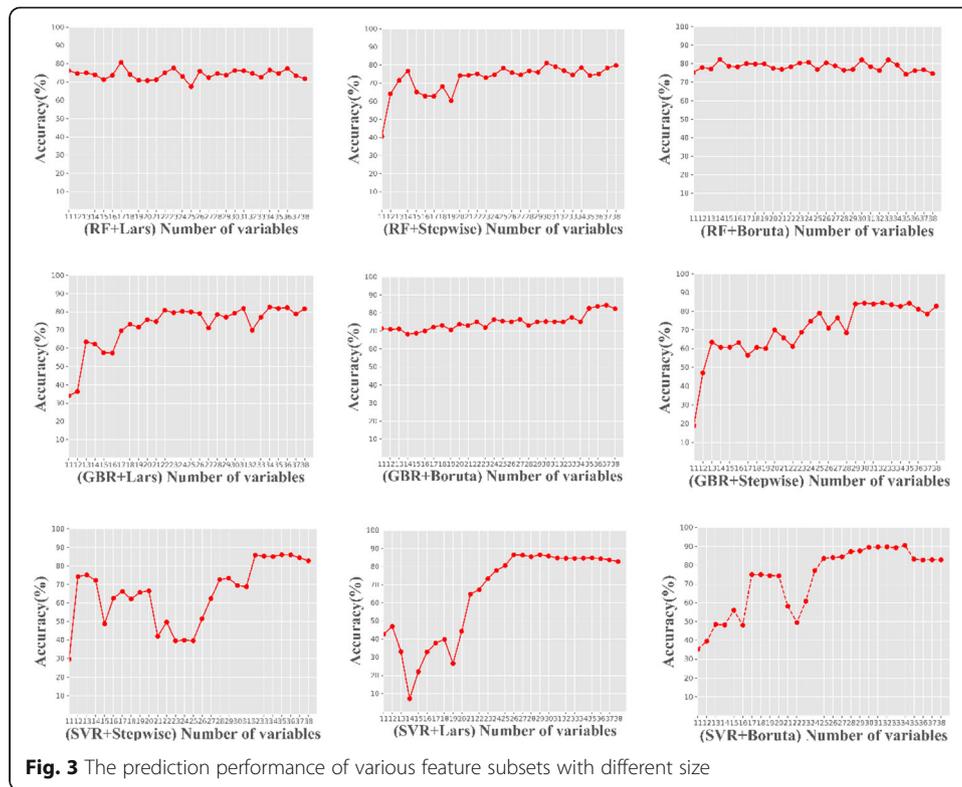
### 4.3 Evaluation indicators

In this study, we have chosen four evaluation indicators to conduct the performance comparison of the models. The error indicators are the mean absolute error (MAE), the mean squared error (MSE), the root mean squared error (RMSE), and the R-squared ( $R^2$ ). The R-squared ( $R^2$ ) was used to assess the predict accuracy of the models. The formulas were expressed as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \tag{9}$$

$$MSE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|^2 \tag{10}$$

$$RMSE = \sqrt{\frac{1}{2} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \tag{11}$$



$$R^2 = 1 - \frac{\sum_{i=1}^n |y_i - \hat{y}_i|^2}{\sum_{i=1}^n |y_i - \bar{y}_i|^2} \tag{12}$$

where the  $y_i$  is the real output variable, the  $\hat{y}_i$  is the predicted output variable, the  $\bar{y}_i$  is the average value of real output variable, and  $n$  represents the amounts of samples in the testing set.

### 5 Results and discussion

The proposed hybrid model was implemented by the software JetBrains PyCharm 2019 and with the system environment of Intel Core i5-8300H and 8.00GB RAM. In this paper, the hybrid model of FS algorithms and machine learning methods was applied to perform the energy consumption prediction. The optimal feature subset was

**Table 3** The optimal feature size of nine hybrid models

Models	Optimal size of feature subset
RF + Lars	17
RF + stepwise	30
RF + Boruta	30
GBR + Lars	34
GBR + Boruta	37
GBR + stepwise	32
SVR + stepwise	35
SVR + Lars	26
SVR + Boruta	34

**Table 4** Optimized parameters of RF, GBR, and SVR

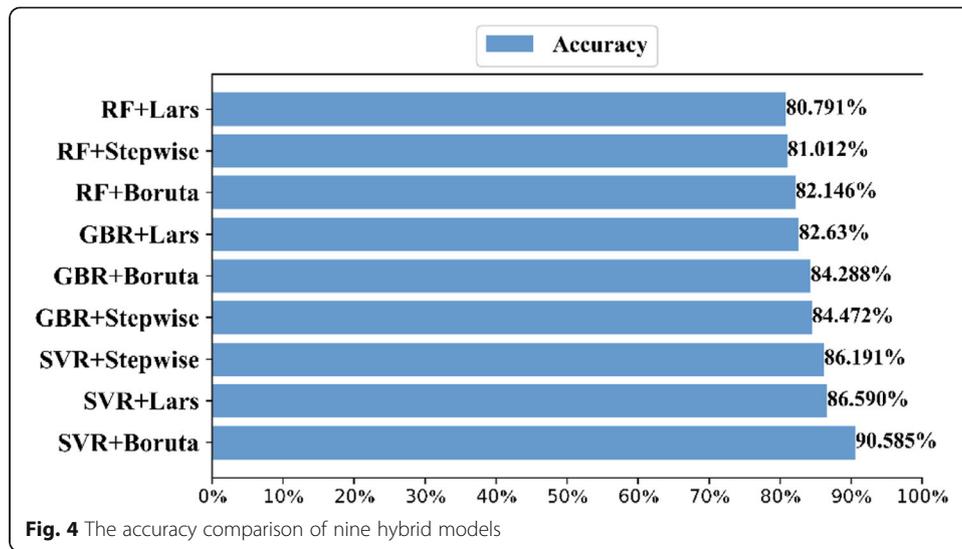
Methods	Parameters	Parameter description	Terminal values
RF	n_estimators	The number of trees in the forest	10
	max_depth	The depth of a tree	7
GBR	max_depth	The depth of a tree	7
	learning_rate	The learning rate	0.19
	n_estimators	The number of trees	40
SV	C	The penalty parameter	150
	$\epsilon$	Insensitive loss coefficient	0.08
	$\gamma$	$\gamma$ defines how much influence a single training example has. The larger the $\gamma$ is, the closer other example must be to be affected.	0.1

extracted by FS algorithms in advance, totally nine hybrid models were mentioned in this study, and the dataset was split into 83% for training and 17% for testing. The prediction results of nine hybrid models were displayed in Table 5, and four evaluation indicators were utilized to measure the prediction performance. The error indicators are MAE, MSE, and RMSE, and smaller errors generally indicate that smaller deviation was detected between prediction and real values.  $R^2$  is always used to measure the fitting degree of the model, and the closer the value is to 1, the better the performance of the model is.

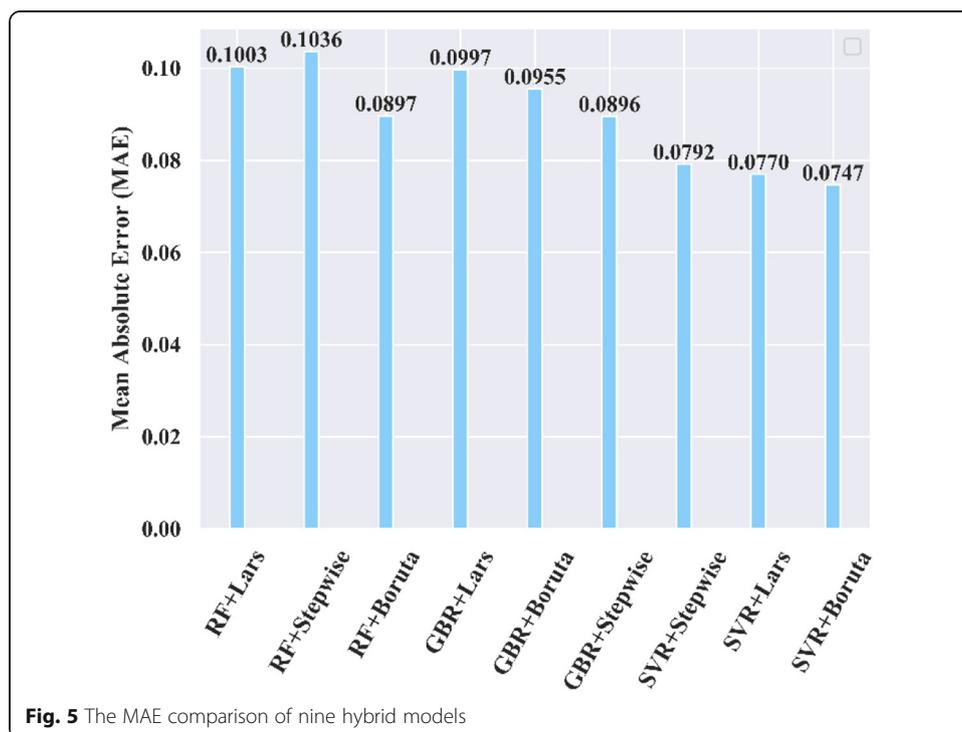
Figure 4 has exhibited the accuracy of nine hybrid models graphically using horizontal bar chart; as can be seen from Fig. 4, the hybrid model SVR-Boruta is the best prediction model with an accuracy of 90.585%, and next are hybrid model SVR-Lars and SVR-stepwise, with values of accuracy 86.59% and 86.191%. The error comparison of nine hybrid models was represented in Figs. 5, 6, and 7, and it is found that the smallest errors were also obtained in hybrid model SVR-Boruta, which further verified the superiority of hybrid model SVR-Boruta in forecast aspect. What is more, we observed that in terms of machine learning methods, the prediction performance of SVR was always higher than GBR and RF algorithms in this study, and it is possibility due to the size of dataset used in this paper that is small and with high dimension. The

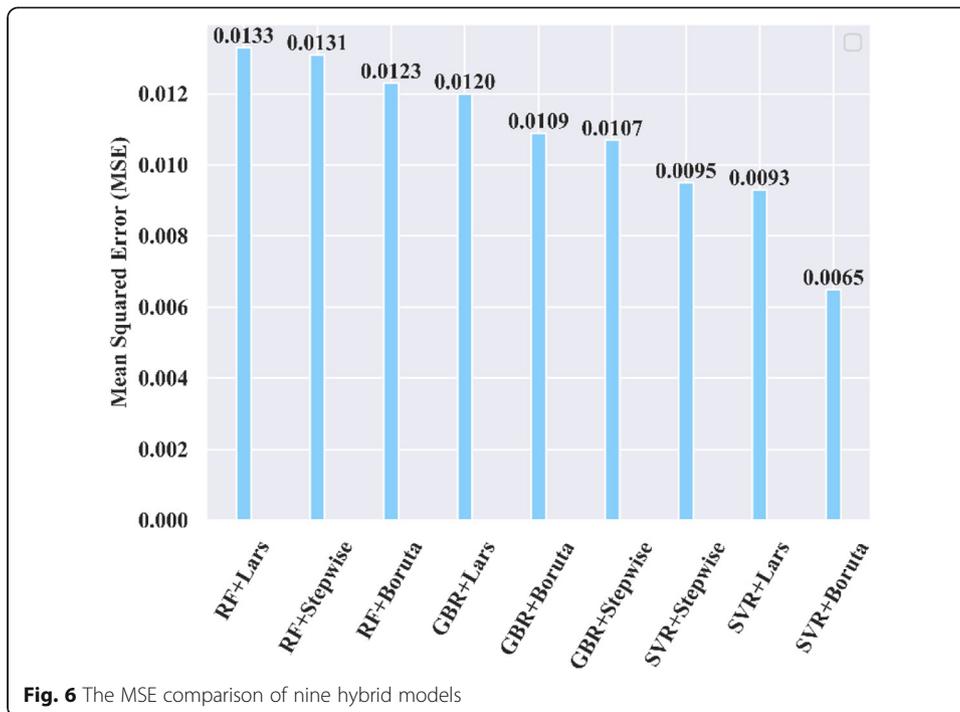
**Table 5** The prediction performance of nine different hybrid models

Models	MAE	MSE	RMSE	$R^2$ (%)
RF + Lars	0.100304	0.013272	0.115202	80.791
RF + stepwise	0.103614	0.013119	0.114538	81.012
RF + Boruta	0.089666	0.012335	0.111064	82.146
GBR + Lars	0.099679	0.012001	0.109548	82.63
GBR + Boruta	0.095545	0.010855	0.104188	84.288
GBR + stepwise	0.089575	0.010729	0.103579	84.472
SVR + stepwise	0.079161	0.009541	0.097676	86.191
SVR + Lars	0.077012	0.009265	0.096255	86.590
SVR + Boruta	0.074670	0.006505	0.080651	90.585



second good performance was obtained by the GBR algorithm, and the worst is the RF algorithm. In terms of FS algorithms, it can be seen from the Fig. 8 that the FS method Boruta works well on both SVR and RF algorithms, but with respect to GBR algorithm, the FS method stepwise works better than Boruta and Lars method. Therefore, we took SVR algorithms as the recommended machine learning, and the Boruta will be taken as the best FS method. Furthermore, the hybrid model SVR-Boruta was the best prediction model in this study, and it could be applied in energy intelligent system to improve the performance of the energy forecasting.

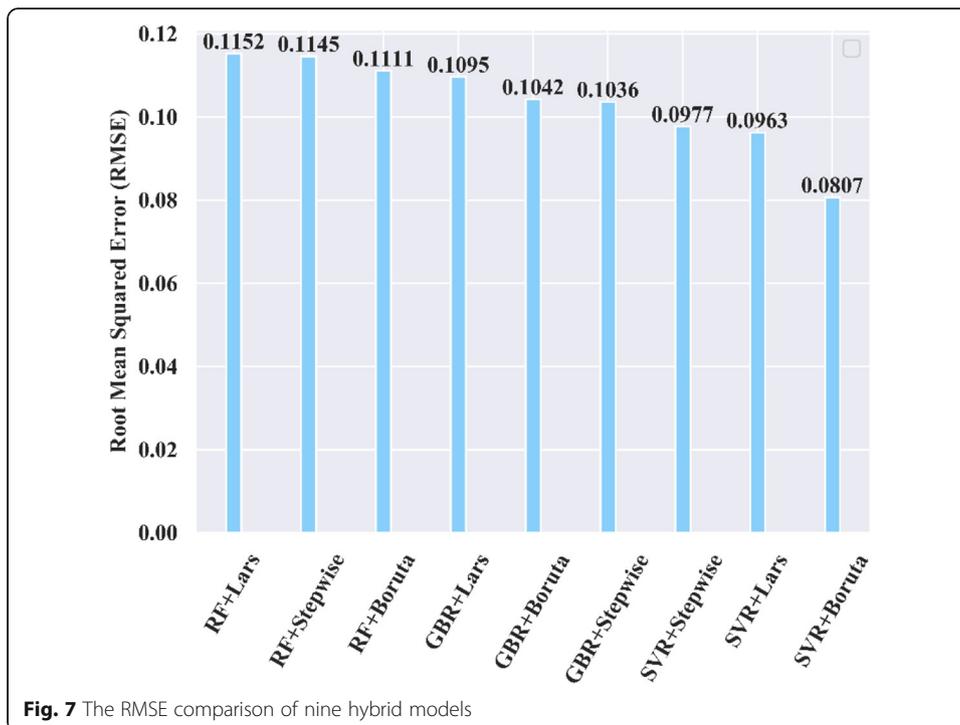




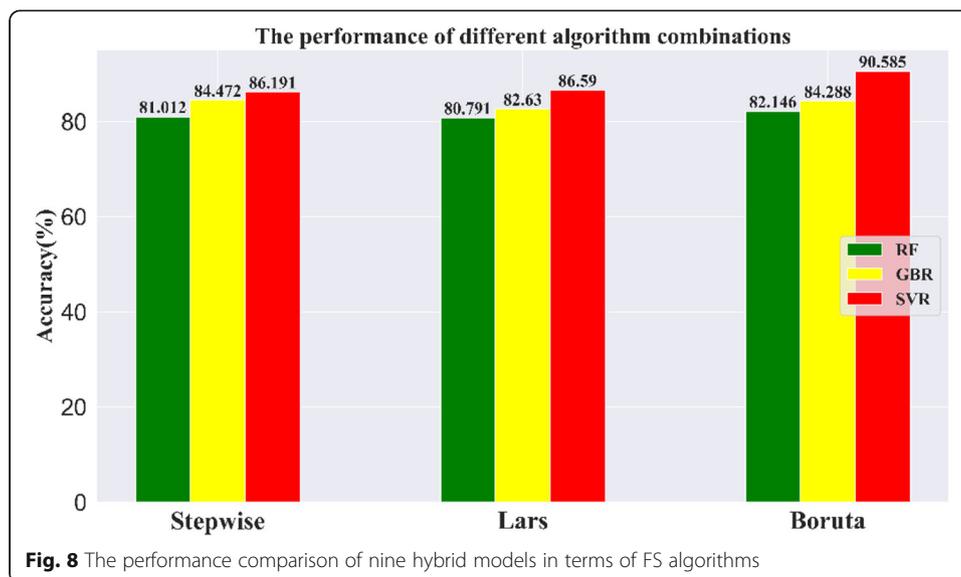
**Fig. 6** The MSE comparison of nine hybrid models

### 6 Conclusions and future work

This paper has proposed a new hybrid model to improve the performance of energy management system. In this paper, the FS algorithms and machine learning methods were integrated to develop the hybrid model. The building process of hybrid model contains two stages. In the first stage, optimal feature subset will be extracted using FS



**Fig. 7** The RMSE comparison of nine hybrid models



algorithms, respectively, stepwise, Lars, and Boruta methods, and it is implemented in the  $R$  environment. In the second stage, optimal feature subset will be used to conduct model learning, RF, GBR, and SVR algorithms that were applied in this study. The efficiency of the models was examined by measures of MAE, MSE, RMSE, and  $R^2$ , and the results presented that the SVR-Boruta has the best performance on energy prediction, with values of accuracy 90.585%. What is more, we also noted that the SVR was the best machine learning in this study, and Boruta was the better FS algorithm. The results of this study were further verified the importance of feature selection and the superiority predictive ability of the hybrid model. The recommended predictive model SVR-Boruta could be employed to conduct energy prediction according to the weather forecast and provide constructive suggestions for the energy manager to make a more reasonable decision in energy distribution.

In our future work, we will pay more attention and time to obtain more richer and larger research data to develop the predictive models. Besides, more accurate hybrid model is needed to be explored in the future. This research was conducted in the northwest region of China regarding the supermarket buildings, and we will also consider to investigate more types of buildings in other climate areas of China with aims to develop a more intelligent energy system.

#### Abbreviations

WSN: Wireless sensor network; HVAC: Heating, ventilation and air conditioning; FS: Feature selection; Lars: Least angle regression; RF: Random forest; GBR: Gradient boosting regression; SVR: Support vector regression; RBF: Radial basis function; MAE: Mean absolute error; MSE: Mean squared error; RMSE: Root mean squared error;  $R^2$ :  $R$ -squared

#### Acknowledgements

The authors thank the person who provided meticulous and valuable suggestions for improving the paper.

#### Authors' contributions

Ni Guo came up with the prediction model of improving support vector regression, Weifeng Gui, Zijian Tian, and Wei Chen analyzed the measurement results, Xin Tian, and Weiguo Qiu have given help in experiment process, Xiangyang Zhang, Zijian Tian, Wei Chen, and Ni Guo were the major contributors in writing the manuscript and Weifeng Gui did a lot of work in the process of manuscript revision. All authors read and approved the final manuscript.

### Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 51674269, in part by the National Natural Science Foundation of China under Grant 51874300 and 51874299, in part by the National Natural Science Foundation of China and Shanxi Provincial People's Government Jointly Funded Project of China for Coal Base and Low Carbon under Grant U1510115 and in part by the Open Research Fund of Key Laboratory of Wireless Sensor Network and Communication, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, under Grant 20190902.

### Availability of data and materials

The research data used to support the finding of this study are available from the corresponding author upon request.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>School of Mechanical Electronic and Information Engineering, China University of Mining and Technology (Beijing), No.11, Xueyuan Road, Haidian District, Beijing 100083, China. <sup>2</sup>School of Physics and Electronic Information Engineering, Henan Polytechnic University, Shiji Road 2001, Jiaozuo City 454003, Henan, China. <sup>3</sup>School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China. <sup>4</sup>Mine Digitization Engineering Research Center of the Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China. <sup>5</sup>Information Engineering College, Beijing Institute of Petrochemical Technology, Beijing 102617, China. <sup>6</sup>Beijing Science and Technology Institute of Housing and Urban-Rural Development, No.9 Daji Street, Tongzhou District, Beijing 101160, China.

Received: 8 January 2020 Accepted: 14 May 2020

Published online: 08 June 2020

### References

1. H. Cheng, D. Feng, X. Shi et al., "Data quality analysis and cleaning strategy for wireless sensor networks," *EURASIP J. Wirel. Commun. Netw.*, vol. 2018, no. 1, pp. 1-11, 2018.
2. N. Liu, and J.-S. Pan, "A bi-population QUasi-Affine TRansformation Evolution algorithm for global optimization and its application to dynamic deployment in wireless sensor networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, pp. 175, 2019.
3. H. Cheng, Z. Xie, L. Wu et al., "Data prediction model in wireless sensor networks based on bidirectional LSTM," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, pp. 203, 2019.
4. H. Cheng, Z. Su, N. Xiong et al., "Energy-efficient node scheduling algorithms for wireless sensor networks using Markov Random Field model," *Information sciences*, vol. 329, pp. 461-477, 2016.
5. X. Chen, J. Chen, B. Liu et al., "AndroidOff: Offloading android application based on cost estimation," *J. Syst. Softw.*, vol. 158, pp. 110418, 2019.
6. R. Cheng, W. Yu, Y. Song et al., "Intelligent safe driving methods based on hybrid automata and ensemble cart algorithms for multihigh-speed trains," *IEEE Trans Cybernetics*, vol. 49, no. 10, pp. 3816-3826, 2019.
7. R. Cheng, D. Chen, B. Cheng et al., "Intelligent driving methods based on expert knowledge and online optimization for high-speed trains," *Expert Syst. Appl.*, vol. 87, pp. 228-239, 2017.
8. C. Parmesan, T.L. Root, and M.S. Willig, "Impacts of extreme weather and climate on terrestrial biota," vol. 81, no. 3, pp. 443-450, 2000.
9. C. Rosenzweig, A. Iglesias, X. Yang et al., "Climate change and extreme weather events; implications for food production, plant diseases, and pests," vol. 2, no. 2, pp. 90-104, 2001.
10. G. A. Meehl, T. Karl, D. R. Easterling et al., "An introduction to trends in extreme weather and climate events: observations, socioeconomic impacts, terrestrial ecological impacts, and model projections," vol. 81, no. 3, pp. 413-416, 2000.
11. T. Stocker, *Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change: Cambridge University Press* (2014)
12. R. K. Pachauri, M. R. Allen, V. R. Barros et al., *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change: Ippc*, 2014.
13. P. A. Stott, N. Christidis, F. E. Otto et al., "Attribution of extreme weather and climate-related events," *Wiley Interdiscip. Rev. Clim. Chang.*, vol. 7, no. 1, pp. 23-41, Jan, 2016.
14. C.-L. Lo, C.-H. Chen, J.-L. Hu et al., "A Fuel-Efficient Route Plan Method Based on Game Theory," *J Internet Technol*, vol. 20, no. 3, pp. 925-932, 2019.
15. C.-Y. Zhang, D. Chen, J. Yin et al., "A flexible and robust train operation model based on expert knowledge and online adjustment," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 15, no. 03, pp. 1750023, 2017.
16. C.-H. CHEN, "A cell probe-based method for vehicle speed estimation," *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 103, no. 1, pp. 265-267, 2020.
17. S.-M. Deng, J. J. E. Burnett, and Buildings, "A study of energy performance of hotel buildings in Hong Kong," vol. 31, no. 1, pp. 7-12, 2000.
18. K. Amasyali, and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renew. Sust. Energ. Rev.*, vol. 81, pp. 1192-1205, 2018.
19. M. Kolokotroni, Z. Mylona, J. Evans et al., "Supermarket Energy Use in the UK," vol. 161, pp. 325-332, 2019.
20. A. Behfar, D. Yuill, and Y. Yu, "Supermarket system characteristics and operating faults (RP-1615)," *Sci Technol Built Environ*, vol. 24, no. 10, pp. 1104-1113, 2018.

21. M. R. Braun, S. B. M. Beck, P. Walton et al., "Estimating the impact of climate change and local operational procedures on the energy use in several supermarkets throughout Great Britain," *Energy Buildings*, vol. 111, pp. 109-119, 2016.
22. J. C. Lam, K. K. W. Wan, D. Liu et al., "Multiple regression models for energy use in air-conditioned office buildings in different climates," *Energy Convers. Manag.*, vol. 51, no. 12, pp. 2692-2697, 2010.
23. Y. Wang, F. Wang, and H. J. P. E. Wang, "Influencing factors regression analysis of heating energy consumption of rural buildings in China," vol. 205, pp. 3585-3592, 2017.
24. B. Venkatesh, and J. Anuradha, "A Review of Feature Selection and Its Methods," *Cybernetics Inform Technol*, vol. 19, no. 1, pp. 3-26, 2019.
25. B. Thompson, "Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial," Sage Publications Sage CA: Thousand Oaks, CA, 1995.
26. T. Hesterberg, N. H. Choi, L. Meier et al., "Least angle and  $\ell_1$  penalized regression: A review," *Statistics Surveys*, vol. 2, no. 0, pp. 61-93, 2008.
27. B. Efron, T. Hastie, I. Johnstone et al., "Least angle regression," vol. 32, no. 2, pp. 407-499, 2004.
28. R Tibshirani, "Regression shrinkage and selection via the lasso," vol. 58, no. 1, pp. 267-288, 1996.
29. L. M. Candanedo, V. Feldheim, and D. Deramaix, "Data driven prediction models of energy use of appliances in a low-energy house," *Energy Buildings*, vol. 140, pp. 81-97, 2017.
30. M. B. Kursu, and W. R Rudnicki, "Feature selection with the Boruta package," vol. 36, no. 11, pp. 1-13, 2010.
31. C.-H. Chen, F. Song, F.-J. Hwang et al., "A probability density function generator based on neural networks," *Physica A: Statistical Mechanics and its Applications*, pp. 123344, 2019.
32. C.-H. Chen, F.-J. Hwang, and H.-Y. Kung, "Travel time prediction system based on data clustering for waste collection vehicles," *IEICE Trans. Inf. Syst.*, vol. 102, no. 7, pp. 1374-1383, 2019.
33. C.-H. Chen, "An arrival time prediction method for bus system," *IEEE Internet Things J*, vol. 5, no. 5, pp. 4231-4232, 2018.
34. X. Zheng, D. An, X. Chen et al., "Interest prediction in social networks based on Markov chain modeling on clustered users," *Concurrency and Computation: Practice and Experience*, vol. 28, no. 14, pp. 3895-3909, 2016.
35. Z. Ma, C. Ye, and W. J. E. P. Ma, "Support vector regression for predicting building energy consumption in southern China," vol. 158, pp. 3433-3438, 2019.
36. Y. Chen, P. Xu, Y. Chu et al., "Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings," vol. 195, pp. 659-670, 2017.
37. H. Guo, B. Liu, D. Cai et al., "Predicting protein-protein interaction sites using modified support vector machine," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 3, pp. 393-398, 2018.
38. G. P. Herrera, M. Constantino, B. M. Tabak et al., "Long-term forecast of energy commodities price using machine learning," vol. 179, pp. 214-221, 2019.
39. M. Yahşi, E. Çanakoğlu, and S. J. C. M. Ağralı, "Carbon price forecasting models based on big data analytics," vol. 10, no. 2, pp. 175-187, 2019.
40. R. Wang, S. Lu, Q. J. S. C. Li et al., "Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings," vol. 49, pp. 101623, 2019.
41. C. Li, Y. Tao, W. Ao et al., "Improving forecasting accuracy of daily enterprise electricity consumption using a random forest based on ensemble empirical mode decomposition," vol. 165, pp. 1220-1227, 2018.
42. A. Chokor, and M. El Asmar, "Data-Driven Approach to Investigate the Energy Consumption of LEED-Certified Research Buildings in Climate Zone 2B," *Journal of Energy Engineering*, vol. 143, no. 2, 2017.
43. I. J. I. t. o. p. a. Barandiaran, and m. intelligence, "The random subspace method for constructing decision forests," vol. 20, no. 8, 1998.
44. L. J. M. I. Breiman, "Random forests," vol. 45, no. 1, pp. 5-32, 2001.
45. L. J. S. D. U. o. C. B. Breiman, CA, USA, "Manual on setting up, using, and understanding random forests v3. 1," vol. 1, 2002.
46. J. H. J. A. o. s. Friedman, "Greedy function approximation: a gradient boosting machine," pp. 1189-1232, 2001.
47. V. Vapnik, S. E. Golowich, and A. J. Smola, "Support vector method for function approximation, regression estimation and signal processing," pp. 281-287.
48. A. J. Smola, B. J. S. Schölkopf, and computing, "A tutorial on support vector regression," vol. 14, no. 3, pp. 199-222, 2004.
49. D. Basak, S. Pal, D. C. J. N. I. P.-L. Patrnanabis et al., "Support vector regression," vol. 11, no. 10, pp. 203-224, 2007.
50. U. Stańczyk, L.C. Jain, *Feature selection for data and pattern recognition: Springer* (2015)
51. A. Liaw, and M Wiener, "Classification and regression by randomForest," vol. 2, no. 3, pp. 18-22, 2002.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.