Hindawi
Geofluids
Volume 2021, Article ID 7271383, 12 pages
https://doi.org/10.1155/2021/7271383

Research Article

Research on the Combined Prediction Model of Residential Building Energy Consumption Based on Random Forest and BP Neural Network

Xuenan Zhang,[1,2] Jinxin Zhang ,[1,2] Jinhua Zhang ,[3] and YuChuan Zhang[1]

[1]*Business School, Hubei University, Wuhan 430062, China*
[2]*Research Center for China Agriculture Carbon Emission Reduction and Carbon Trading, Hubei University, Wuhan 430062, China*
[3]*School of Economics and Management, Fuzhou University, Fuzhou 350108, China*

Correspondence should be addressed to Jinxin Zhang; zhangjinxin@hubu.edu.cn

Received 6 April 2021; Revised 24 June 2021; Accepted 21 July 2021; Published 26 September 2021

Academic Editor: Zhang Xudong

As the energy consumption of residential building takes a large part in the building energy consumption, it is important to promote energy efficiency in residential building for green development. In order to evaluate the energy consumption of residential building more effectively, this paper proposes a combined prediction model based on random forest and BP neural network (RF-BPNN). To verify the prediction effect of the RF-BPNN combined model, experiments were performed by using the energy efficiency data set in the UCI database, and the model was evaluated with five indicators: mean absolute error, root mean square deviation, mean absolute percentage error, correlation coefficient, and coincidence index. Compared with the random forest, BP neural network model, and other existing models, respectively, it is proven by the experimental results that the RF-BPNN model possesses higher prediction accuracy and better stability.

1. Introduction

Global warming has become an important environmental problem that needs to be solved urgently worldwide. Some data shows that the average temperature has increased by 0.74°C in the last 10 decades, and the temperature increase rate in the past 5 decades is twice that of the past 100 years [1]. Climate change severely affects human life and health. The main cause of climate change is that human beings emit a large amount of greenhouse gases into the air, and the sources of these gases are mainly the energy consumption of transportation, industry, and building. With the development of urbanization, the energy consumption of construction has increased significantly. According to the report *Buildings and Climate Change: Summary for Decision-Makers* published by the United Nations Environment Programme in 2009, building energy consumption accounts for 40% of all global energy consumption, and one-third of global greenhouse gas emission is related to building energy consumption [2]. It is reported that building energy consumption in Europe and North America has increased at a rate of 1.5% and 1.9%, respectively, from 1999 to 2004 [3]. In China, the growth rate in building energy consumption is even more dramatic with an annual growth rate of 10% in the past 2 decades [4]. The increasing building energy consumption has a serious impact on human's living environment. One of the ways to reduce the extra energy consumption effectively is to adopt a building design that is energy efficient so that the interior environment of the construction is comfortable and the energy consumption is effectively reduced. In *The 13th Five-Year Plan for Economic and Social Development of the People's Republic of China*, China has determined that the green building area will be increased by more than 2 billion square meters at the end of the period of 2016 to 2020; therefore, the building industry, which is one of the largest energy-consuming industries, is facing a great challenge [5].

Buildings include commercial buildings, office buildings, and residential buildings. The unit energy consumption area of residential buildings is small, but the total amount is large;

Table 1: Comparison of the advantages and disadvantages of machine learning algorithms.

Algorithm	Advantages	Disadvantages
DT [17]	Simple structure; suitable for handling large amount of data; fast running speed	Not easy to deal with missing data and prone to overfitting; ignore the association between attributes in the data set
KNN [17]	No requirement for data distribution, faster training phase	Not easy to find the relationship between features; large calculation amount and slow speed
SVM [18]	Solve small sample and nonlinear problems; better handling of high-dimensional data; better generalization ability	Poor interpretation of the high-dimensional mapping ability of kernel functions, especially radial basis kernel functions; more sensitive to missing data values; longer training time
BPNN [17, 18]	Strong learning ability; strong robust and fault-tolerant to noisy data; can handle nonlinear problems well	Difficult to determine the network structure; more parameters; objectiveness of the selection of training data
RF [19]	Can handle higher dimensional problems with higher prediction accuracy; insensitive to noisy data and less prone to overfitting	Belong to the black box model; difficult to explain the internal operation mechanism

thus, it cannot be ignored. For residential buildings, energy consumption is mainly due to the use of heating, ventilation, and air conditioning systems (HVAC) [6]. The parameters that affect the energy consumption of residential building are the environment and the structure of the house. Environmental parameters include temperature, humidity, intensity of sunlight exposure, etc. The structure of the building includes the relative compactness (RC), surface area (SA), wall area (WA), roof area (RA), overall height (OH), orientation (O), glazing area (GA), and glazing area distribution (GAD). These factors have influence on the energy-saving performance of buildings by affecting the heating load (HL) and cooling load (CL) of buildings. HL and CL are energy assessment methods that increase or remove part of the thermal energy from the room through the HVAC system to maintain the comfort of the indoor environment [7]. Therefore, accurate prediction of HL and CL is very important for designing energy-efficient buildings.

Recently, many scholars have proposed simulation tools to predict the HL and CL of buildings. Evcil [8] estimated the energy consumption of houses in Cyprus, Turkey, by calculating the average specific heat loss coefficient of houses in this region. Koo et al. [9] used the finite element theory to estimate the energy consumption of residential buildings. Li et al. [10] estimated the energy consumption of residential buildings in Chongqing City in China by taking the structure, weather conditions, and the age of the house into account. Simson et al. [11] established an overheating assessment method for a single-zone model, multizone apartment model, and multizone building model, which can dynamically and timely simulate the energy consumption of residential buildings. With the development of artificial intelligence technology, more and more researchers tend to use artificial intelligence methods to predict energy consumption, such as artificial neural network (ANN) [12–14], random forest (RF) [15], and Extreme Learning Machine (ELM) [16]. In residential building, the complex nonlinear relationships among the features affecting building energy consumption determine the prediction results of building energy consumption. Machine learning methods can solve nonlinear problems well, and typical machine learning methods mainly include decision tree (DT), K-nearest neighbor (KNN), support vector machine (SVM), BP neural network (BPNN), random forest, etc., whereas all of the single models have advantages and disadvantages and the advantages and disadvantages of each machine learning algorithm are listed in Table 1.

Different prediction methods reflect the changing trend of the objects and their influencing factors from different aspects; meanwhile, different information will be provided according to their respective principles. Therefore, any single prediction method is faced with the fluctuation of incomplete information and high prediction accuracy [20]. To overcome the problems that single machine learning models are prone to overfitting and sensitive to noisy data and have low prediction accuracy, Chou and Bui [7] proposed a combined model of SVR+ANN (support vector machine+artificial neural network). Their experimental results showed that the proposed SVR+ANN combined model has higher accuracy and is more efficient compared with the single models SVR and ANN. Kumar et al. [16] improved the Extreme Learning Machine (ELM) to obtain OSELM (Online Sequential ELM) and B-ELM (Bidirectional ELM) and combined these two methods to predict residential energy consumption. The above models have obtained satisfactory results for the accuracy of residential building energy consumption prediction. However, for energy saving and emission reduction in buildings, more accurate methods are needed to estimate building energy consumption, which can be used as a reference for building engineers to design energy-saving buildings.

The energy consumption of residential buildings is affected by the area of the house, the orientation of the house, the relative compactness of the house, and other factors. Meanwhile, the data distribution is complex with more discrete attribute variables and noisy data. As shown in Table 1, the BP neural network and random forest are more suitable for residential energy consumption prediction as they can handle nonlinear problems and are insensitive to noisy data compared with other machine learning algorithms. But it is pitiful that there is almost no research which combines the RF method and the BPNN method to predict building energy consumption in the current research. In order to improve the accuracy of machine learning models

in predicting building energy consumption, RF and BPNN are combined in this paper to obtain the RF-BPNN model, and the energy consumption of heating and cooling systems in different residential buildings in the energy efficiency data set in UCI is predicated. The selection of the appropriate weighted average coefficients of a single model in the combined model is also a key issue which will affect the model performance. Compared with the arithmetic average [21] and induced ordered weighted averaging (IOWA) [22], the variance-covariance (VC) [23] has better robustness. Therefore, in this paper, the VC is used to combine RF and BPNN to solve the problem of the dynamic weight allocation of a single model.

2. Models and Methods

2.1. Benchmark Prediction Model

2.1.1. Random Forest (RF). The random forest (RF) evolved from a classification and regression tree (CART), which is a collection of many trees. The CART method is a powerful nonlinear machine learning method with simple principle, which usually yields more accurate prediction results. Using a dichotomous recursive partitioning method, CART splits the sample set into two subsets so that there will be two branches at each nonleaf node above. The training process of RF is the same as in CART, with the difference that a randomly selected subset of candidate variables can be used to select the best variables for each segmentation. Being flexible, robust, usable, and efficient, RF can be used for analysis such as classification, regression, prediction, and clustering [24]. The model has been widely used in recent years due to its obvious advantages in parameter optimization, variable ranking, and subsequent variable analysis and interpretation [25–27]. Many experiments have shown that the RF algorithm can get better prediction results in many different applications [28]. However, the random forest belongs to the black box models; therefore, researchers cannot understand the internal operating mechanism of the random forest. Besides, the random forest is sensitive to noise.

2.1.2. BP Neural Network (BPNN). The artificial neural network (ANN) is a supervised machine learning algorithm proposed earlier. ANN is based on biological learning and has a structure similar to the human nervous system. A typical ANN structure envelops three layers, including an input layer, implicit layer, and output layer. In ANN, the BP neural network (BPNN) is one of the widely used neural networks, which is a multilayer feedforward neural network with backpropagation by error. During the training process, the connection weights and thresholds between neurons are continuously adjusted until a set target value is reached. Since the BP neural network can handle a large number of samples and can deal with nonlinear problems effectively and quickly, it is widely used in the fields of disease diagnosis [29, 30], traffic flow prediction [31], and service quality evaluation [32]. However, there is an obvious drawback in using BPNN alone to predict. Namely, the BP neural network is subjective in the selection of training samples with poor prediction accuracy and scalability. When solving problems with a larger scale and more features, it cannot get a higher accuracy rate. The main solution is to integrate models which can significantly improve the generalization ability of the BP neural network by integrating multiple machine learning models together [33].

2.2. RF-BPNN-Based Combined Prediction Model of Energy Consumption.

The prediction results of a single machine learning method are not accurate, and in order to make use of the advantages as well as overcome the shortcomings of a single model, this paper combines the random forest (RF) and BP neural network (BPNN) together to obtain the RF-BPNN model. Actually, the combined model is a heterogeneous integration model which combines and supplements the classification information provided by multiple single models through integrated thought and finally obtains an integration model. Therefore, both the prediction accuracy and generalization performance of the combined model can be further improved theoretically. The RF-BPNN model refers to a combined model obtained by weighted combination of the RF model's and the BPNN model's respective building energy consumption prediction results. The scope of application of the BF-BPNN model and the single model (RF model, BPNN model) is the same.

The flowchart of the RF-BPNN combined model for predicting building energy consumption is shown in Figure 1. The specific process is as follows.

Step 1. Preprocessing of raw data of residential building energy consumption. Data preprocessing includes data normalization and data set partition

(1) Data normalization. The input data is normalized in order to eliminate the dimension of the original data. This article applies the widely used standard deviation standardization method. And the normalization process is shown in

$$x* = \frac{x - \bar{x}}{\sigma}. \tag{1}$$

In Equation (1), $x*$ refers to the normalized data, x denotes the original data, $\bar{x}$ refers to the mean of the data, and σ means the standard deviation of the data.

(2) Data set partitioning. In order to test the generalization ability of the model, the data set is divided into a training set and test set with a ratio of about 9 : 1. The training set is used to estimate the model, and the test set is used to test the performance of the model

Step 2. The random forest (RF) model and BP neural network (BPNN) model are, respectively, used to predict the residential building energy consumption.

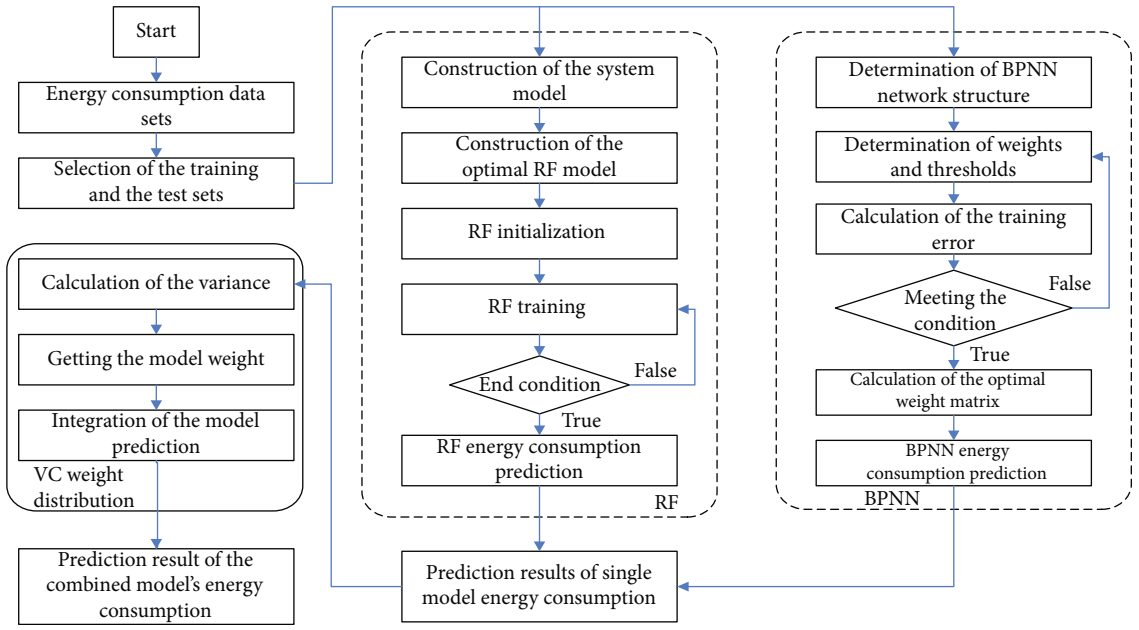

Figure 1: RF-BPNN energy consumption prediction flowchart.

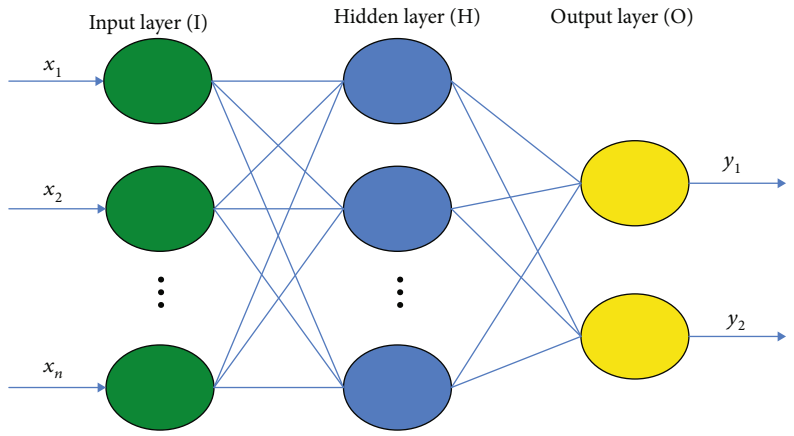

Figure 2: The structure of the BP neural network.

(1) Prediction of energy consumption in residential buildings through the random forest. The specific process is as follows

(i) Determine the number of decision trees in the random forest N. A random forest is a collection of many decision trees

(ii) Select n samples from the training set by the bootstrap method and select k attributes among all attributes

(iii) Select the optimal segmentation attribute. In the regression problems, the segmentation principle is the sum of error squares

(iv) Establish the CART decision tree

(v) Keep iterating processes (ii)–(iii) until all the trees in the random forest have split. Each decision tree will have a prediction result of a test set, and the average of the prediction results of all decision trees will be taken as the random forest prediction results

(2) Prediction of energy consumption in residential buildings through the BP neural network. The BP neural network is usually a three-layer network structure: input layer, hidden layer, and output layer. The structure of the BP neural network is shown in Figure 2

In this figure, $(x_1, x_2, \cdots, x_n)$ refer to the input sample of energy consumption data sets. (y_1, y_2) refer to the heat load

of two output variables for energy consumption prediction, respectively. The specific steps of the BP neural network for energy consumption prediction are as follows.

(i) Determine the structure of the BP neural network. The structure of the BP neural network mainly refers to the number of neurons in the hidden layer

(ii) Determine the connection of weights w and thresholds b. The neurons, which exist in the input and implicit layers and the implicit and output layers, are fully connected, and each connection has a corresponding weight w_i. Moreover, the threshold b is set to fit the data better

(iii) Calculate the training error. If the error between the training value and the real value is not within a reasonable range, it has to return back to step (ii) until the training error is within a reasonable range

(iv) Output the optimal weight matrix after training, apply it to the test set, and output the prediction results

Step 3. Combine the prediction results of the random forest and BPNN energy consumption obtained in Step 2 to obtain the final energy consumption prediction results. Given that residential buildings are characterized by a small area, large total volume, and various types, the single machine learning model cannot effectively predict the energy consumption of residential buildings. Combined prediction models can make use of the advantages of the single models to improve the prediction accuracy. The prediction results of different model combination methods vary greatly. The advantage of the variance-covariance (VC) combination method solves the dynamic weight assignment problem; namely, the optimal combination of weight coefficients can be found, and thus, it can improve the robustness and prediction accuracy of the model.

The variance of each prediction model is calculated by the following equation.

$$\delta_i = \frac{1}{n}\left[(e_1 - \bar{e})^2 + (e_2 - \bar{e})^2 + \cdots + (e_n - \bar{e})^2\right]. \tag{2}$$

In Equation (2), n denotes the number of training samples, $e_1, e_2, \cdots e_n$ represents the absolute percentage error of each training sample, and $\bar{e}$ refers to the average absolute percentage error of all training samples.

The dynamic weights for each model are calculated as follows.

$$w_1 = \frac{1}{[\delta_1 * (1/\delta_1 + 1/\delta_2)]}, \tag{3}$$

$$w_2 = \frac{1}{[\delta_2 * (1/\delta_1 + 1/\delta_2)]}. \tag{4}$$

The energy consumption prediction result of the combined model is obtained by multiplying the weights obtained from the above equation with the corresponding energy consumption prediction values and then summing the values.

$$p = w_1 * p_1 + w_2 * p_2. \tag{5}$$

In Equation (5), p denotes the energy consumption prediction results of the combined model and p_1, p_2 refers to the energy consumption prediction results of the two single models, respectively. To obtain better adaptability of the combined energy consumption prediction results, corresponding weights are dynamically adjusted through different training and testing results.

3. Empirical Results and Analysis

All experiments in this section are implemented in a unified experimental environment. In the experiments, the operating system is Windows 7, the CPU is Intel 1.60 GHz with 4 GB RAM, and the programming tool is PyCharm 2018.2.

3.1. Data Description and Statistical Analysis. To verify the effectiveness of the proposed RF-BPNN combined model for predicting energy consumption in residential buildings, the energy efficiency data set from the UCI, an authoritative database for machine learning, was used for the experiments. The energy efficiency data set consists of 768 data, 8 input variables ($_{X1\text{-}X8}$), and 2 output variables ($_{Y1}$, $_{Y2}$). The specific data descriptions are shown in Table 2. More information on the data description is given in Reference [15].

Figures 3 and 4 represent the scatter plots of HL and CL, respectively. It can be seen from the figures that HL and CL have similar trends and periodicity.

To further explore the strength of correlation between each input attribute ($_{X1\text{-}X8}$) and output variables ($_{Y1}$ and $_{Y2}$) in residential buildings, the Pearson correlation coefficient test was conducted, and the specific results are shown in Table 3. It can be seen from the figures that the input features $_{X1}$(RC), $_{X2}$(SA), $_{X4}$(RA), and $_{X5}$(OH) have strong linear correlations with the output variables $_{Y1}$(HL) and $_{Y2}$(CL). Also, some of them are highly correlated with two output features. For example, the correlation coefficient between $_{X1}$(RC) and $_{X5}$(OH) is 0.87, and the correlation coefficient between $_{X2}$(SA) and $_{X4}$(RA) is also 0.87, which indicates that there is a multicollinearity relationship between $_{X1}$(RC) and $_{X5}$(OH) as well as $_{X2}$(SA) and $_{X4}$(RA), while the correlation coefficient between $_{X4}$(RA) and $_{X5}$(OH) is -0.94, which is because the roof area calculation needs to be calculated by the roof height, so they show a negative correlation relationship. Additionally, some of them do not have an obvious linear correlation, such as $_{X6}$(OR) and $_{X8}$(GAD). Moreover, the relationship between the input features is also complicated. For example, $_{X1}$(RC) and $_{X2}$(SA) have a linear correlation coefficient of -1, because the volume of the building (V) is assumed to be constant here, and the relationship between them is shown in

$$\mathrm{RC} = V^{2/3} \cdot (\mathrm{SA})^{-1}. \tag{6}$$

Table 2: Description of the data set.

Variables	Property name	Abbreviations	Number of possible values	Minimum	Maximum	Average	Median	Standard deviation
X1	Relative compactness	RC	12	0.62	1.00	0.76	0.76	0.10
X2	Surface area	SA	12	2.00	808.50	670.83	661.50	91.28
X3	Wall area	WA	7	3.00	416.50	318.08	318.50	45.05
X4	Roof area	RA	4	4.00	220.50	176.37	147.00	45.56
X5	Overall height	OH	2	3.50	7.00	5.24	5.00	1.75
X6	Orientation	OR	4	2.00	6.00	3.50	4.00	1.12
X7	Glazing area	GA	4	0.00	7.00	0.24	0.25	0.27
X8	Glazing area distribution	GAD	6	0.00	8.00	2.81	3.00	1.56
Y1	Heating load	HL	583	1.00	43.10	22.27	18.89	10.11
Y2	Cooling load	CL	636	2.00	48.03	24.55	22.07	9.54

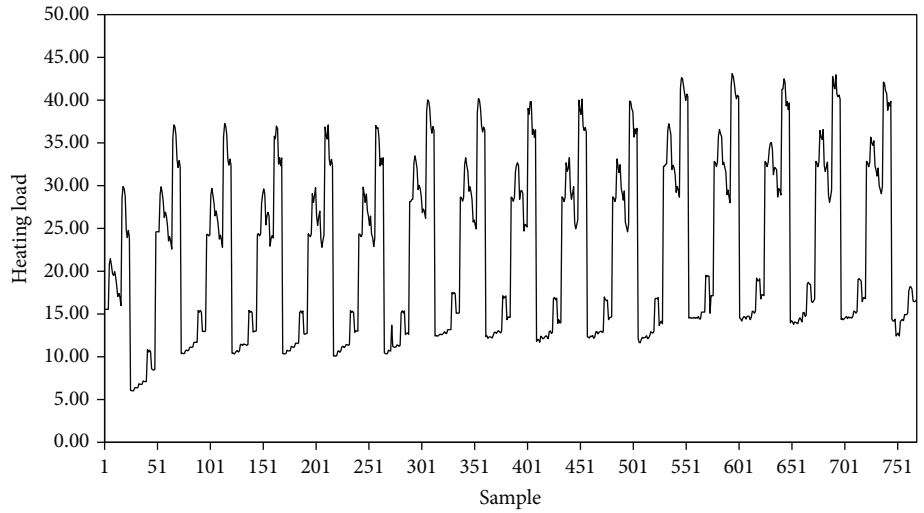

Figure 3: Heating load (HL).

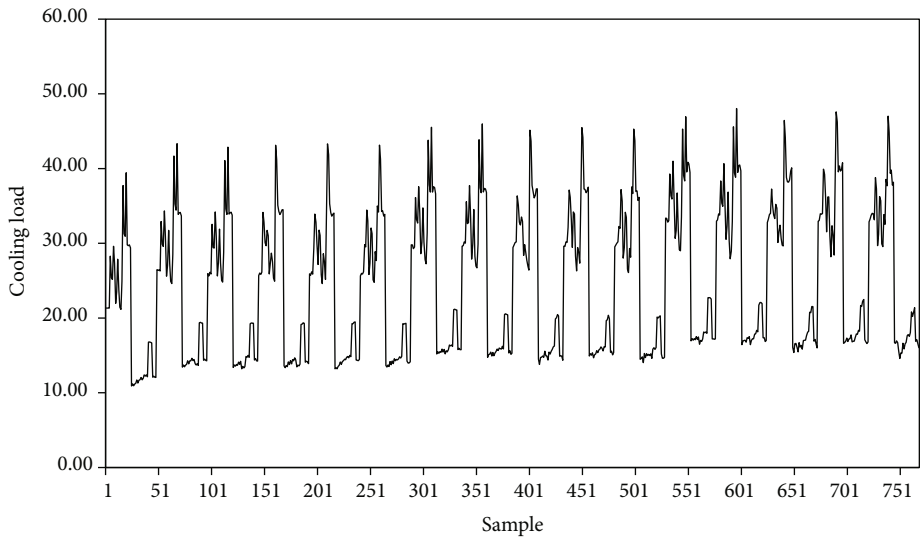

Figure 4: Cooling load (CL).

Table 3: Pearson correlation coefficient.

	X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2
X1	1.00	-1.00	-0.26	-0.87	0.87	0.00	0.00	0.00	0.62	0.65
X2	-1.00	1.00	0.26	0.87	-0.87	0.00	0.00	0.00	-0.62	-0.65
X3	-0.26	0.26	1.00	-0.19	0.22	0.00	0.00	0.00	0.47	0.42
X4	-0.87	0.87	-0.19	1.00	-0.94	0.00	0.00	0.00	-0.80	-0.80
X5	0.87	-0.87	0.22	-0.94	1.00	0.00	0.00	0.00	0.86	0.86
X6	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.02
X7	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.19	0.32	0.29
X8	0.00	0.00	0.00	0.00	0.00	0.00	0.19	1.00	0.07	0.05
Y1	0.62	-0.62	0.47	-0.80	0.86	0.00	0.32	0.07	1.00	0.97
Y2	0.65	-0.65	0.42	-0.80	0.86	0.02	0.29	0.05	0.97	1.00

It can be seen from Equation (6) that the relative compactness of the house and the surface area of the house are inversely related. In summary, the relationship between the input attributes of residential buildings is very complex and cannot be predicted accurately with simple linear models; thus, more complex nonlinear models, such as neural networks, random forests, and support vector machines, are needed.

3.2. Parameter Setting. The number of hidden layer nodes of the BPNN has a crucial impact on the experimental results. As there are different problems to be solved, there is no accurate method to guide the selection of the appropriate number of hidden layer nodes for BPNN. Therefore, according to the study, a typical 3-layer BP neural network is established, in which eight input attributes are used as input units and two prediction targets are used as the number of neurons in the output layer, and the method to determine the number of nodes in the hidden layer is referred to in the literature [34] with the following equation.

$$m = \sqrt{n + l} + a. \tag{7}$$

In Equation (7), n is the input layer neural unit, l refers to the output layer neural unit, and a refers to the arbitrary constant between 0 and 20. After several experiments, it is finally determined that the best result is obtained by taking 20. Therefore, the topology of BPNN is 8-23-2. And the remaining parameters are determined by extensive experiments: the maximum training number is 500, the minimum error rate of the training target is 0.0001, and the training speed is 0.1. After several experiments, when the number of random trees in the random forest is set to 10, it predicts the best results.

To prevent overfitting, the 10-fold cross-validation method is used and the average result of 100 repeated runs is taken as the final result.

3.3. Model Assessment. In order to accurately assess the predictive performance of the model, five regression model evaluation criteria, mean absolute error (MAE), root mean square deviation (RMSD), mean absolute percentage error (MAPE), correlation coefficient (R), and index of agreement (IA), were used. And the calculation equations are as follows.

(1) Mean absolute error (MAE):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|. \tag{8}$$

(2) Root mean square deviation (RMSD):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y\wedge_i)^2}. \tag{9}$$

(3) Mean absolute percentage error (MAPE):

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{y}_i|}{y_i} \times 100\%. \tag{10}$$

(4) Correlation coefficient (R):

$$R = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y\wedge_i - \bar{y})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{11}$$

(5) Coincidence indicator (IA):

$$\mathrm{IA} = 1 - \frac{\sum_{i=1}^{n}(y\wedge_i - y_i)^2}{\sum_{i=1}^{n}(|y\wedge_i - \bar{y}| + |y_i - \bar{y}|)^2}. \tag{12}$$

In abovementioned equations, n refers to the number of samples in the test set, y_i refers to the true value, $\bar{y}$ denotes the average of the true value, and $\widehat{y}_i$ means the predicted value. MAE, RMSD, and MAPE all indicate the error between the predicted and true values, so the smaller the value, the better the model. R denotes the degree of correlation between the predicted value and the true value, and the larger the value, the stronger the correlation between them. In addition, when $R = 1$, it means that the predicted value is

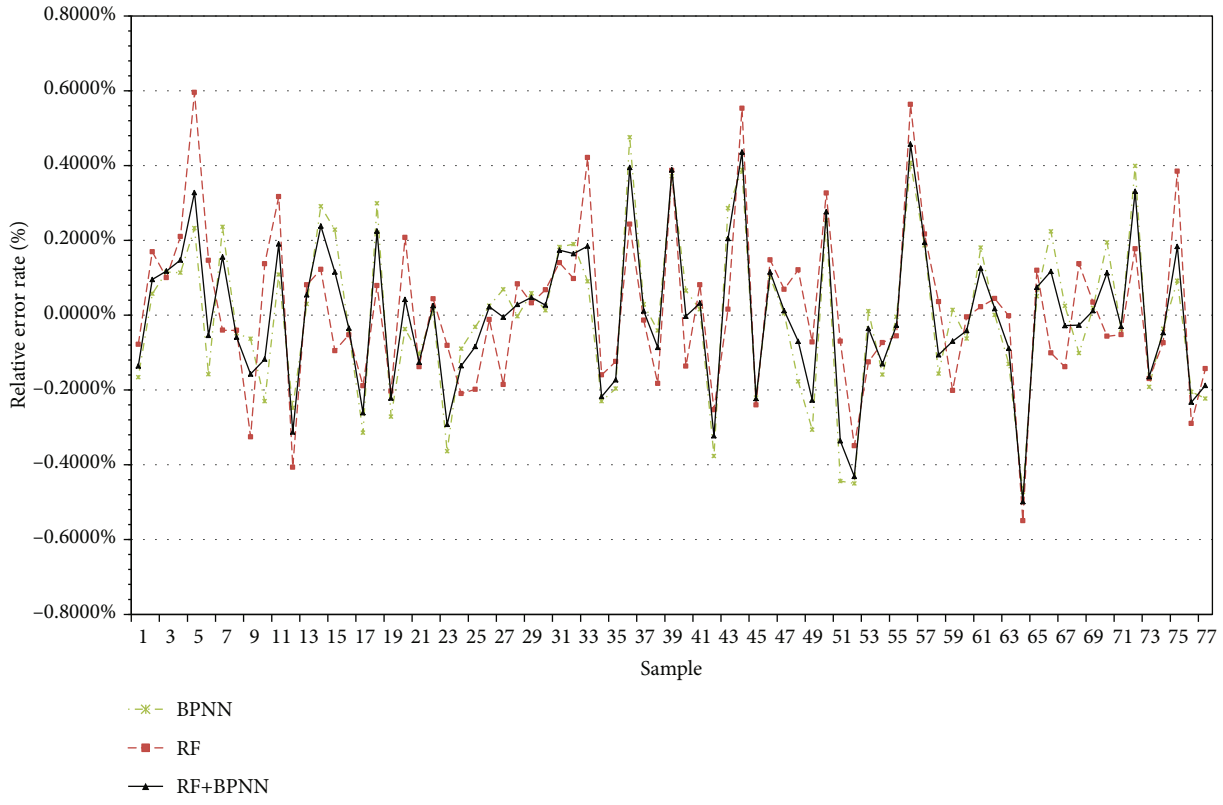

Figure 5: Relative error rate curves of HL predicted by each model.

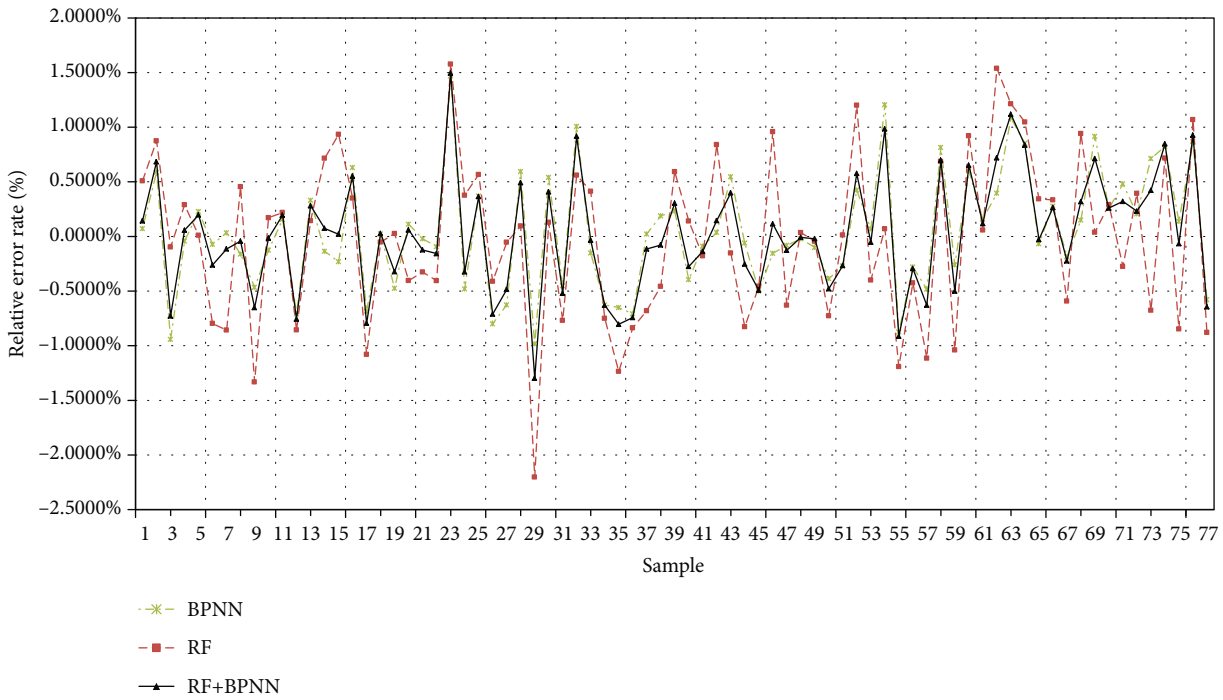

Figure 6: Relative error rate curves of CL predicted by each model.

completely correlated with the true value. Namely, there is a linear relationship when the viability is 1. According to Equation (12), IA is to eliminate the effects of variable dimension for comparison between different models, which usually has a value between 0 and 1. For almost perfect models, IA will generally be close to 1 [35, 36].

Table 4: Comparison of prediction results of RF, BPNN, and RF-BPNN models.

	Evaluation criteria	RF	BPNN	RF-BPNN
HL	MAE	0.3243 ± 0.0384	0.3451 ± 0.0475	**0.3199 ± 0.0900**
	RMSE	0.4870 ± 0.0679	0.4799 ± 0.0685	**0.4550 ± 0.0601**
	MAPE (%)	**1.3971 ± 0.0019**	1.6306 ± 0.0031	1.4591 ± 0.0023
	R	0.9988 ± 0.0004	0.9989 ± 0.0004	**0.9990 ± 0.0003**
	IA	0.9994 ± 0.0002	0.9994 ± 0.0004	**0.9995 ± 0.0001**
CL	MAE	0.9783 ± 0.1472	0.8101 ± 0.1309	**0.7765 ± 0.1212**
	RMSE	1.6338 ± 0.2200	1.1911 ± 0.2092	**1.1614 ± 0.1841**
	MAPE (%)	3.3800 ± 0.0042	3.2300 ± 0.0050	**3.0000 ± 0.0043**
	R	0.9853 ± 0.0039	0.9919 ± 0.0003	**0.9924 ± 0.0026**
	IA	0.9922 ± 0.0020	0.9958 ± 0.0016	**0.9960 ± 0.0013**

Note: the results in the table represent the mean of the results of 100 repeated runs and standard deviation.

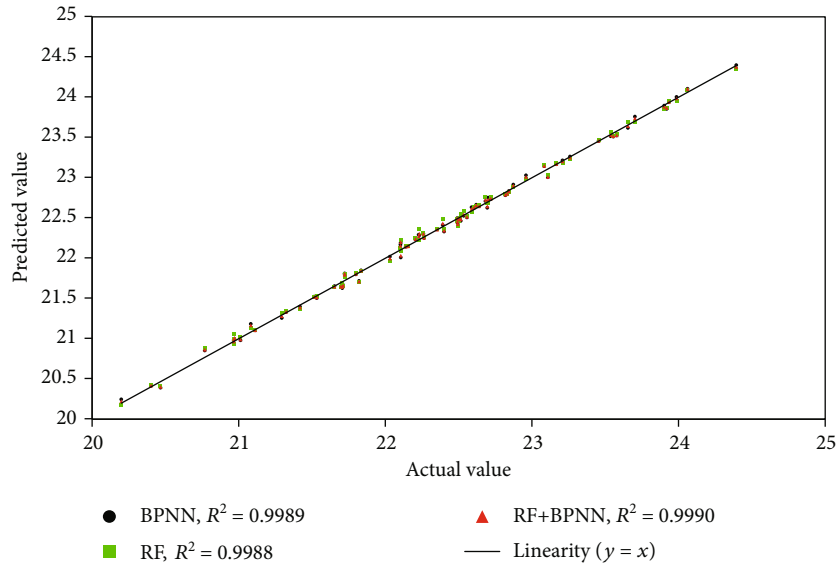

Figure 7: Comparison of HL's true value and predicted value.

3.4. Analysis of Results. To fully illustrate that the RF-BPNN combined model has higher accuracy than a single model in predicting building energy consumption, the relative accuracy of the prediction results of the two test sets (HL and CL) is compared, and the results are shown in Figures 5 and 6.

As can be seen in Figures 5 and 6, the volatility of the single models RF and BPNN is greater than that of the combined model RF-BPNN, especially in predicting CL. The VC combination method is used to solve the dynamic weight assignment problem, and these models are combined to obtain the RF-BPNN model. The combined weights of the single models of RF and BPNN are 0.236 and 0.764, respectively. And the combination results get better adaptability of the training results by dynamically adjusting the corresponding weights according to different training results. Therefore, the volatility of RF-BPNN is the least, which indicates that the prediction results of RF-BPNN are more accurate compared with RF and BPNN.

To further compare the prediction accuracy of the proposed combined model RF-BPNN with that of the single models RF and BPNN, the regression model performance evaluation methods introduced in Section 3.3 were applied, respectively, and the results are shown in Table 4 with the best model performance results indicated in bold.

As can be seen from Table 4, in terms of the prediction result error—MAE and RMSE, the prediction result of BPNN is the worse and that of RF-BPNN is better. In terms of the correlation coefficient R, RF-BPNN obtained the largest result, indicating that the predicted value of RF-BPNN has the strongest correlation with the true value. Finally, in terms of the evaluation criterion IA, RF-BPNN also obtained the largest value, which is closer to 1, indicating that this model is almost perfect and the prediction accuracy is relatively

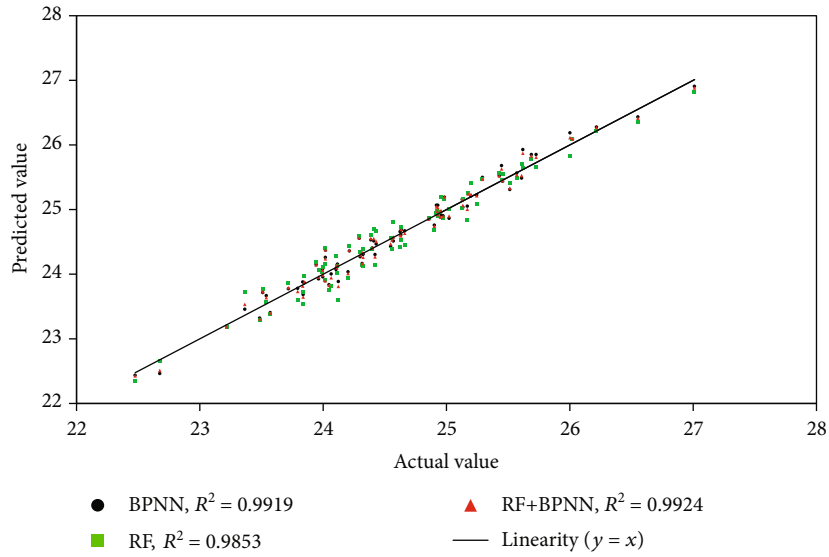

Figure 8: Comparison of the true and predicted values of CL.

high. In terms of the standard deviation of the results of 100 runs, except for MAE and MAPE, RF-BPNN has the smallest standard deviation, which indicates that this model is more stable. Overall, the combined model RF-BPNN has more accurate and stable prediction results than the single models RF and BPNN, and this advantage of RF-BPNN is more obvious especially in the prediction of CL.

Figures 7 and 8 represent the comparison of the true values of HL and CL with the predicted values of each model, respectively. The black dots in the figures indicate the predicted values of the BPNN model, the green squares refer to the predicted values of RF, the red triangles refer to the predicted values of the combined RF-BPNN model, and the black lines denote $y = x$. Namely, the closer the predicted values are to the $y = x$ line, the more accurate the prediction results are.

As seen in Figure 7, almost all the points are close to the $y = x$ straight line, but the RF-BPNN model gets the largest R of 0.9990. The advantage of the RF-BPNN model is more obvious in Figure 8, where the red triangles are almost all gathered in the $y = x$ straight line, which indicates that the RF-BPNN model has the most accurate prediction results.

In addition, to further verify the effectiveness of the RF-BPNN method proposed in this paper, the experimental results obtained in this paper are compared with the results of existing models. Because the results of the existing models are not publicly available, the results from the original literature are cited. And the results of the comparison are shown in Table 5, and the best model performance is indicated in bold.

As can be seen in Table 5, as far as HL prediction is concerned, the prediction results of the SVR+ANN model proposed in the literature [7] have the smallest MAE and RMSE values of 0.3000 and 0.4280, respectively, but the evaluation criteria MAE and RMSM obtained by the RF-BPNN model are 0.3199 and 0.4550, respectively, whose numerical gap is very small from that of the SVR+ANN. The other methods obtained MAE and RMSE which differ greatly from the results of SVR+ANN, especially the IRLS (iteratively reweighted least squares) method in the literature [15] that obtained very poor results, which further indicates that the common regression methods are not applicable to the prediction of building energy HL and CL. In terms of the indicator MAPE, the HYBRID-LIN method proposed in the literature [6] predicts more accurate results. For MAE, RMSE, and MAPE of CL prediction results, RF-BPNN obtains the smallest values. Overall, RF-BPNN has a higher prediction accuracy compared to the existing models.

Table 5: Comparison of RF-BPNN with existing model results.

	Methods	MAE	RMSE	MAPE (%)
HL	IRLS [15]	2.1400	3.1400	10.0900
	SVR [16]	0.4320	0.6100	—
	CART [16]	0.4370	0.8000	—
	ANN-SVR [7]	**0.3000**	**0.4280**	1.5570
	HYBRID-LIN [6]	0.5100	0.7874	**0.4700**
	KNN	1.9529	2.3329	8.4504
	RF-BPNN	0.3199	0.4550	1.4591
CL	IRLS [15]	2.2100	3.3900	8.4100
	SVR [16]	0.8900	1.6470	—
	CART [16]	1.1570	1.8410	—
	ANN-SVR [7]	0.973	1.5660	3.4550
	HYBRID-LIN [6]	1.1800	2.0372	3.3300
	KNN	1.8193	2.2651	7.1413
	RF-BPNN	**0.7765**	**1.1614**	**3.0000**

4. Conclusion

Recently, in order to effectively curb global warming, China has put forward new requirements for energy saving and emission reduction in succession. Building energy consumption,

the main cause of climate change, is facing great challenges. Therefore, it is very important to predict building energy consumption accurately. To this end, this paper predicts the heating and cooling energy consumption of different residential buildings in the energy efficiency data set of the UCI database by proposing the RF-BPNN combined model. The 10-fold cross-validation method is applied to prevent overfitting, and the average result of 100 model runs is used as the final prediction result to eliminate the effect of random data selection on the generalization ability of the model. Five model evaluation metrics which include mean absolute error, root mean square deviation, mean absolute percentage error, correlation coefficient, and coincidence index are applied to verify the performance of the RF-BPNN combined model. The experimental results show that the RF-BPNN combined model proposed in this paper can accurately predict the HL and CL of residential building energy consumption compared with the prediction results of the single RF and BPNN models. In addition, the advantages of the RF-BPNN combined model are further illustrated by comparing the prediction results with those of existing models (Table 5). It is demonstrated that the RF-BPNN combined model is easy to apply and has great value in building energy consumption.

Data Availability

Statistics used in this paper are from the UCI. The data can be downloaded from http://archive.ics.uci.edu/ml/datasets/Energy+efficiency.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work is partially supported by the General Fund of National Natural Science Foundation of China (71871086).

References

[1] J. X. Gao, H. Ren, W. G. Cai, X. R. Ma, and M. H. Tang, "Study on 'dilution effect' of per unit area energy consumption in residential building in China," *System Engineering-Theory & Practice*, vol. 39, no. 3, pp. 569–577, 2019.

[2] UNEP SBCI, *Buildings and Climate Change: Summary for Decision-Makers*, United Nations Environmental Programme, Paris, 2009.

[3] Z. Yu, F. Haghighat, B. C. Fung, and H. Yoshino, "A decision tree method for building energy demand modeling," *Energy & Buildings*, vol. 42, no. 10, pp. 1637–1646, 2010.

[4] W. G. Cai, Y. Wu, Y. Zhong, and H. Ren, "China building energy consumption: situation, challenges and corresponding measures," *Energy Policy*, vol. 37, no. 6, pp. 2054–2059, 2009.

[5] Y. Wang and L. P. Zhang, "A simulation analysis of optimized incentive policies in green housing market -a case study of Xi'an," *Systems Engineering*, vol. 36, no. 5, pp. 37–46, 2018.

[6] M. Castelli, L. Trujillo, L. Vanneschi, and A. Popovič, "Prediction of energy performance of residential buildings: a genetic programming approach," *Energy and Buildings*, vol. 102, pp. 67–74, 2015.

[7] J. S. Chou and D. K. Bui, "Modeling heating and cooling loads by artificial intelligence for energy-efficient building design," *Energy and Buildings*, vol. 82, pp. 437–446, 2014.

[8] A. Evcil, "An estimation of the residential space heating energy requirement in Cyprus using the regional average specific heat loss coefficient," *Energy and Buildings*, vol. 55, pp. 164–173, 2012.

[9] C. Koo, S. Park, T. Hong, and H. S. Park, "An estimation model for the heating and cooling demand of a residential building with a different envelope design using the finite element method," *Applied Energy*, vol. 115, pp. 205–215, 2014.

[10] X. Li, R. Yao, W. Yu et al., "Low carbon heating and cooling of residential buildings in cities in the hot summer and cold winter zone - a bottom-up engineering stock modeling approach," *Journal of Cleaner Production*, vol. 220, pp. 271–288, 2019.

[11] R. Simson, J. Kurnitski, and K. Kuusk, "Experimental validation of simulation and measurement-based overheating assessment approaches for residential buildings," *Architectural Science Review*, vol. 60, no. 3, pp. 192–204, 2017.

[12] A. Yezioro, B. Dong, and F. Leite, "An applied artificial intelligence approach towards assessing building performance simulation tools," *Energy & Buildings*, vol. 40, no. 4, pp. 612–620, 2008.

[13] C. Turhan, T. Kazanasmaz, I. E. Uygun, K. E. Ekmen, and G. G. Akkurt, "Comparative study of a building energy performance software (KEP-IYTE-ESS) and ANN-based building heat load estimation," *Energy and Buildings*, vol. 85, pp. 115–125, 2014.

[14] Y. Wei, L. Xia, S. Pan et al., "Prediction of occupancy level and energy consumption in office building using blind system identification and neural networks," *Applied Energy*, vol. 240, pp. 276–294, 2019.

[15] T. Athanasios and X. Angeliki, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy and Buildings*, vol. 49, pp. 560–567, 2012.

[16] S. Kumar, S. K. Pal, and R. P. Singh, "Intra ELM variants ensemble based model to predict energy performance in residential buildings," *Sustainable Energy, Grids and Networks*, vol. 16, pp. 177–187, 2018.

[17] J. F. Yang, P. R. Qiao, Y. M. Li, and N. Wang, "A review of machine-learning classification and algorithms," *Statistics & Decision*, vol. 6, pp. 36–40, 2019.

[18] J. Li and L. L. Xu, "Comparison and analysis of research trend prediction models based on machine learning algorithm-BP neural network, support vector machine and LSTM model," *Journal of Modern Information*, vol. 39, no. 4, pp. 24–34, 2019.

[19] J. M. Ding, G. Q. Liu, and H. Li, "The application of improved random forest in the telecom customer churn prediction," *Pattern Recognition and Artificial Intelligence*, vol. 28, no. 11, pp. 1041–1049, 2015.

[20] D. Niu, Y. Liang, H. Wang, M. Wang, and W. C. Hong, "Icing forecasting of transmission lines with a modified back propagation neural network-support vector machine-extreme learning machine with kernel (BPNN-SVM-KELM) based on the variance-covariance weight determination method," *Energies*, vol. 10, no. 8, pp. 1196–1217, 2017.

[21] H. L. Lu, "Applications of the combination weighted arithmetic averaging operator in the delivery merchandise sale

forecast of material resources," *Journal of Natural Science of Heilongjiang University*, vol. 29, no. 1, pp. 22–28, 2016.

[22] W. Han, J. Wang, and X. H. Zhang, "Application research of combined forecasting based on induced ordered weighted averaging operator," *Management Science and Engineering*, vol. 8, pp. 23–26, 2014.

[23] H. Zhang, Y. J. Zhu, L. L. Fan, and Q. L. Wu, "Mid-long term load interval forecasting based on Markov modification," *East China Electric Power*, vol. 41, no. 1, pp. 33–36, 2013.

[24] X. Fang, Z. Wen, J. Chen, S. Wu, Y. Huang, and M. Ma, "Remote sensing estimation of suspended sediment concentration based on random forest regression model," *Journal of Remote Sensing*, vol. 23, no. 4, pp. 756–772, 2019.

[25] W. J. Wu, P. Jing, H. F. Jia, and M. H. Zhang, "Low carbon travel intention data mining for residents based on K-means clustering and random forest algorithm," *Journal of South China University of Technology (Natural Science Edition)*, vol. 47, no. 7, pp. 105–111, 2019.

[26] C. Y. Lai, Y. Q. Guo, and H. F. Yu, "Fuel metering unit performance degradation detection and remaining useful life estimation methods based on FR-SVR," *Journal of Aerospace Power*, vol. 34, no. 7, pp. 1624–1632, 2019.

[27] Z. Y. Wang, Y. Ni, and J. Zhang, "Research on patent value assessment method of network platform based on grey relational analysis and random forest regression," *Information Studies: Theory & Application*, vol. 42, no. 10, pp. 109–116, 2019.

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

[29] A. Tsanas, M. A. Little, P. E. Mcsharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *Journal of the Royal Society Interface*, vol. 8, no. 59, pp. 842–855, 2011.

[30] T. Zhang, X. L. Hao, Y. J. Zhang, and Y. H. Zhang, "A medical diagnosis model based on BP-AsymBoost algorithm," *System Engineering-Theory & Practice*, vol. 37, no. 6, pp. 1654–1664, 2017.

[31] S. Li, L. J. Liu, and M. Zhai, "Prediction for short-term traffic flow based on modified PSO optimized BP neural network," *System Engineering-Theory & Practice*, vol. 32, no. 9, pp. 2045–2049, 2012.

[32] Y. Liu, M. Yu, and Y. Liu, "Evaluation model of bus route service quality based on passengers perceptions," *Journal of Northeastern University (Natural Science)*, vol. 40, no. 5, pp. 145–150, 2019.

[33] P. K. Srimani and M. S. Koti, "Medical diagnosis using ensemble classifiers-a novel machine-learning approach," *Journal of Advanced Computing*, vol. 1, pp. 9–27, 2013.

[34] J. Jin, L. Zhu, Z. H. Li, X. H. Tong, and C. W. Yang, "Application of variable structure of BPNN in risk evaluation of overseas railway construction in target countries," *Journal of the China Railway Society*, vol. 40, no. 12, pp. 11–16, 2018.

[35] Q. Zhou, H. Jiang, J. Wang, and J. Zhou, "A hybrid model for $PM_{2.5}$ forecasting based on ensemble empirical mode decomposition and a general regression neural network," *Science of the Total Environment*, vol. 496, pp. 264–274, 2014.

[36] S. Zhu, X. Lian, L. Wei et al., "$PM_{2.5}$ forecasting using SVR with PSOGSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors," *Atmospheric Environment*, vol. 183, no. 5, pp. 20–32, 2018.