energies

MDPI

Article

Modular Predictor for Day-Ahead Load Forecasting and Feature Selection for Different Hours

Lin Lin [1,*], Lin Xue [2], Zhiqiang Hu [3] and Nantian Huang [2]

[1] College of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin 132022, China
[2] School of Electrical Engineering, Northeast Electric Power University, Jilin 132013, China; xuelin1428@163.com (L.X.); huangnantian@126.com (N.H.)
[3] Zhejiang Electric Power Corporation Wenzhou Power Supply Company, Wenzhou 325000, China; huzhiqiang19910714@163.com
* Correspondence: jllinlin@126.com; Tel.: +86-139-4420-8142

Received: 25 June 2018; Accepted: 12 July 2018; Published: 20 July 2018

Abstract: To improve the accuracy of the day-ahead load forecasting predictions of a single model, a novel modular parallel forecasting model with feature selection was proposed. First, load features were extracted from a historic load with a horizon from the previous 24 h to the previous 168 h considering the calendar feature. Second, a feature selection combined with a predictor process was carried out to select the optimal feature for building a reliable predictor with respect to each hour. The final modular model consisted of 24 predictors with a respective optimal feature subset for day-ahead load forecasting. New England and Singapore load data were used to evaluate the effectiveness of the proposed method. The results indicated that the accuracy of the proposed modular model was higher than that of the traditional method. Furthermore, conducting a feature selection step when building a predictor improved the accuracy of load forecasting.

Keywords: day-ahead load forecasting; modular predictor; feature selection

1. Introduction

The main idea of short-term load forecasting (STLF) is to predict future loads with horizons of a few hours to several days. Accurate STLF predictions play a vital role in electrical department load dispatch, unit commitment, and electricity market trading [1]. With the permeation of renewable resources in grids and the technological innovation of electric vehicles, load components become more complex and make STLF difficult; therefore, strict requirements of stability and accuracy are needed [2–6].

STLF is an old but worthy theme for research. General forecasting methods can be divided into two branches: the statistical method and the artificial intelligence method. Statistical methods such as regression analysis, exponential smoothing, Kalman filter, and autoregressive integrated moving average (ARIMA) are easy to apply but modeling is difficult for complex loads [7–9]. Artificial intelligence methods show better performance than statistical methods in load forecasting and include fuzzy logic, the artificial neural network (ANN), the support vector machine (SVM), Gaussian process regression (GPR), and random forest (RF) [10–17]. The relationship of input and output is confirmed by a list of rules by fuzzy logic. However, the prior knowledge required to select the parameters in the membership function and the rules makes the modeling process complex [18]. The artificial neural network method is applied to the STLF of power systems owing to its self-learning ability and robustness to data noise. However, shortcomings such as the difficulty in determining initial network parameters and over-fitting still exist [19]. By adopting a structural risk minimization principle, the complexity and the learning ability of an SVM can be balanced. With low-dimension conditions

and few samples, the SVM can maintain its generalization ability. Compared to the artificial neural network, the SVM has many advantages. The parameters of the SVM should be determined through a computational optimization by algorithm such as the genetic algorithm or the particle swarm optimization algorithm [20,21]. GPR is a kernel-function-based algorithm whose transcendental function is established in the form of probability distribution, and the posterior function can be acquired by Bayesian logic. The parameter of kernel function in GPR is obtained automatically in the process of training [22]. RF is a type of integrated machine-learning algorithm based on a decision tree. The main advantages of RF are immunity to noise and insensitivity to its parameters [23].

In addition to the forecasting method, input feature selection is a vital factor that influences the accuracy and efficiency of load forecasting. A model using a few features has difficulty analyzing the effect of external conditions on the load. However, as the complexity of a model increases, the accuracy and efficiency will be influenced. Feature selection is a process of selecting a subset of variables from an original high-dimensionality variable set that retains the most efficient variables while reducing the effects of the irrelevant variables [24]. Feature selection methods can be classified as wrapper, filter, and embedded [25]. In the wrapper method, the performance of a predictor is chosen as the criterion for feature selection. An exhaustive search is performed to identify the optimal feature subset from numerous combinations of features at which the predictor performs best. However, the wrapper method needs to evaluate 2N subsets which leads to an NP-hard problem with too many features [26]. Therefore, evolutionary algorithms such as the memetic algorithm [27], the genetic algorithm [28], and the particle swarm optimization algorithm [29] can reduce the complexity of computation. Filter methods, such as mutual information (MI) and RreliefF, are ranking methods that evaluate features by analyzing the relationship between the inputs and outputs and a feature score or weight is given to each feature for ranking. To acquire an optimal feature subset, the accuracy of the predictor is used as the criterion [30]. Compared to wrapper methods, filter methods do not rely on other learning algorithms and the computational cost is light [31–33]. Embedded methods, such as the classification and regression tree (CART) and RF, which combine feature selection with a learning algorithm, analyze and compute the importance value of features in a training process [25]. Experiments need to be performed according to a specific forecasting case that considers the advantages and disadvantages of different kinds of feature selection methods, the size of training sets, and the performance of a predictor to determine the most-accurate forecasting method.

Although the performance of a predictor can be provided by feature selection, it should be noted that the load time series presents a day-cycle characteristic, which means the load characteristics at the same time on different days are similar [34]. In addition, the load at different hours of a day is affected by consumption behavior and leads to significantly different feature responses. A single predictor with a feature selection for forecasting all future load periods may not reach the load requirement of different hours, and the accuracy of the total forecast result will decrease. Therefore, a modular model that consists of several single predictors used for forecasting the load of different hours is needed. The relation of the load at different hours to be forecast and a feature could be analyzed by a modular predictor with a feature selection for a specific hour of load, and thus the accuracy can be improved [35]. In addition, in electric power dispatching, for different electric power departments, the demand of the time of submission of the STLF result is different. Therefore, when constructing a candidate feature set for STLF, the time factor should be considered.

Considering the construction of a feature set, feature selection, and modeling objects, a novel modular parallel forecasting model with feature selection for day-ahead load forecasting was proposed. First, to meet the requirement of the dispatch department and electricity market, the load time series which records every hour according to different forecasting moments was reconstructed to a different load sub-time series. Second, the candidate feature set included 173 features extracted from historic load and calendar. Then, five feature selection methods—MI, conditional mutual information (CMI), RreliefF, CART, and RF—were used to analyze the importance between each feature and different prediction targets and to rank the features in descending order. Third, combined with various

predictors, the sequential forward-selection algorithm and a decision criterion based on the mean absolute percentage error (MAPE) were utilized to obtain optimal feature subsets corresponding to different prediction targets. Finally, the optimal modular predictor including several optimal sub-predictors with optimal feature subsets for different forecasting periods was built. The optimal combination method was determined by comparing the forecast results. The proposed method was tested through a day-ahead load forecasting experiment using actual load data from New England and Singapore.

2. Feature Selection

The input feature (variable), as one of the key factors in a predictor build, has a significant influence on the accuracy of the predictor in day-ahead load forecasting. In this study, the filter method and embedded method were adopted for feature selection before building the predictor.

2.1. Filter Method of Feature Selection

The filter method is a feature ranking method that computes a feature's numerical value to evaluate its importance. Therefore, the estimation of a feature is important to the feature selection result. MI, CMI, and RreliefF methods were used as filters in this study.

2.1.1. Mutual Information

The Mutual Information (MI) method measures the common information between two random variables. For two random variables X and Y, the MI between X and Y can be estimated as:

$$I(X,Y) = \sum_{X,Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

where $P(x)$ and $P(y)$ are the marginal density functions corresponding to X and Y, respectively. $P(x,y)$ is the joint probability density function. In load forecasting, the feature is defined as X, the target variable is defined as Y, and $I(X,Y)$ represents their strength of relevance. The larger $I(X,Y)$ is, the more dependent X is. If $I(X,Y)$ is zero, X and Y are independent. The MI method can measure the relevance between a feature and a target variable effectively; however, the redundancy is analyzed differently.

2.1.2. Conditional Mutual Information

The Conditional Mutual Information (CMI) method measures the relevance of two variables when the variable Z is known. In the feature selection of load forecasting, let us suppose the selected feature set is S and the CMI between feature X_i and target Y is defined as:

$$I(Y;X_i|S) = I(Y;S|X_i) - I(Y;S) \tag{2}$$

where $I(Y;X_i \mid S)$ represents the new information that X_i supplies to S. The larger $I(Y;X_i \mid S)$ is, the more information X_i can supply, and the less is the redundancy to S. Compared to the MI method, the redundancy among features can be reduced by CMI.

2.1.3. RreliefF

RreliefF is the extended version of relief for regression [36]. By evaluating the feature weight, the feature quality is measured. Relief works by randomly selecting an instance and searching the nearest neighbor from the same class and from a different class. The weight $W[X_i]$ of feature X_i estimated by relief is an approximation of the difference of probabilities:

$$\begin{aligned} W[X_i] &= P(diff,\ value\ of\ X_i | nearest\ inst.\ from\ diff.\ class) \\ &- P(diff,\ value\ of\ X_i | nearest\ inst.\ from\ same.\ class) \end{aligned} \tag{3}$$

For RreliefF, the probability of two instances belonging to different classes can be evaluated by their relative distances for classification. However, for STLF, the predicted value is continuous; therefore, Equation (3) should be reformulated. By using Bayes' theorem, $W[X_i]$ can be obtained as:

$$W[X_i] = \frac{P_{diffC|diffX_i} P_{diffX_i}}{P_{diffC}} - \frac{(1 - P_{diffC|diffX_i}) P_{diff_{X_i}}}{1 - P_{diffC}} \tag{4}$$

2.2. Embedded Method for Feature Selection

In the embedded method, feature selection is performed during the training process where the contribution of the feature combination is efficiently evaluated. The embedded method can be directly applied to STLF and can collaborate with other feature selection methods according to their estimated importance.

2.2.1. Classification and Regression Tree

The Classification and Regression Tree (CART) method uses a binary recursive partitioning algorithm [37]. By splitting the current samples into two sub-samples, a father node generates two child nodes. The final model of CART is a simple binary tree.

The generation of the CART can be divided into two steps:

Step one: first, the root node is split. A best feature X^{bset} chosen from the feature set serves as the criterion for node splitting. To select the best feature, the minimum variance of child nodes is the objective function. The variance of the child node of X_i is defined as:

$$\text{var}(q) = \sum_{X_i \in q} (y_i - \overline{y}_q)^2 \tag{5}$$

where $\overline{y}_q$ is the average of observation values y_i at node q. The importance of feature X_i according to the variance is defined as:

$$V_C(X_i) = \frac{1}{\sum_{X_i \in q} (y_i - \overline{y}_q)^2} \tag{6}$$

Step two: for each child node, repeat Step one until the CART grows completely. The predictive model can be expressed as $t(\boldsymbol{x}, T)$, where $T = (\boldsymbol{x}_i, y_i)$, $i = 1,2, \ldots ,n$ and $\boldsymbol{x} \in \boldsymbol{R}$ is the training set. For STLF, the forecasting value of load $\hat{y}$ is obtained when inputting the new $\hat{\boldsymbol{x}}$.

$$\hat{y} = t(\hat{\boldsymbol{x}}, T) \tag{7}$$

2.2.2. Random Forest

Random Forest (RF) is a machine-learning algorithm that uses a combination of CART with a bootstrap sample for classification and regression [38]. For a training set T with n samples, the bootstrap sample means randomly selecting n samples from T replacements. The probability that each sample selected is $1/n$, means one sample may appear several times. After a complete bootstrap sample, the samples that were not sampled form the out-of-bag (OOB) dataset. Different from CART, the feature for node splitting in RF is selected from m features which are chosen from the original feature set. The basis of selecting the best feature for node splitting is Equation (5). The predictive output of RF is obtained by averaging the results of the trees:

$$\hat{y} = \frac{1}{N_t} \sum_{i=1}^{N_t} t(\hat{\boldsymbol{x}}, T^i) \tag{8}$$

where N_t is the number of trees.

In addition, the OOB error and the importance of each feature are computed in the process of modeling. Each tree has an OOB dataset, and the OOB error is evaluated by predicting the OOB dataset using the tree model corresponding to the OOB dataset. The OOB error is defined as:

$$e = \frac{1}{N_t}\sum_{i=1}^{N_t}(y_i - \hat{y}_i)^2 \tag{9}$$

A feature's importance is estimated by permutating the feature and averaging the difference of OOB errors before and after the permutation of all trees. For instance, for the ith tree whose OOB data is OOB_i and OOB error is e_i, after permutation, the new OOB data will be OOB_i' and the OOB error will be e_i'. The feature's importance in this tree is computed as:

$$VI_i = e_i' - e_i \tag{10}$$

3. The Short-Term Load Forecasting (STLF) Predictor

Selecting an appropriate predictor is key to improving the accuracy of STLF. Five state-of-the-art predictors were applied in this study: support vector regression (SVR), back-propagation neural network (BPNN), CART, GPR, and RF. The SVR, BPNN, and GPR are introduced briefly in this section. The detailed mathematical theories of these algorithms are shown in the references [39–41].

3.1. Support Vector Regression

By using the non-sensitive loss function, an Support Vector Regression (SVM), which is used only for classification, is extended for regression to be applied for load forecasting in power systems and is called support vector regression (SVR).

Given a training set T, the model for the load that decreases the difference between the predictive value $f(\boldsymbol{x})$ and the true load y as much as possible is expected to be:

$$f(\boldsymbol{x}) = \omega^T \boldsymbol{x} + b \tag{11}$$

In SVR, the maximum difference that can be tolerated between $f(\boldsymbol{x})$ and y is ε. The mathematical model can be expressed as:

$$\begin{cases} \max\limits_{\alpha,\alpha^*}\left[-\frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(\boldsymbol{x}_i,\boldsymbol{x}_j) - \sum\limits_{j=1}^{n}(\alpha_i+\alpha_i^*)\varepsilon + \sum\limits_{i=1}^{n}(\alpha_i-\alpha_i^*)y_i\right] \\ s.t.\begin{cases} \sum\limits_{i=1}^{n}(\alpha_i-\alpha_i^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le C \end{cases} \end{cases} \tag{12}$$

where C is the regularization parameter, $K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \boldsymbol{\varphi}(\boldsymbol{x}_i)\boldsymbol{\varphi}(\boldsymbol{x}_j)$ is the kernel function, and α_i, α_i^* are Lagrange factors.

The radial basis function selected in this study is expressed as:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right) \tag{13}$$

where σ^2 is the kernel width.

The SVR model is obtained by solving Equation (12):

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(\boldsymbol{x}_i,\boldsymbol{x}) + b \tag{14}$$

where b is the bias value.

3.2. Back-Propagation Neural Network

The Back-Propagation Neural Network (BPNN) is a type of artificial neural network consisting of an input layer, a hidden layer, and an outer layer trained by a back-propagation algorithm with the mean squared error (MSE) as the objective function. The main idea of the BPNN is to deliver the output-layer error from back to front by which the error of the hidden layer is computed. The learning process of BPNN is divided into two steps:

Step 1: The output of each neural unit in the input and hidden layers is estimated.

Step 2: By using the output error, the error of each neural unit which is used for updating the former layer weight is computed.

The objective function of the gradient minimization is based on:

$$e_f = \frac{1}{2}\sum_i (y_i - \hat{y}_i)^2 \tag{15}$$

where y_i is the actual value of neural unit i and $\hat{y}_i$ is the predictive value. To compute the minimum value of e_f, a modification value is needed to correct the weight. The modification value is defined as:

$$\Delta w_{ij}(t) = -\eta \frac{\partial e}{\partial w_{ij}} + \alpha \Delta w_{ij}(t-1) = -\eta \frac{\partial e}{\partial net_i}\frac{\partial net_i}{\partial w_{ij}} + \alpha \Delta w_{ij}(t-1) = -\eta \delta_i O_j + \alpha \Delta w_{ij}(t-1) \tag{16}$$

$$net_i = \sum_j w_{ij} O_j \tag{17}$$

$$O_i = \frac{1}{1 + e^{-net_i}} \tag{18}$$

where η is the learning rate, net_i is the input of neuron i, O_i is the output of neuron i, and α is the momentum factor.

The modified weight is:

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij} \tag{19}$$

The final output $\hat{y}_i$ of neuron i can be estimated by the iteration of weight w_{ij} when meeting precision requirements.

3.3. Gaussian Process Regression (GPR)

Gaussian Process Regression (GPR) is a random process in which the random variables obey the Gaussian distribution and is used to establish the input and output maps. For STLF, the load data collected is polluted by noise. Assuming that the noise follows a normal distribution $\varepsilon \sim N(0, \sigma_n^2)$, then the joint prior distribution of observation y and the predictive value f^* are defined as:

$$\begin{bmatrix} \boldsymbol{y} \\ f^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 \boldsymbol{I}_n & K(X, \boldsymbol{x}^*) \\ K(\boldsymbol{x}^*, X) & k(\boldsymbol{x}^*, \boldsymbol{x}^*) \end{bmatrix}\right) \tag{20}$$

where n is the number of training samples, $K(X,X)$ is the covariance matrix, and $\boldsymbol{I}_n$ is the unit matrix.

The posterior distribution of f^* is defined as:

$$f^*|X, \boldsymbol{y}, \boldsymbol{x}^* \sim N\left[\overline{f}^*, \mathrm{cov}(f^*)\right] \tag{21}$$

where $\overline{f}^*$ is the mean value of f^* and $\mathrm{cov}(f^*)$ is the variance of f^*.

Then, $\overline{f}^*$ and $\mathrm{cov}(f^*)$ can be computed as:

$$\begin{cases} \overline{f}^* = K(x,X)\left[K(X,X) + \sigma_n^2 \boldsymbol{I}_n\right]^{-1} \boldsymbol{y} \\ \mathrm{cov}(f^*) = k(\boldsymbol{x}^*, \boldsymbol{x}^*) - K(\boldsymbol{x}^*, X) \times \left[K(X,X) + \sigma_n^2 \boldsymbol{I}_n\right]^{-1} K(X, \boldsymbol{x}^*) \end{cases} \tag{22}$$

The covariance function of GPR is the squared exponential function:

$$k(\boldsymbol{x}, \boldsymbol{x\prime}) = \sigma_f^2 exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x\prime})\boldsymbol{M}^{-1}(\boldsymbol{x} - \boldsymbol{x\prime})\right] \tag{23}$$

where $\theta = \left\{M, \sigma_f^2, \sigma_n^2\right\}$ is a hyper-parameter that can be solved by the maximum likelihood method [41].

4. Data Analysis

4.1. Load Analysis

Affected by different factors, load sequence appears as a type of complicated non-linear time series. To construct a reasonable original feature set and achieve better forecasting for a region, the load characteristics and other factors should be analyzed.

Figure 1 shows the power load of New England in different time lengths. By observing Figure 1a,b, the load patterns from 2011 to 2013 are similar. Influenced by climate, load patterns differ by season. In Figure 1c, the load curves of two continuous weeks in four seasons are presented (the first day is Monday). It is easy to see that the weekday and weekend load demands differ, and the load demand presented a cycling mode with a period of seven days. The Tuesday load curves of the different seasons shown in Figure 1d shows that the Tuesday load pattern of different weeks is similar. The load increased rapidly from 6:00 am to 11:00 am, which corresponds to the beginning of work, and reached the first peak load. The second peak load occurred from 19:00 pm to 20:00 pm.

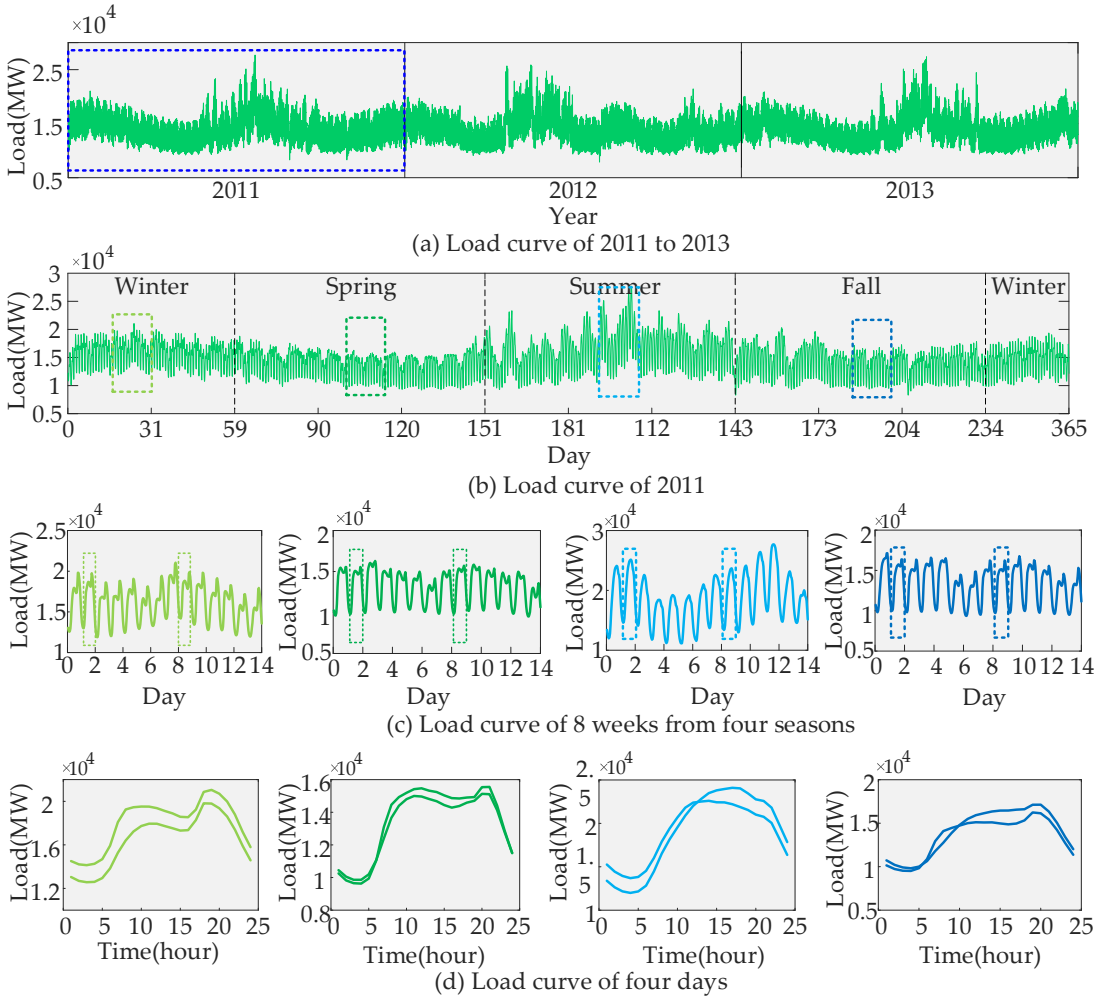


Figure 1. The power load of New England.

As analyzed above, the load characteristics can be summarized as

(1) The same-day load patterns are similar and represent the week-cycle of the load.

(2) The weekday and weekend load patterns were similar respectively and represent the day-cycle of the load.

4.2. Candidate Feature Set

An appropriate feature set plays a significant role in modeling an uncomplicated but outstanding predictor. However, a candidate feature set that contains sufficient information must be found to ensure that effective features can be selected by the feature selection method. The two main feature types are the endogenous predictor (load feature) and the exogenous predictor (calendar feature).

The time interval before the predictive moment before submission of the dispatch department's forecasting result should be considered when extracting features. To ensure the universality of the original feature set, we used the interval time $p = 24$. A feature set consisting of 145 internal historic load features (from lag 24 to lag 168) from a one-week data window was chosen as a part of the candidate feature set. The maximum, minimum, and mean loads were also included. Except for the load feature, calendar features such as hour of day, day type, working day, and non-working day were also considered. The candidate feature set with 173 features was formed as shown in Table 1.

Table 1. The feature information.

Feature Type	Feature Name	Feature Number
Endogenous predictor	$F_{L(t-i)}$, i = 24, 25, . . . , 168	145
	$F_{L(\max,d-k)}$, $F_{L(\min,d-k)}$, $F_{L(\mathrm{mean},d-k)}$, k = 2, 3, 4, 5, 6, 7	18
Exogenous predictor	F_D^W, W = 1, 2, 3, 4, 5, 6, 7	7
	F_W	2
	F_{Hour}	1

Feature explanation:

Endogenous predictor:

$F_{L(\max,d-k)}$ is the maximum power load k days before, k = 2, 3, 4, 5, 6, 7.

$F_{L(\min,d-k)}$ is the minimum power load k days before, k = 2, 3, 4, 5, 6, 7.

$F_{L(\mathrm{mean},d-k)}$ is the average power load k days before, k = 2, 3, 4, 5, 6, 7.

$F_{L(t-i)}$ is the historic power load i hours before the forecasting hour t, and i = 24, 25, 26, . . . , 168.

Exogenous predictor:

F_D^W is the day of week, which is signed by 0 or 1 (W = 1, 2, 3, 4, 5, 6, 7 represents Monday to Sunday).

F_W is work day or non-work day (0 is a work day and 1 is a non-work day).

F_{Hour} is the moment of hour (1 to 24).

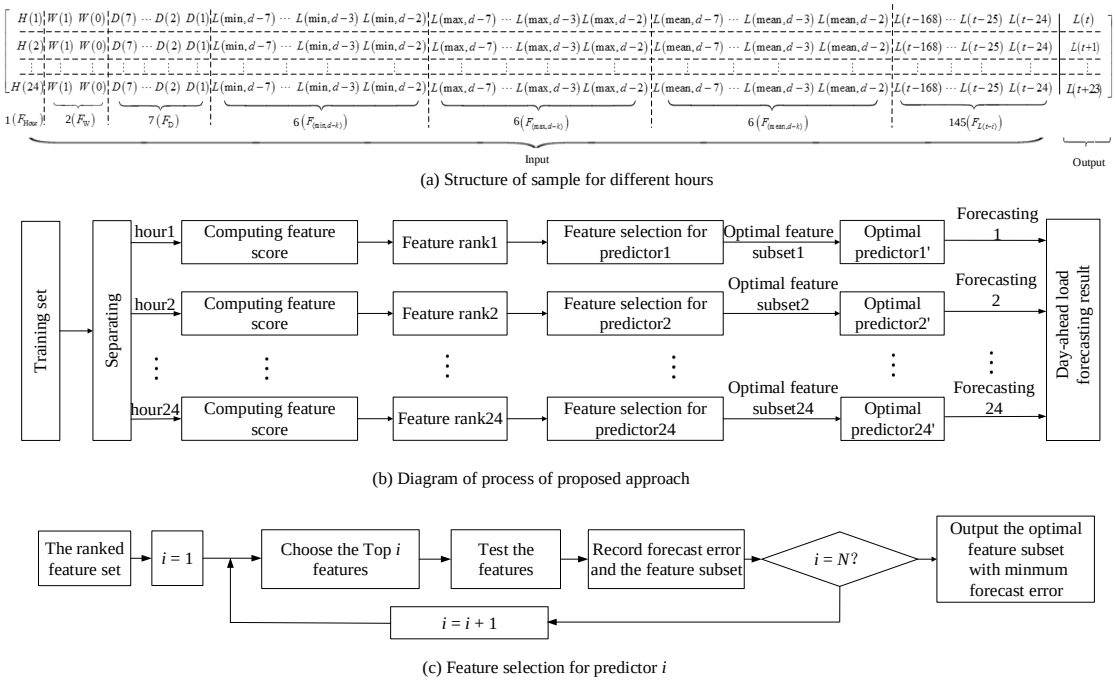
5. Experimental Setup

5.1. Proposed STLF Process with Feature Selection

Figure 2 provides an overview of the proposed method which covers the construction of the feature set, the dataset separation, and the feature selection for the load with respect to the different hours and the modeling for different hours. Figure 2a shows the one-day structure of a sample. The inputs include 173 features, and the output is the predicted load.

The diagram of the proposed method is displayed in Figure 2b. The training set was separated into 24 training subsets corresponding to each hour. The features in each training subset were ranked in descending order according to their feature scores as computed by the feature selection method. Then, the optimal feature subset was selected using the predictor and the MAPE-based criteria. Finally, the modular predictor was constructed based on 24 predictors with the obtained optimal subsets.

The process of selecting the optimal feature subset in modeling is shown in Figure 2c. According to the ranked feature order, the predictor was used to test the feature subset consisting of the top i features, and the criteria based on the MAPE was used to select the optimal feature subset.



(a) Structure of sample for different hours

(b) Diagram of process of proposed approach

(c) Feature selection for predictor *i*

Figure 2. Overview of the proposed method.

5.2. Dataset Split

The data used in this study were from New England [42] and Singapore [43]. The New England data were recorded every hour from 2011 to 2013 for a total of 26,304 data points. The Singapore data were recorded every half hour from 2014 to 2015 for a total number of 35,040 data points. To apply the proposed method, the hourly load of Singapore was extracted to form a new load time series. The data used for training and testing the predictor consisted of the feature set (173 features) and the predictive object (the load corresponding to different hours) as shown in Figure 2.

Each dataset was split into three parts: a training set (14,616 New England samples, 11,712 Singapore samples), a validation set (2928 New England samples, 2094 Singapore samples), and a test set (8760 New England samples, 2904 Singapore samples). The training and the validation sets were used to build the predictor and to select an optimal feature subset. The test set was used to examine the performance of the feature subset and the predictor. Detailed information about the datasets is shown in Table 2.

Table 2. Experimental data description.

Data Set	Detail Information of Experimental Data (New England)			Detail Information of Experimental Data (Singapore)	
	2011	2012	2013	2014	2015
Training set	Jan., Feb., Mar., Apr., May, Jun., Jul., Aug., Sept., Oct., Nov., Dec.	Jan., Feb., Apr., Jun., Jul., Aug., Oct., Dec.	-	Jan., Feb., Mar., Apr., May, Jun., Jul., Aug., Sept., Oct., Nov., Dec.	Jan., Apr., Aug., Dec.
Validation set	-	Mar., May, Sept., Nov.	-	-	Feb., May, Jul., Oct.
Test set	-	-	Jan., Feb., Mar., Apr., May, Jun., Jul., Aug., Sept., Oct., Nov., Dec.	-	Mar., Jun., Sept., Nov.

5.3. Evaluation Criterion

To evaluate the performance of the proposed method, three criteria, the MAPE, the mean absolute error (MAE), and the root mean square error (RMSE) were used as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{24}$$

$$MAE = \frac{1}{n}|y_i - \hat{y}_i| \tag{25}$$

$$RMSE = \sqrt{\frac{1}{n}(y_i - \hat{y}_i)^2} \tag{26}$$

where y_i is the actual load, $\hat{y}_i$ is the predictive load, and n is the number of predictive loads.

6. Results

The software used were MATLAB 2016b (Version 9.1.0.441655, Mathworks Inc., Natick, MA, USA) and Rx64 3.3.2 (Version 3.3.2, GUN Project, developed at Bell Laboratories). It is noted that the CART algorithm in the rpart package in R identifies part of the features whose total importance value is 100. The parameter of each predictor was set by:

BPNN: the number of neurons in the hidden layer was $N_{neu} = 2 \times N_{feature} + 1$, iteration $T = 1000$ [44].

SVR: the regularization parameter $C = 1$, the non-sensitive loss function $\varepsilon = 0.1$, the kernel width $\delta^2 = 2$ [15].

RF: $m = N_{feature}/3$ and $N_{Tree} = 500$ [16,23].

CART: no pruning parameter was set because the tree grows completely.

GPR: the parameter of GPR was tuned by learning the training data.

6.1. Load Forecasting for New England

6.1.1. Feature Selection for Different-Hour Loads

Feature Score for Feature Analysis

Feature selection methods rate the importance of a feature by assigning a numerical value to represent the relation between the feature and the target. For example, the value of a feature computed by MI is called an MI value, while that computed by RF and CART is called its permutation importance. The feature score is used for easy description. Parts of normalized feature score curves computed by different feature selection methods are shown in Figure 3. The feature score curves of typical hours (hour 5, hour 6, hour 10, and hour 11 when the valley and peak loads appear) were chosen for analysis. The feature score curves that used the same feature score calculation method were different at various hours. For example, the MI curves were much different for hour 5, hour 6, hour 10, and hour 11, and the features with the highest scores were different from each other (marked by a red circle).

The feature score shows the importance between the feature and the target variable. When selecting a feature subset, the feature with the highest score should be retained and one with the lowest should be eliminated.

The top 10 features after ranking are shown in Table 3, where it is clear that the top 10 features for the same hour were similar. For example, for hour 5, the same top 10 features were selected by the various methods such as $F_{L(t-24)}$, $F_{L(t-25)}$, $F_{L(t-26)}$, and $F_{L(t-27)}$ and similar features such as $F_{L(t-28)}$, $F_{L(t-29)}$, $F_{L(t-30)}$, and $F_{L(t-31)}$. However, there was an obvious difference in the features of hour 5 and hour 6 which may have been caused by the different load characteristics shown in Figure 1d.

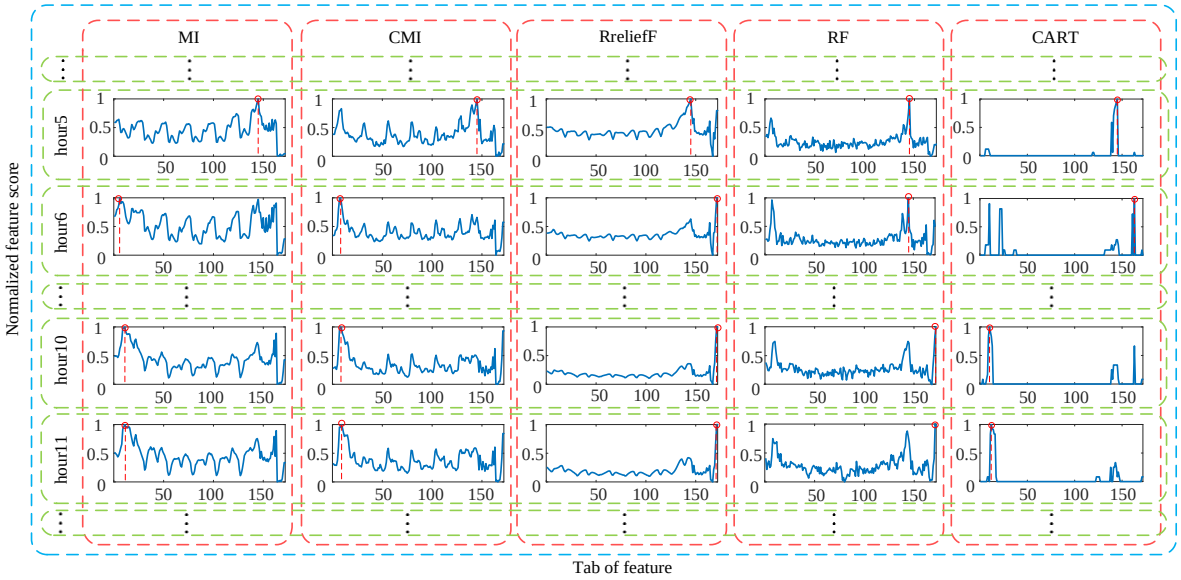


Figure 3. Normalized feature score of features evaluated by kinds of feature selection methods.

Table 3. Top 10 features of ranked of feature by different feature selection corresponding to Figure 3.

	MI	CMI	RreliefF	RF	CART
Hour 5	$F_{L(t-24)}$, $F_{L(t-25)}$, $F_{L(t-26)}$, $F_{L(t-27)}$, $F_{L(t-28)}$, $F_{L(t-29)}$, $F_{L(\min,\,d-2)}$, $F_{L(t-30)}$, $F_{L(\text{mean},\,d-2)}$, $F_{L(t-44)}$	$F_{L(t-24)}$, $F_{L(t-25)}$, $F_{L(t-29)}$, $F_{L(t-28)}$, $F_{L(t-160)}$, $F_{L(t-26)}$, $F_{L(t-161)}$, $F_{L(t-162)}$, $F_{L(t-27)}$, $F_{L(\max,\,d-2)}$	$F_{L(t-24)}$, $F_{L(t-25)}$, $F_{L(t-26)}$, $F_{L(t-27)}$, $F_{L(t-28)}$, F_W^0, F_W^1, $F_{L(t-28)}$, $F_{L(\max,\,d-2)}$, $F_{L(t-31)}$	$F_{L(t-24)}$, $F_{L(t-25)}$, $F_{L(t-163)}$, $F_{L(t-162)}$, $F_{L(t-26)}$, $F_{L(t-164)}$, $F_{L(t-30)}$, $F_{L(t-29)}$, $F_{L(t-160)}$, $F_{L(t-27)}$	$F_{L(t-24)}$, $F_{L(t-25)}$, $F_{L(t-26)}$, $F_{L(t-27)}$, $F_{L(t-28)}$, $F_{L(t-30)}$, $F_{L(t-163)}$, $F_{L(t-160)}$, $F_{L(t-161)}$, $F_{L(t-162)}$
Hour 6	$F_{L(t-160)}$, $F_{L(t-162)}$, $F_{L(t-161)}$, $F_{L(t-24)}$, $F_{L(t-164)}$, $F_{L(\text{mean},\,d-7)}$, $F_{L(t-163)}$, $F_{L(t-159)}$, $F_{L(t-28)}$, $F_{L(t-29)}$	$F_{L(t-161)}$, $F_{L(t-162)}$, $F_{L(t-160)}$, $F_{L(t-163)}$, $F_{L(t-159)}$, $F_{L(t-29)}$, $F_{L(t-145)}$, $F_{L(t-158)}$, $F_{L(t-141)}$, $F_{L(t-65)}$	F_W^0, F_W^1, F_D^7, $F_{L(t-24)}$, $F_{L(t-25)}$, $F_{L(t-26)}$, F_D^1, $F_{L(t-28)}$, $F_{L(t-27)}$, $F_{L(t-29)}$	$F_{L(t-24)}$, $F_{L(t-162)}$, $F_{L(t-161)}$, $F_{L(t-160)}$, $F_{L(t-30)}$, $F_{L(t-29)}$, $F_{L(t-25)}$, F_W^0, $F_{L(t-163)}$, $F_{L(\text{mean},\,d-7)}$	$F_{L(\text{mean},\,d-7)}$, $F_{L(t-159)}$, $F_{L(t-147)}$, $F_{L(t-146)}$, $F_{L(t-148)}$, $F_{L(\max,\,d-7)}$, $F_{L(t-24)}$, $F_{L(t-25)}$, $F_{L(t-30)}$, $F_{L(t-26)}$
Hour 10	$F_{L(t-158)}$, $F_{L(t-159)}$, $F_{L(t-157)}$, $F_{L(\text{mean},\,d-7)}$, $F_{L(t-160)}$, $F_{L(t-156)}$, $F_{L(t-24)}$, $F_{L(t-154)}$, $F_{L(t-147)}$, $F_{L(t-153)}$	$F_{L(t-161)}$, $F_{L(t-160)}$, $F_{L(t-162)}$, F_W^0, F_W^1, $F_{L(t-159)}$, $F_{L(t-158)}$, $F_{L(t-157)}$, $F_{L(t-154)}$, $F_{L(t-155)}$, $F_{L(t-159)}$	F_W^0, F_W^1, F_D^7, F_D^6, $F_{L(t-26)}$, $F_{L(t-25)}$, $F_{L(t-27)}$, $F_{L(t-24)}$, $F_{L(t-28)}$, F_D^1	F_{WW}^1, F_W^0, $F_{L(t-159)}$, $F_{L(t-25)}$, $F_{L(t-160)}$, $F_{L(t-24)}$, $F_{L(t-161)}$, $F_{L(t-26)}$, $F_{L(t-28)}$, $F_{L(t-27)}$	$F_{L(t-159)}$, $F_{L(t-158)}$, $F_{L(t-160)}$, $F_{L(t-157)}$, $F_{L(\text{mean},\,d-7)}$, $F_{L(t-156)}$, $F_{L(t-25)}$, $F_{L(t-27)}$, $F_{L(t-28)}$, $F_{L(t-26)}$
Hour 11	$F_{L(t-159)}$, $F_{L(t-157)}$, $F_{L(t-158)}$, $F_{L(t-156)}$, $F_{L(\text{mean},\,d-7)}$, $F_{L(t-153)}$, $F_{L(t-155)}$, $F_{L(t-152)}$, $F_{L(t-160)}$, $F_{L(t-154)}$	$F_{L(t-160)}$, $F_{L(t-162)}$, $F_{L(t-161)}$, $F_{L(t-159)}$, $F_{L(t-158)}$, F_W^0, F_W^1, $F_{L(t-154)}$, $F_{L(t-156)}$, $F_{L(t-155)}$	F_W^0, F_W^1, F_D^7, $F_{L(t-26)}$, $F_{L(t-27)}$, $F_{L(t-25)}$, $F_{L(t-33)}$, $F_{L(t-24)}$, $F_{L(t-34)}$, $F_{L(t-28)}$	F_W^1, F_W^0, $F_{L(t-26)}$, $F_{L(t-27)}$, $F_{L(t-25)}$, $F_{L(t-161)}$, $F_{L(t-157)}$, $F_{L(t-160)}$, $F_{L(t-24)}$, $F_{L(t-158)}$	$F_{L(t-157)}$, $F_{L(t-156)}$, $F_{L(t-155)}$, $F_{L(t-153)}$, $F_{L(t-154)}$, $F_{L(t-158)}$, $F_{L(t-26)}$, $F_{L(t-25)}$, $F_{L(t-27)}$, $F_{L(t-28)}$

Therefore, a feature analysis for each hour is required to choose the best features for improving the accuracy of STLF.

Optimal Feature Subset Selection Process

According to the trend of feature score curves of diverse feature selection methods, the first 36 to 50 features are chosen as the optimal features for modeling [30]. By analyzing the autocorrelation of the lag variables, 50 features were selected for very-short-term load forecasting [41]. When selecting a feature subset, most studies did not give a specific threshold for selecting the optimal feature subset. In this study, the performance of features which ranked in descending order based on feature score were estimated by the MAPE which was chosen as the threshold for selecting the optimal feature subset by adding features one-by-one to the feature subset.

Figure 4 shows the MAPE curves of different feature selection methods and predictor-based feature selection processes. As shown in each subplot in Figure 4, the MAPE was reduced and reached a minimum value with an increase in the number of features. For example, the MAPE of MI for hour 5 and the MAPEs of BPNN, CART, GPR, RF, and SVR when using the top feature were 4.587%, 4.743%, 4.618%, 5.196%, and 4.718%, respectively. When 20 features were used, the MAPEs were reduced to 3.901%, 4.555%, 4.008%, 4.160%, and 3.831%, respectively. The MAPEs of different predictors decreased in different levels, indicating that the 20 features made a positive contribution to a better prediction model build. A similar conclusion can be summarized by analyzing other curves. The dimension of each optimal feature subset and its MAPE is marked by different colored circles corresponding to different predictors.

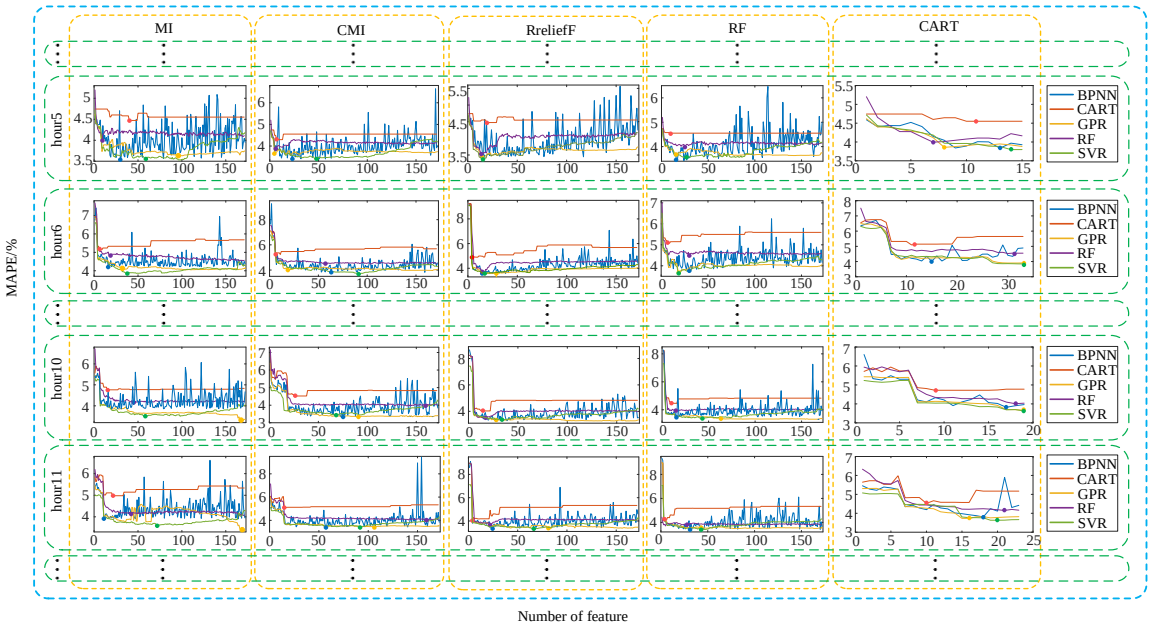


Figure 4. Mean absolute percentage error (MAPE) curves of combinations of feature selection methods and forecasting methods for selecting feature subset.

The following conclusions can be drawn from Table 3 and Figure 4:

(1) The feature permutation estimated by different feature selection methods varies.

(2) The dimension of the optimal feature subset and its MAPE depends on the predictor-based feature selection method.

(3) The optimal feature subset selected by the same predictor-based feature selection method for the predictive target of different hours is different.

Table 4 shows the MAPE and the dimension of the optimal feature subset corresponding to using MI as the feature selection method and RF as the prediction model. The Table shows that 1 to 6 am, the dimension of optimal feature subset is less than at 7 to 19 pm, as the same as the forecasting error. This is because people are less active at night and there are fewer factors affecting the load than during the day.

The MAPE and the dimension of optimal feature subset corresponding to 1:00 were carried out by different feature selection methods and forecasting methods shown in Table 5. The MAPEs are in 3% to 4% which means the performance of forecasters were similar after feature selection. By analysis of the feature dimension, we could find there is huge difference between the number of the feature of the optimal feature subset that selected by different feature selection methods, which caused by the different evaluation criterions.

Table 4. Optimal feature subset construction of different hours with mutual information (MI) + random forest (RF) for New England.

Time	MAPE	FD	Time	MAPE	FD
1	3.294	34	13	4.663	41
2	3.419	22	14	4.926	33
3	3.632	9	15	5.190	38
4	3.783	30	15	5.351	46
5	4.008	9	17	5.547	31
6	4.828	18	18	5.358	98
7	5.456	61	19	5.117	136
8	5.314	59	20	4.506	23
9	4.526	64	21	4.376	28
10	4.171	45	22	4.779	9
11	4.147	42	23	4.794	41
12	4.414	67	24	4.847	72

Remark: FD means the feature dimension.

Table 5. Optimal feature subset construction of 1:00 with different methods for New England.

Method	CART		RF		SVR		ANN		GPR	
	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD
MI	3.741	7	3.294	34	3.064	10	3.226	8	3.087	119
CMI	3.729	2	3.447	20	3.043	13	3.062	47	3.052	134
CART	3.729	3	3.422	11	3.068	11	3.270	9	3.245	8
RF	3.741	7	3.533	51	3.140	41	3.069	18	3.099	81
RreliefF	3.741	10	3.310	26	3.043	18	3.269	9	3.019	134

The details of the dimension of the optimal feature subset and its MAPE are shown in Appendix A Table A1 to Table A2. Based on a longitudinal comparison, the dimension of optimal feature subsets selected by different feature selection methods with same-hour predictors were different. For instance, the horizon of the hour-2 MAPE calculated by various methods was from 3.107% to 4.050%. The combination RreliefF + SVR method had the smallest MAPE and lower feature subset dimension.

By the horizontal comparison, the dimension of optimal feature subsets selected by the same feature selection method with the same-hour predictor varied. For example, the horizon of dimension of the feature subset corresponding to different hours selected by the RreliefF + SVR method ranged from 13 to 109 and the MAPE range was 3.043% to 4.558%. In addition, the number of features for a night hour was less than the day hour, indicating that the day load components were more complex and more difficult to forecast.

In conclusion, the characteristic of the load to predict for different hours varies; therefore, the load needs a special feature set to build a predictor for special hours. The necessity of using one kind of structure of modular time-scale prediction and feature selection for the load of different hours was verified.

6.1.2. Forecasting Result of Method Combinations with Optimal Feature Subsets for New England Load Data

To test the performance of diverse method combinations with the optimal feature subset, we used a special week for our experiment.

The effect of temperature on the loads in summer and winter is large, and severe fluctuations make accurate forecasting difficult. Therefore, two weeks were chosen randomly from the summer and winter of 2013 for testing. The summer period was from 28 July to 3 August, and the winter period was from 22 to 28 December. As shown in Figure 5, the predictive load of each combined

method was fit with the true summer load. The average error of the various methods are shown in Table 6. The top-three combined methods were CART + SVR, RreliefF + RF, and RreliefF + SVR, and the MAPEs were 3.634%, 3.710%, and 4.204%, respectively. The predictive load of each combined method in winter is shown in Figure 6, each of the predicted loads matched the actual load except for Tuesday and Wednesday which corresponded to Christmas day and the day before. As is shown in Table 7, the first three combined methods were RreliefF + SVR, CART + GPR, and CART + SVR, and the MAPEs were 4.207%, 4.754%, and 4.770%, respectively.

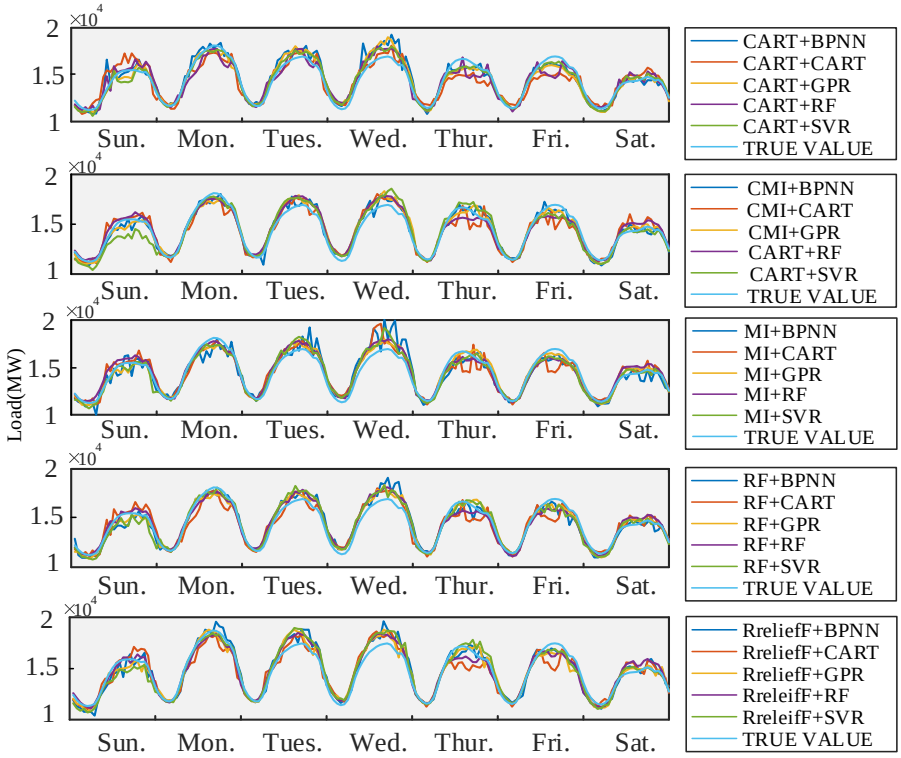


Figure 5. Prediction from 28 July to 3 August 2013.

Table 6. Comparison of different combined methods.

Method		CART	RF	GPR	BPNN	SVR
MI	**MAPE**	5.027	4.376	4.223	5.705	4.286
	MAE	849.194	732.926	709.848	962.649	720.361
	RMSE	1191.968	871.897	988.378	1323.916	921.862
CMI	**MAPE**	4.672	4.423	4.299	4.457	4.880
	MAE	784.550	719.337	699.910	566.609	809.988
	RMSE	1016.001	931.743	942.492	715.524	1027.936
CART	**MAPE**	6.179	4.936	4.449	4.910	3.634
	MAE	1034.009	833.712	752.653	823.088	599.284
	RMSE	1282.515	1077.501	961.304	1142.275	753.655
RF	**MAPE**	4.936	4.231	4.381	4.291	4.262
	MAE	833.712	711.268	815.776	711.438	705.789
	RMSE	1077.501	855.686	915.139	969.156	916.701
RreliefF	**MAPE**	4.577	3.710	4.239	4.270	4.204
	MAE	786.561	629.120	717.094	710.419	700.174
	RMSE	1072.662	781.775	1045.609	922.320	910.103

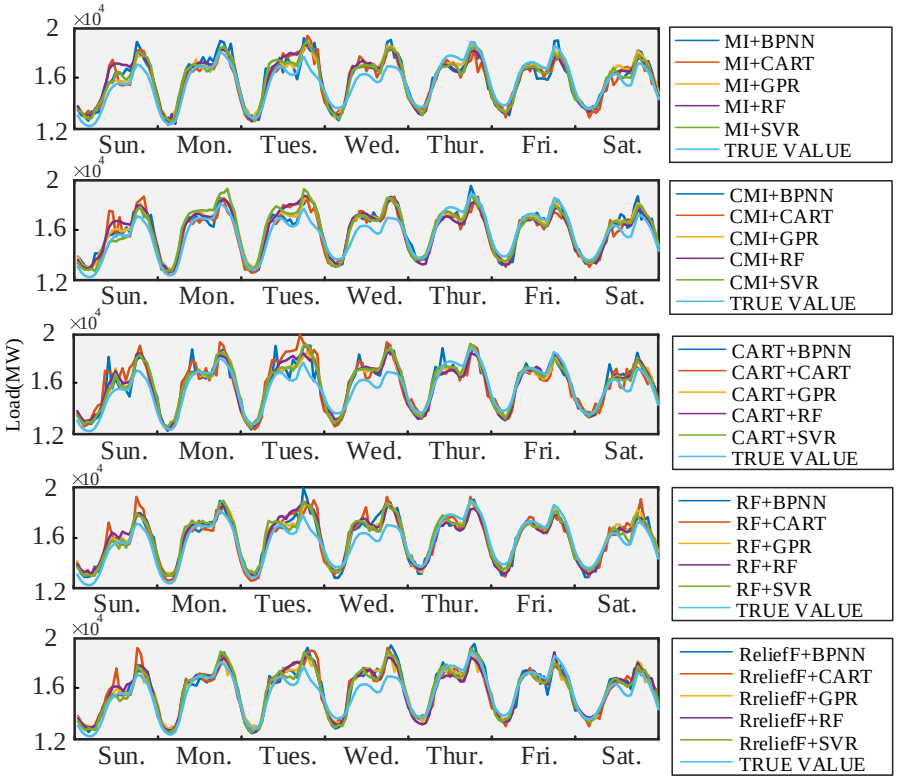


Figure 6. Prediction from 22 to 28 December 2013.

Table 7. Comparison of different combined methods.

Method		CART	RF	GPR	BPNN	SVR
MI	**MAPE**	5.420	5.783	4.862	5.823	4.977
	MAE	809.153	855.560	706.073	868.632	734.877
	RMSE	1052.017	1038.861	875.357	1059.056	897.331
CMI	**MAPE**	5.479	5.515	4.862	5.072	5.262
	MAE	814.890	821.482	710.464	733.701	788.030
	RMSE	1029.141	983.674	867.158	941.800	956.224
CART	**MAPE**	6.876	5.154	4.754	5.206	4.770
	MAE	1027.157	763.678	704.088	776.566	705.472
	RMSE	1307.768	1031.547	892.356	1055.224	911.921
RF	**MAPE**	5.154	5.421	4.817	5.190	4.540
	MAE	763.678	795.999	697.702	757.221	666.295
	RMSE	1031.547	955.704	858.767	961.667	849.553
RreliefF	**MAPE**	4.985	4.830	5.026	4.689	4.207
	MAE	741.379	713.534	749.809	702.243	628.159
	RMSE	1019.697	893.103	1034.086	931.176	810.417

For the full verification of different method combinations, the entire test set was used for the contrast experiment. The results of different evaluated criteria for the proposed forecasting approach applied by 25 method combinations are presented for day-ahead load forecasting in Table 8. The forecast errors of the different methods varied. For example, the error of MI-based SVR was close to that of the GPR. The MAPEs for the MI-based SVR and GPR were 4.872% and 4.785%, the RMSEs were 1196.775 MW and 1141.372 MW, and the MAEs were 773.447 MW and 755.325 MW, respectively. Based on these observations, the forecast errors of the SVR with any feature selection method was

below 5% (marked in bold) except with RF. In addition, the MAPEs of GPR with CMI and RF were below 5% as well.

Table 8. Error of load forecasting of different methods with proposed forecasting approach for the whole test set.

Feature Selection Method	Forecaster	Evaluated Criterion		
		MAPE (%)	RMSE (MW)	MAE (MW)
MI	CART	6.021	1360.445	934.560
	RF	5.536	1260.281	864.385
	SVR	**4.872**	1196.775	773.447
	BPNN	5.491	1320.809	865.842
	GPR	**4.785**	1141.372	755.325
CMI	CART	6.088	1371.643	945.217
	RF	5.364	1235.216	841.376
	SVR	**4.870**	1225.231	776.654
	BPNN	5.054	1179.931	793.064
	GPR	**4.758**	1135.260	750.937
CART	CART	6.495	1493.344	1013.322
	RF	5.364	1228.542	837.765
	SVR	**4.794**	1158.022	758.601
	BPNN	5.414	1270.671	847.104
	GPR	5.018	1176.996	790.088
RF	CART	5.883	1322.730	911.334
	RF	5.385	1236.724	843.334
	SVR	5.534	1260.281	834.385
	BPNN	5.287	1248.014	827.752
	GPR	**4.839**	1244.614	761.119
RreliefF	CART	5.804	1898.190	1305.192
	RF	5.202	1220.145	816.788
	SVR	**4.746**	1229.229	759.143
	BPNN	5.175	1244.537	812.642
	GPR	5.543	1410.293	883.576

By comparison of the results, the RreliefF + SVR method showed the best performance with the least MAPE.

6.2. Load Forecasting for Singapore

To further verify the applicability of the proposed approach, the load data from Singapore was used to perform the load forecasting experiments.

6.2.1. Feature Selection for Hour Loads

First, using the same method used in Section 6.1.1, the score of the feature corresponding to the predictive target at different hours was computed by different feature selection methods. Then, the optimal feature subset was obtained based on the MAPE of different subsets forecast by a predictor.

Table 9 shows the MAPE and the dimension of the optimal feature subset corresponding to using MI as the feature selection method and RF as the prediction model. The Table shows that 1 to 7 am, the dimension of optimal feature subset is less than at 8 to 19 pm, as the same as the forecasting error. Similar to the analysis result of 4, this is because people are less active at night and there are fewer factors affecting the load than during the day.

Table 9. Optimal feature subset construction of different hours with MI + RF for Singapore.

Time	MAPE	FD	Time	MAPE	FD
1	1.349	72	13	2.353	49
2	1.138	64	14	2.376	42
3	1.112	61	15	2.387	48
4	1.137	66	15	2.486	44
5	1.201	79	17	2.534	57
6	1.453	75	18	2.258	62
7	1.836	57	19	2.049	49
8	2.229	55	20	1.793	43
9	2.389	55	21	1.632	64
10	2.359	52	22	1.526	59
11	2.379	59	23	1.485	45
12	2.332	58	24	1.529	55

The MAPE and the dimension of optimal feature subset corresponding to 1:00 were carried out by different feature selection methods and forecasting methods shown in Table 10. The MAPEs are in 1.0% to 1.6% which means the performance of forecasters were similar after feature selection. While by analysis the feature dimension, we could find there is huge difference between the number of the feature of the optimal feature subset and that selected by different feature selection methods, which is caused by the different evaluation criteria.

Table 10. Optimal feature subset construction of 1:00 with different methods for Singapore.

Method	CART		RF		SVR		ANN		GPR	
	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD
MI	1.595	59	1.349	72	1.225	74	1.349	47	1.170	75
CMI	1.528	11	1.266	31	1.209	43	1.239	17	1.148	122
CART	1.559	14	1.303	26	1.103	56	1.210	16	1.169	60
RF	1.594	72	1.371	5	1.186	58	1.242	17	1.163	38
RreliefF	1.530	7	1.300	10	1.197	21	1.242	17	1.159	95

As is shown in Appendix A Table A3 to Table A4, considering both the MAPEs and the dimensions, the optimal feature subsets were used for the load forecasting of the Singapore data. Similar to the conclusion summarized in Table 4, the different optimal feature subsets employed various feature selection methods and forecasters.

6.2.2. Forecasting Results of Method Combinations with Optimal Feature Subsets for Singapore Load Data

To test the performance of diverse combined methods with the optimal feature subset, the data of special weeks were used for the experiment.

Two weeks were chosen randomly from the summer and winter of 2015 for testing as is shown in Figures 7 and 8. The summer week included the days from 21 to 27 June and the winter week included days from 8 to 14 November. The results are shown in Figure 7 and Table 11. It was found that the GPR, RF, and SVR methods showed a better ability to forecast the summer loads. The MAPEs of the combinations of MI + GPR, CMI + GPR, RF + GPR, RreliefF + GPR, CMI + SVR, and RreliefF + SVR were less than 1.5%. The outstanding combined method was RreliefF + GPR whose MAPE was 1.402%, MAE was 74.400 MW, and RMSE was 93.092 MW. By observing Figure 8 and Table 12, the RreliefF + GPR method showed the best performance with an MAPE of 3.567%, an MAE of 200.711 MW, and an RMSE of 224.017 MW. The predictive results of GPR and SVR with different feature selection methods were better than those of the CART, BPNN, and RF methods.

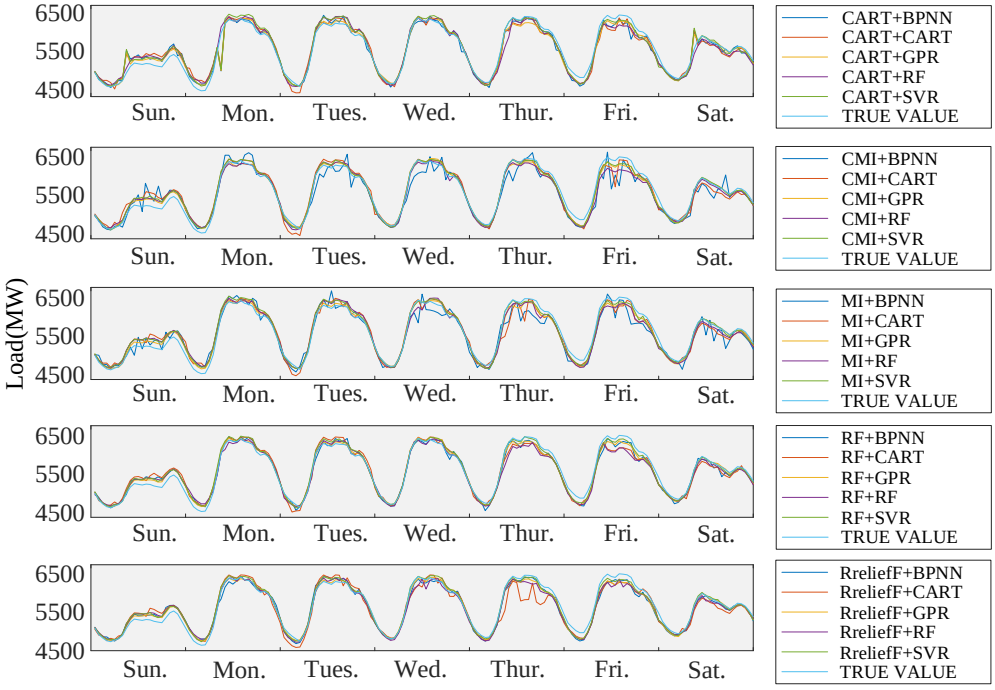


Figure 7. Prediction from 21 to 27 June 2015.

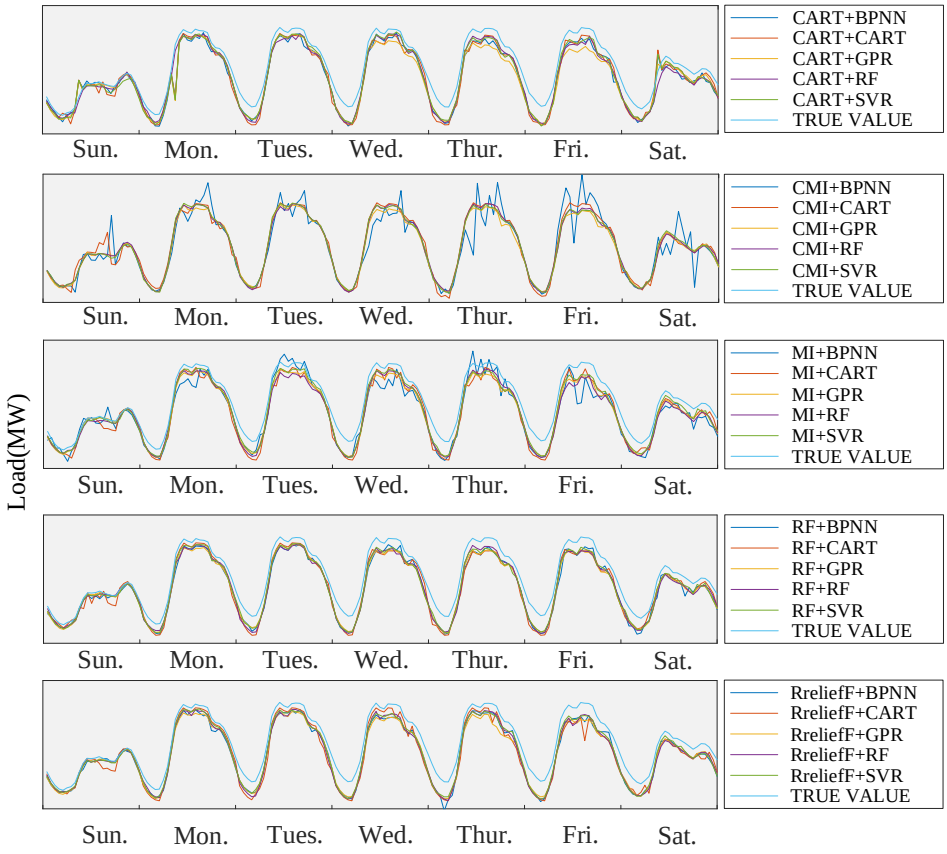


Figure 8. Prediction from 8 to 14 November 2015.

Table 11. Comparison of different combined methods.

Method		CART	RF	GPR	BPNN	SVR
MI	**MAPE**	2.321	2.145	1.439	2.719	1.662
	MAE	128.596	119.058	79.453	153.693	91.493
	RMSE	162.801	137.462	99.346	202.410	110.360
CMI	**MAPE**	2.117	1.867	1.419	3.165	1.482
	MAE	115.395	103.810	78.407	180.786	81.177
	RMSE	150.781	134.390	99.425	322.873	102.596
CART	**MAPE**	2.420	2.136	1.645	1.963	1.911
	MAE	132.823	118.571	91.358	108.851	106.369
	RMSE	175.615	143.584	139.408	160.930	152.349
RF	**MAPE**	2.213	2.000	1.435	1.702	1.404
	MAE	123.568	112.369	77.803	94.085	77.236
	RMSE	148.988	146.759	97.686	117.295	95.627
RreliefF	**MAPE**	2.720	1.862	1.428	1.917	1.402
	MAE	154.605	103.586	79.134	105.902	74.400
	RMSE	201.458	128.291	101.035	127.631	93.092

Table 12. Comparison of different combined methods.

Method		CART	RF	GPR	BPNN	SVR
MI	**MAPE**	3.895	3.854	3.806	4.273	3.637
	MAE	217.339	217.647	215.942	243.454	204.362
	RMSE	250.640	240.934	236.196	283.816	232.913
CMI	**MAPE**	3.573	3.518	3.899	5.023	3.585
	MAE	200.803	197.387	221.095	288.472	200.942
	RMSE	229.837	217.891	239.055	390.638	229.780
CART	**MAPE**	3.868	3.587	4.115	3.897	3.599
	MAE	215.523	200.915	234.630	219.650	201.124
	RMSE	260.178	225.501	272.193	254.684	235.158
RF	**MAPE**	3.799	3.711	3.851	3.871	3.599
	MAE	212.788	209.019	218.327	218.218	201.083
	RMSE	245.087	231.296	236.664	241.936	230.831
RreliefF	**MAPE**	3.981	3.895	4.104	3.935	3.567
	MAE	222.013	219.243	233.919	221.717	200.711
	RMSE	262.683	242.705	254.076	247.552	224.017

To further verify the superiority of the proposed method based on feature subsets of different hours, the entire test data from Singapore was used for validation. Detailed information about the test data is shown in Table 2 in Section 5.2. Table 13 shows the average predictive error of the different combined methods. It indicates that, based on MI, the CMI, RF, RreliefF, and SVR methods achieved the minimum errors with MAPEs of 1.471%, 1.440%, 1.387%, and 1.373%, respectively. Of all the combined methods, the RreliefF + SVR method worked best with an MAPE of 1.373%, an MAE of 75.118 MW, and an RMSE of 147.585 MW.

Table 13. Error of load forecasting of different methods with proposed forecasting strategy for the whole test set.

Feature Selection Method	Forecaster	Evaluated Criterion		
		MAPE (%)	RMSE (MW)	MAE (MW)
MI	CART	2.019	172.293	112.003
	RF	1.668	157.946	92.817
	SVR	**1.474**	**154.191**	**80.67**
	BPNN	2.551	218.916	145.116
	GPR	1.492	147.726	82.693
CMI	CART	2.174	189.964	121.050
	RF	1.623	156.450	90.309
	SVR	**1.440**	**151.230**	**78.764**
	ANN	3.072	332.424	177.185
	GPR	1.538	148.127	85.497
CART	CART	2.219	201.990	123.030
	RF	**1.733**	**164.604**	**96.589**
	SVR	1.748	188.225	96.562
	BPNN	1.954	192.515	109.282
	GPR	1.774	183.266	99.119
RF	CART	2.012	172.188	111.418
	RF	1.641	160.659	91.235
	SVR	**1.387**	**148.926**	**75.885**
	BPNN	1.663	158.088	92.355
	GPR	1.461	145.833	81.011
RreliefF	CART	2.075	177.441	116.199
	RF	1.608	155.962	89.551
	SVR	**1.373**	**147.585**	**75.118**
	BPNN	1.669	157.988	92.890
	GPR	1.446	144.170	80.283

By analyzing the load forecasting results for Singapore, the combination of RreliefF and SVR was the most accurate method.

6.3. Comparison and Discussion

6.3.1. Comparison of Forecasting Methods without Feature Selection for New England and Singapore

In this section, a comparison of the proposed method and the traditional method (which only builds a single predictor for forecasting without feature selection) based on the data of New England and Singapore was carried out to verify the necessity of forecasting by a modular predictor.

The histograms of the error and the training time duration of different forecasting methods using New England data are displayed in Figure 9. As shown in Figure 9a, the MAPE of the SVR that adopted the proposed method was almost half that of the SVR using the traditional method. The MAPE of other predictors employing the proposed method without the feature selection step decreased in different levels compared with the predictors employing the traditional method. By analyzing the MAE in Figure 9b and the RMSE in Figure 9c, a similar conclusion can be obtained. In addition, it is noted that the model training time of the proposed method decreased because of the smaller modeling training set. Therefore, the decreased error and training time reflect the advantages of the proposed method and confirms the necessity of employing a modular predictor.

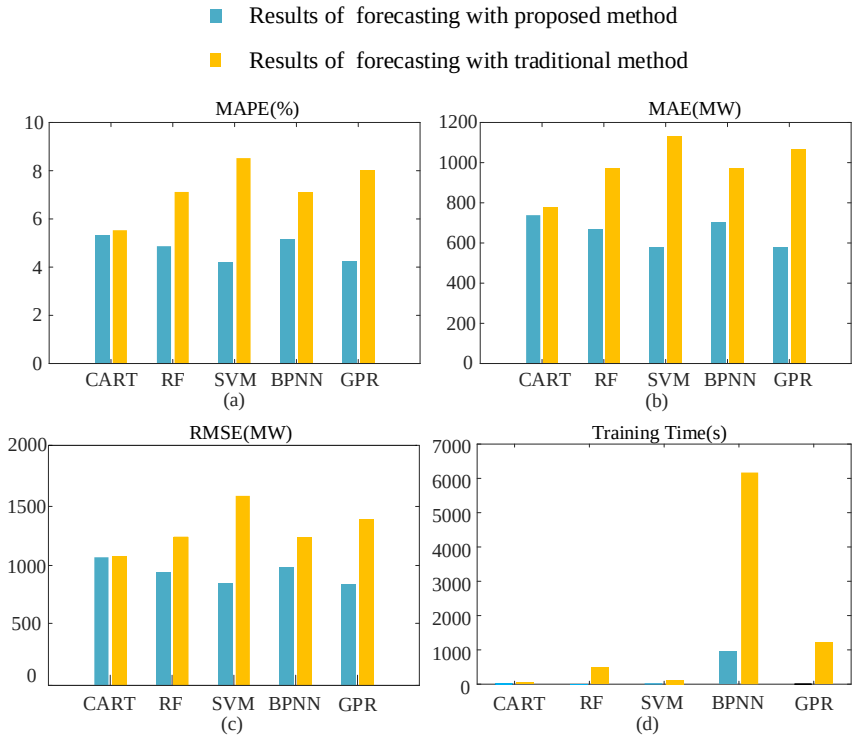


Figure 9. Comparison of error and time of training a model with traditional and proposed approaches.

The values of MAPE, MAE, and RMSE and the training time of each forecaster with different approaches based on the data of New England and Singapore are shown in Table 14. The results for New England indicate that the MAPE values of CART, RF, SVR, BPNN, and GPR with the proposed method were reduced by 0.182%, 2.253%, 4.294%, 1.953%, and 3.775% compared with the CART, RF, SVR, BPNN, and GPR with the traditional approach, respectively. Similarly, the results for Singapore also verified the superior performance of the proposed method.

Table 14. Comparison of the error of different forecasting approaches with original feature set.

Method	Forecaster	Test for New England				Test for Singapore			
		MAPE (%)	MAE (MW)	RMSE (MW)	Time (s)	MAPE (%)	MAE (MW)	RMSE (MW)	Time (s)
The proposed method	CART	5.348	738.641	1067.723	0.106	2.166	116.742	209.413	0.275
	RF	4.867	671.261	941.661	10.445	1.930	105.306	199.974	10.776
	SVR	**4.228**	**580.80**	**849.806**	**0.431**	**1.914**	**103.356**	**196.145**	**0.405**
	BPNN	5.167	705.324	986.974	962.457	3.133	174.104	285.083	844.257
	GPR	4.242	581.889	844.0766	2.102	1.523	82.573	170.478	1.569
The traditional method	CART	5.530	778.083	1076.316	7.976	3.597	196.391	273.112	2.601
	RF	7.120	975.272	1235.783	486.263	2.088	114.732	209.064	402.743
	SVR	8.522	1130.870	1556.371	123.394	5.067	267.048	361.547	91.623
	BPNN	7.120	975.272	1235.783	6170.835	4.864	267.416	408.305	4686.007
	GPR	8.017	1065.700	1387.252	1219.056	5.072	287.277	405.181	1054.359

6.3.2. Comparison of Forecasting Approaches with Feature Selection for New England and Singapore

A comparison between the proposed method and traditional method with feature selection was performed on the New England and Singapore datasets. The results of the proposed method with feature selection are shown in Table 8 (New England) and Table 13 (Singapore), and the results of the traditional method with feature selection are shown in Table 15. The results indicate that the error was reduced in different levels by adopting the proposed method. The largest reduction in

MAPE resulted from the CMI + SVR and CART + BPNN methods with MAPEs of 2.799% and 3.072%, respectively. The minimum error was achieved by the RreliefF + SVR combination with MAPEs of 4.746% (New England) and 1.373% (Singapore). In conclusion, the forecasted results obtained by the proposed method were better than those of the traditional method regardless of the predictor used. The most accurate combined method was RreliefF + SVR.

Table 15. Error of load forecasting of different methods with traditional forecasting approach for the whole test set.

Feature Selection Method	Forecaster	Test for New England			Test for Singapore		
		MAPE (%)	MAE (MW)	RMSE (MW)	MAPE (%)	MAE (MW)	RMSE (MW)
MI	CART	8.452	1269.711	1701.808	3.247	178.082	239.891
	RF	5.911	920.201	1339.227	1.855	103.612	168.744
	SVR	7.587	1116.529	1521.691	4.246	222.376	314.547
	BPNN	5.854	909.553	1390.574	2.103	115.674	176.764
	GPR	5.680	881.310	1296.119	2.161	118.833	180.018
CMI	CART	8.420	1267.213	1705.926	3.320	182.186	241.884
	RF	5.645	878.479	1281.361	1.838	102.269	164.790
	SVR	7.669	1134.308	1560.853	4.206	219.965	313.680
	BPNN	7.697	1173.160	1929.675	2.053	113.328	173.493
	GPR	6.562	1029.708	1558.377	2.104	115.482	175.308
CART	CART	8.420	1267.213	1705.926	3.212	175.834	238.396
	RF	5.970	921.976	1318.980	1.940	108.871	175.704
	SVR	7.635	1127.940	1506.504	4.170	217.423	312.793
	BPNN	6.044	922.497	1404.137	5.026	278.908	462.199
	GPR	5.904	920.084	1372.375	2.860	161.948	250.843
RF	CART	8.056	1212.114	1653.079	3.262	179.431	242.135
	RF	5.483	858.934	1306.136	1.833	102.516	167.013
	SVR	7.316	1081.404	1482.864	4.147	216.703	310.201
	BPNN	5.348	831.493	1196.481	1.790	99.181	160.264
	GPR	5.774	902.872	1321.057	1.951	108.686	169.148
RreliefF	CART	8.056	1212.114	1653.079	3.188	174.799	237.592
	RF	5.506	866.377	1333.577	2.003	111.688	176.002
	SVR	7.350	1081.259	1464.366	4.319	226.854	319.170
	BPNN	5.789	894.686	1320.901	1.958	107.762	168.592
	GPR	6.015	967.163	1682.298	2.130	117.884	188.138

7. Conclusions

Accurate day-ahead load forecasting enhances the stability of grid operations and improves the social benefits of power systems. To improve the accuracy of day-ahead load forecasting, a novel modular parallel forecasting model with feature selection was proposed. Load data from New England and Singapore were used to test the proposed method. The experimental results show the advantages of the proposed method as follows:

(1) A modular predictor consisting of 24 independent predictors can efficiently capture load characteristics with respect to different hours and thereby avoid the inaccurate analysis of a single predictor.

(2) The feature selection adopted for the load corresponding to different hours analyzes the relevance between the feature and a special load. Each optimal feature subset of different dimension benefits the building of a more-accurate predictor.

(3) To serve the demand of dispatch departments of different regions, the interval time $p = 24$ was chosen for structuring a general candidate feature set that met the requirements of the power system.

Future work will concentrate on predictor parameter optimization and improve the efficiency of forecasting in the modeling process and applying the proposed method to probabilistic load forecasting.

Author Contributions: L.L. put forward to the main idea and design the whole venation of this paper. L.X. and Z.H. did the experiments and prepared the manuscript. N.H. guided the experiments and paper writing. All authors have read and approved the final manuscript.

Acknowledgments: This work is supported by the National Nature Science Foundation of China (No. 51307020), the Science and Technology Development Project of Jilin Province (No. 20160411003XH), the Science and Technology Project of Jilin Province Education Department (No. JJKH20170219KJ), Major science and technology projects of Jilin Institute of Chemical Technology (No. 2018021), and Science and Technology Innovation Development Plan Project of Jilin City (No. 201750239).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Optimal feature subset construction of different hours from 1:00 to 12:00 with different methods for New England.

Time Point		1:00		2:00		3:00		4:00		5:00		6:00		7:00		8:00		9:00		10:00		11:00		12:00	
Error		MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD
MI	CART	3.741	7	3.769	2	4.071	12	4.083	84	4.472	40	5.140	6	4.949	164	4.748	164	4.627	164	4.765	15	4.978	21	5.529	130
	RF	3.294	34	3.419	22	3.632	9	3.783	30	4.008	9	4.828	18	5.456	61	5.314	59	4.526	64	4.171	45	4.147	42	4.414	67
	SVR	3.064	10	3.167	28	3.189	27	3.314	23	3.553	59	3.852	38	4.353	57	4.327	47	3.682	61	3.510	58	3.598	72	3.789	23
	ANN	3.226	8	3.329	7	3.422	26	3.897	27	3.521	30	4.215	16	4.889	40	4.345	29	3.891	23	3.848	31	3.911	11	4.342	32
	GPR	3.087	119	3.226	115	3.359	9	3.381	99	3.629	96	4.087	36	4.476	54	4.432	102	3.805	65	3.645	49	3.781	36	3.852	31
CMI	CART	3.729	2	3.769	2	4.058	7	4.192	16	4.296	8	5.242	6	4.926	99	4.523	27	5.076	15	4.523	27	5.076	15	5.413	106
	RF	3.447	20	3.447	23	3.590	12	3.717	13	3.848	6	4.505	57	4.750	40	4.531	130	4.136	49	3.949	159	4.009	159	4.281	124
	SVR	3.043	13	3.126	12	3.238	12	3.341	4	3.375	48	3.722	91	4.008	60	3.972	73	3.469	88	3.351	68	3.448	93	3.667	88
	ANN	3.062	47	3.123	42	3.134	28	3.329	53	3.365	23	3.821	64	4.178	83	4.167	35	3.590	78	3.341	75	3.418	57	3.576	63
	GPR	3.052	134	3.189	23	3.288	18	3.366	16	3.593	21	4.017	18	4.128	150	3.911	158	3.517	168	3.352	91	3.455	106	3.612	88
CART	CART	3.729	3	4.050	6	4.071	4	4.134	5	4.558	6	5.596	13	4.958	9	4.751	4	4.634	7	4.725	10	4.524	19	5.512	10
	RF	3.422	11	3.511	5	3.589	6	3.615	12	3.963	7	4.511	32	4.512	20	4.367	11	4.062	22	3.989	18	4.151	21	4.546	11
	SVR	3.068	11	3.167	11	3.548	11	3.433	12	3.798	14	3.846	33	3.804	20	3.870	18	3.524	22	3.629	19	3.633	20	4.260	18
	ANN	3.270	9	3.301	11	3.670	5	3.483	12	3.836	13	4.012	15	3.974	19	4.081	16	3.921	17	3.775	17	3.806	18	4.397	11
	GPR	3.245	8	3.280	11	3.526	11	3.458	8	3.858	8	3.911	33	3.753	20	3.872	18	3.707	22	3.659	19	3.732	16	4.360	11
RF	CART	3.741	7	3.769	2	4.059	9	4.128	44	4.552	9	5.084	7	4.807	52	4.656	5	4.337	6	4.464	10	4.147	3	4.155	6
	RF	3.533	51	3.679	27	3.790	28	3.777	11	3.991	25	4.522	30	4.624	12	4.416	9	3.973	26	3.922	15	3.801	28	4.227	7
	SVR	3.140	41	3.512	14	3.312	42	3.469	11	3.5518	26	3.6554	19	3.9069	27	3.8594	46	3.3732	31	3.3966	44	3.376	43	3.7329	51
	ANN	3.069	18	3.255	20	3.097	21	3.469	17	3.486	16	3.816	27	4.262	11	4.278	13	3.682	11	3.542	19	3.424	31	3.923	14
	GPR	3.099	81	3.239	80	3.338	64	3.447	150	3.632	16	3.679	19	4.208	15	3.964	56	3.522	86	3.359	63	3.484	43	3.787	37
RreliefF	CART	3.741	10	3.769	2	4.059	8	4.156	42	4.475	20	4.917	4	4.729	38	4.646	15	4.443	7	4.030	15	4.417	4	4.448	20
	RF	3.310	26	3.390	20	3.466	17	3.560	19	3.528	14	3.764	14	4.534	30	4.294	23	3.643	10	3.514	17	3.680	22	4.006	18
	SVR	3.043	18	3.107	19	3.233	19	3.351	30	3.434	16	3.928	14	3.716	34	3.648	21	3.320	34	3.205	34	3.407	66	3.594	53
	ANN	3.269	9	3.306	14	3.338	13	3.368	17	3.445	16	3.555	15	4.193	35	3.760	23	3.399	19	3.329	20	3.427	25	3.820	24
	GPR	3.019	134	3.156	152	3.329	32	3.346	117	3.460	16	3.578	29	3.811	34	3.715	24	3.414	22	3.333	28	3.358	81	3.807	29

Table A2. Optimal feature subset construction of different hours from 13:00 to 24:00 with different methods for New England.

Time Point		13:00		14:00		15:00		16:00		17:00		18:00		19:00		20:00		21:00		22:00		23:00		24:00	
Error		MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPEE	FD	MAPEE	FD	MAPEE	FD	MAPE	FD	MAPEE	FD
MI	CART	4.654	164	5.143	164	5.764	164	5.406	164	6.170	41	6.313	6	6.066	12	5.388	9	5.166	14	5.105	12	5.198	6	5.086	8
	RF	4.663	41	4.926	33	5.190	38	5.351	46	5.547	31	5.358	98	5.117	136	4.506	23	4.376	28	4.779	9	4.794	41	4.847	72
	SVR	3.927	23	4.263	35	4.456	49	4.518	82	4.645	68	4.516	69	4.418	69	4.165	21	3.867	24	3.901	73	3.860	154	4.039	115
	ANN	4.042	19	4.362	14	4.626	35	4.799	29	4.746	37	5.262	46	4.849	33	4.371	15	3.919	27	4.599	42	4.261	27	4.203	35
	GPR	3.974	24	4.177	36	4.524	29	4.689	28	4.616	32	4.823	21	4.770	32	4.272	24	4.010	26	4.157	24	4.291	22	4.446	18
CMI	CART	5.459	14	5.045	53	5.519	24	5.406	91	5.985	18	6.382	5	5.973	5	5.299	19	5.134	20	5.178	19	5.164	8	4.994	14
	RF	4.445	146	4.608	151	4.843	150	5.064	162	5.326	164	5.352	157	5.136	126	4.496	136	4.394	126	4.176	145	4.731	153	4.804	133
	SVR	3.923	107	4.102	97	4.339	99	4.539	86	4.501	105	4.540	70	4.398	87	3.944	113	3.717	90	3.763	94	3.895	93	3.972	111
	ANN	3.846	64	3.827	53	4.304	54	4.452	49	4.698	86	4.500	72	4.416	94	4.273	156	3.780	32	3.943	76	3.971	78	3.902	117
	GPR	3.916	105	4.080	40	4.474	53	4.686	94	4.469	172	4.455	168	4.709	76	4.508	16	4.116	39	4.075	49	4.304	103	4.432	118
CART	CART	4.655	11	5.141	8	5.768	12	5.412	9	6.320	11	6.955	8	6.393	9	5.978	13	5.324	11	5.742	9	5.214	6	5.182	7
	RF	4.587	19	4.846	15	5.113	28	5.409	22	5.390	25	5.720	22	5.444	17	4.634	15	4.665	24	5.001	29	4.700	7	4.806	7
	SVR	4.105	15	4.371	21	4.393	28	4.882	22	4.804	24	5.515	22	4.804	26	4.236	30	4.099	24	4.256	21	4.228	18	4.168	20
	ANN	4.205	15	4.713	18	4.666	18	4.931	20	5.248	16	5.205	22	4.986	26	4.302	26	4.248	24	4.441	18	4.152	17	3.988	22
	GPR	4.174	13	4.507	21	4.573	28	5.109	22	5.054	23	5.124	21	4.864	28	4.424	32	4.304	21	4.545	18	4.405	17	4.321	21
RF	CART	4.448	12	5.022	23	5.476	6	5.350	7	5.678	5	6.129	5	6.040	22	5.273	11	4.820	26	5.175	98	4.980	5	5.080	45
	RF	4.445	7	4.617	55	4.845	109	5.045	52	5.226	73	5.296	70	5.101	85	4.484	68	4.450	76	4.780	75	4.711	64	4.810	50
	SVR	3.971	100	3.983	35	4.122	82	4.635	94	4.556	32	4.589	40	4.725	101	4.089	37	7.848	25	3.847	97	3.936	90	3.996	141
	ANN	4.101	17	4.638	16	4.585	32	4.882	13	5.205	17	5.275	11	5.227	11	4.413	17	3.890	21	4.411	10	4.328	23	4.336	19
	GPR	4.069	48	4.091	74	4.515	42	4.662	41	4.912	32	4.577	159	4.886	56	4.491	43	4.026	70	4.182	81	4.367	79	4.497	18
RreliefF	CART	4.574	20	4.949	21	5.529	29	5.405	108	5.817	22	6.238	23	5.860	17	5.321	24	4.796	37	4.721	23	5.005	9	4.986	20
	RF	4.365	17	4.716	20	4.908	74	5.065	136	5.293	15	5.166	16	5.029	81	4.403	109	4.368	140	4.613	14	4.612	20	4.619	16
	SVR	3.693	65	3.876	66	4.104	52	4.426	49	4.496	67	4.528	41	4.558	68	3.983	52	3.809	44	3.803	109	4.008	43	4.056	38
	ANN	4.089	19	4.246	26	4.644	20	4.902	39	5.000	25	5.216	16	5.161	31	4.633	30	4.379	23	4.218	14	4.566	24	4.291	21
	GPR	3.980	40	4.161	34	4.411	33	4.632	36	4.748	45	4.472	160	4.391	170	4.301	45	3.976	43	4.284	34	4.001	172	4.324	16

Remark. The FD in Table 4 means feature dimension.

Table A3. Optimal feature subset construction of different hours from 1:00 to 12:00 with different methods for Singapore.

Time Point		1:00		2:00		3:00		4:00		5:00		6:00		7:00		8:00		9:00		10:00		11:00		12:00	
Error		MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD
MI	CART	1.595	59	1.479	57	1.402	4	1.482	10	1.535	6	1.624	108	2.041	29	2.727	62	2.515	48	2.526	43	2.615	59	2.451	58
	RF	1.349	72	1.138	64	1.112	61	1.137	66	1.201	79	1.453	75	1.836	57	2.229	55	2.389	55	2.359	52	2.379	59	2.332	58
	SVR	1.225	74	1.057	61	1.056	58	1.025	57	1.114	59	1.377	90	1.401	57	1.565	43	1.749	44	1.723	55	1.796	58	1.738	57
	ANN	1.349	47	1.179	56	1.133	10	1.276	59	1.262	76	1.453	33	1.558	30	1.926	48	2.323	58	2.319	53	2.321	48	2.424	21
	GPR	1.170	75	0.955	109	0.904	92	0.963	86	1.025	110	1.281	101	1.452	94	1.859	110	2.021	113	2.039	41	1.952	98	1.876	99
CMI	CART	1.528	11	1.470	4	1.386	4	1.396	9	1.475	4	1.632	62	1.998	72	2.752	108	2.709	135	2.534	124	2.531	114	2.485	125
	RF	1.266	31	1.093	15	1.305	33	1.052	30	1.133	50	1.416	44	1.847	23	2.191	49	2.333	42	2.345	38	2.353	43	2.237	43
	SVR	1.209	43	1.027	28	0.950	33	1.082	35	1.123	55	1.316	65	1.387	39	1.508	34	1.651	52	1.732	33	1.692	44	1.631	44
	ANN	1.239	17	1.062	21	1.034	28	1.072	31	1.149	29	1.312	27	1.542	23	1.979	31	2.202	23	2.107	28	2.037	24	2.129	47
	GPR	1.148	122	0.930	138	0.882	163	0.900	167	1.037	73	1.090	122	1.470	27	1.879	36	2.054	43	2.012	48	1.925	135	1.892	50
CART	CART	1.559	14	1.528	14	1.403	3	1.482	12	1.608	2	1.675	143	2.323	50	2.718	7	2.950	18	2.875	19	2.593	102	2.478	102
	RF	1.303	26	1.113	12	1.100	21	1.105	32	1.214	25	1.431	38	1.917	44	2.289	19	2.417	45	2.458	45	2.544	56	2.477	19
	SVR	1.103	56	0.995	36	1.031	17	1.001	73	1.038	22	1.138	32	1.575	83	1.653	25	1.803	73	1.863	46	1.916	53	1.845	53
	ANN	1.210	16	1.061	21	1.043	15	1.037	24	1.122	32	1.252	31	1.876	22	1.981	27	2.082	23	2.158	19	2.254	28	2.274	13
	GPR	1.169	60	1.015	74	0.902	120	0.918	115	1.066	25	1.216	33	1.415	173	1.813	172	1.959	172	1.903	172	1.893	172	1.851	173
RF	CART	1.594	72	1.583	19	1.425	2	1.525	11	1.577	9	1.672	14	2.147	8	2.456	4	2.754	10	2.710	21	2.615	42	2.488	29
	RF	1.371	5	1.089	24	1.081	22	1.105	31	1.190	35	1.381	21	1.608	10	2.042	9	2.336	10	2.239	18	2.334	41	2.238	19
	SVR	1.186	58	0.993	13	0.904	30	0.972	38	1.035	27	1.080	39	1.370	20	1.483	27	1.617	30	1.588	29	1.682	30	1.649	23
	ANN	1.242	17	1.016	18	0.926	28	1.014	44	1.033	24	1.149	23	1.461	11	1.689	14	1.945	25	1.930	31	1.953	9	1.952	15
	GPR	1.163	38	0.949	76	0.897	76	0.897	60	0.946	105	1.115	39	1.427	27	1.835	30	1.982	31	1.966	26	1.952	35	1.882	31
RreliefF	CART	1.530	7	1.506	9	1.283	10	1.395	8	1.513	9	1.574	13	1.754	24	2.579	38	2.579	38	2.464	40	2.412	43	2.295	42
	RF	1.300	10	1.083	18	1.042	38	1.070	31	1.178	12	1.173	14	1.448	18	2.109	19	2.248	57	2.216	59	2.204	64	2.191	9
	SVR	1.197	21	1.019	13	1.003	14	1.047	14	1.098	59	1.080	26	1.343	38	1.464	24	1.620	42	1.671	43	1.679	42	1.678	34
	ANN	1.242	17	1.016	18	0.926	28	1.014	44	1.033	24	1.149	23	1.645	11	1.689	14	1.945	25	1.930	31	1.953	9	1.952	15
	GPR	1.159	95	0.950	94	0.910	95	0.925	96	0.989	88	1.128	16	1.442	22	1.780	18	1.886	34	1.901	37	1.876	41	1.778	42

Table A4. Optimal feature subset construction of different hours from 13:00 to 24:00 with different methods for Singapore.

Time Point		13:00		14:00		15:00		16:00		17:00		18:00		19:00		20:00		21:00		22:00		23:00		24:00	
Error		MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD	MAPE	FD
MI	CART	2.501	75	2.522	37	2.700	43	2.637	45	2.619	58	2.398	51	2.113	87	1.884	49	1.790	64	1.588	90	1.614	66	1.820	13
	RF	2.353	49	2.376	42	2.387	48	2.486	44	2.534	57	2.258	62	2.049	49	1.793	43	1.632	64	1.526	59	1.485	45	1.529	55
	SVR	1.795	49	1.850	42	1.878	43	1.898	43	2.010	54	1.883	54	1.629	111	1.469	39	1.317	94	1.372	40	1.337	41	1.301	75
	ANN	2.280	25	2.362	36	2.466	23	2.334	32	2.259	44	2.324	29	1.936	45	1.797	52	1.682	65	1.482	41	1.450	38	1.390	33
	GPR	1.912	95	2.036	40	2.032	43	2.098	43	2.095	102	1.821	170	1.752	101	1.554	44	1.357	94	1.304	91	1.283	102	1.253	119
CMI	CART	2.368	111	2.749	126	2.517	127	2.665	113	2.481	125	2.423	110	1.983	115	1.884	139	1.779	66	1.587	121	1.580	7	1.667	5
	RF	2.263	37	2.307	36	2.309	36	2.349	28	2.342	35	2.160	31	1.933	40	1.672	41	1.593	30	1.466	29	1.456	8	1.451	53
	SVR	1.702	49	1.768	34	1.792	45	1.914	42	1.937	41	1.806	33	1.646	32	1.442	27	1.359	33	1.301	41	1.324	54	1.352	37
	ANN	2.232	27	2.125	30	2.163	43	2.300	36	2.381	23	2.218	39	1.761	31	1.851	21	1.488	15	1.405	34	1.358	22	1.377	25
	GPR	1.913	48	1.978	45	1.994	44	2.055	40	2.054	102	1.884	128	1.667	171	1.448	172	1.296	172	1.244	171	1.239	125	1.282	92
CART	CART	2.486	103	2.557	4	2.630	4	2.734	4	2.626	151	2.398	104	2.132	15	1.940	20	1.911	86	1.591	58	1.638	156	1.734	7
	RF	2.414	22	2.463	18	2.493	46	2.622	7	2.653	8	2.352	17	1.909	20	1.903	27	1.760	24	1.448	25	1.499	23	1.503	45
	SVR	1.871	52	1.935	54	1.885	54	2.115	40	2.188	39	1.843	54	1.601	87	1.606	87	1.511	69	1.322	29	1.284	47	1.227	85
	ANN	2.218	19	2.213	18	2.202	19	2.387	19	2.404	31	2.179	19	1.832	32	1.773	19	1.754	12	1.438	33	1.453	22	1.351	17
	GPR	1.871	172	1.950	173	1.948	168	1.981	171	2.011	169	1.824	168	1.663	172	1.442	168	1.285	169	1.235	168	1.205	166	1.203	169
RF	CART	2.501	67	2.759	27	2.673	7	2.651	34	2.617	38	2.416	35	1.999	13	1.794	27	1.750	17	1.583	52	1.638	41	1.692	2
	RF	2.272	36	2.323	33	2.328	35	2.364	34	2.396	39	2.098	26	1.881	25	1.628	19	1.527	10	1.427	16	1.410	15	1.478	24
	SVR	1.670	28	1.751	23	1.857	22	1.853	39	1.944	36	1.789	16	1.577	57	1.394	22	1.299	56	1.240	15	1.195	12	1.244	38
	ANN	1.972	11	2.256	10	2.224	8	2.245	10	2.326	13	1.949	21	1.822	33	1.551	7	1.419	12	1.323	17	1.244	19	1.396	32
	GPR	1.876	44	1.981	30	2.125	34	2.023	54	2.108	42	1.893	91	1.767	35	1.502	32	1.377	13	1.269	16	1.244	18	1.265	47
RreliefF	CART	2.323	41	2.532	41	2.662	42	2.533	24	2.366	58	2.220	42	1.949	66	1.766	7	1.674	15	1.605	21	1.624	12	1.642	10
	RF	2.201	84	2.298	9	2.243	17	2.262	14	2.271	35	1.998	14	1.713	18	1.468	21	1.366	17	1.308	15	1.337	17	1.400	24
	SVR	1.714	41	1.768	20	1.799	37	1.801	15	1.815	14	1.646	23	1.564	20	1.409	11	1.258	16	1.265	24	1.252	24	1.299	21
	ANN	1.972	11	2.256	10	2.224	8	2.244	10	2.326	13	1.949	21	1.822	33	1.551	7	1.419	12	1.322	12	1.244	19	1.396	32
	GPR	1.856	29	1.882	41	1.861	31	1.937	41	1.936	30	1.789	33	1.645	95	1.450	35	1.316	56	1.296	41	1.279	72	1.208	170

References

1. He, Y.; Xu, Q.; Wan, J.; Yang, S. Short-term power load probability density forecasting based on quantile regression neural network and triangle kernel function. *Energy* **2016**, *114*, 498–512. [CrossRef]
2. Nikmehr, N.; Najafi-Ravadanegh, S. Optimal operation of distributed generations in micro-grids under uncertainties in load and renewable power generation using heuristic algorithm. *IET Renew. Power Gener.* **2015**, *9*, 982–990. [CrossRef]
3. Duan, Z.Y.; Gutierrez, B.; Wang, L. Forecasting Plug-In Electric Vehicle Sales and the Diurnal Recharging Load Curve. *IEEE Trans. Smart Grid* **2014**, *5*, 527–535. [CrossRef]
4. Ferlito, S.; Adinolfi, G.; Graditi, G. Comparative analysis of data-driven methods online and offline trained to the forecasting of grid-connected photovoltaic plant production. *Appl. Energy* **2017**, *205*, 116–129. [CrossRef]
5. Ferruzzi, G.; Cervone, G.; Delle Monache, L.; Graditi, G.; Jacobone, F. Optimal bidding in a Day-Ahead energy market for Micro Grid under uncertainty in renewable energy production. *Energy* **2016**, *106*, 194–202. [CrossRef]
6. Feng, Y.H.; Ryan, S.M. Day-ahead hourly electricity load modeling by functional regression. *Appl. Energy* **2016**, *170*, 455–465. [CrossRef]
7. Bindiu, R.; Chindris, M.; Pop, G.V. Day-Ahead Load Forecasting Using Exponential Smoothing. *Sci. Bull. Petru Maior Univ. Tîrgu Mureș* **2009**, *6*, 89–93.
8. Al-Hamadi, H.M.; Soliman, S.A. Fuzzy short-term electric load forecasting using Kalman filter. *IEE Proc.-Gener. Transm. Distrib.* **2012**, *153*, 217–227. [CrossRef]
9. Lee, C.M.; Ko, C.N. Short-term load forecasting using lifting scheme and ARIMA models. *Expert Syst. Appl.* **2011**, *38*, 5902–5911. [CrossRef]
10. Luy, M.; Ates, V.; Barisci, N.; Polat, H.; Cam, E. Short-Term Fuzzy Load Forecasting Model Using Genetic–Fuzzy and Ant Colony–Fuzzy Knowledge Base Optimization. *Appl. Sci.* **2018**, *8*, 864. [CrossRef]
11. Xiao, L.Y.; Shao, W.; Liang, L.L.; Wang, C. A combined model based on multiple seasonal patterns and modified firefly algorithm for electrical load forecasting. *Appl. Energy* **2016**, *167*, 135–153. [CrossRef]
12. Khotanzad, A.; Zhou, E.; Elragal, H. A neuro-fuzzy approach to short-term load forecasting in a price-sensitive environment. *IEEE Trans. Power Syst.* **2002**, *17*, 1273–1282. [CrossRef]
13. Felice, M.D.; Yao, X. Short-Term Load Forecasting with Neural Network Ensembles: A Comparative Study Application Notes. *IEEE Comput. Intell. Mag.* **2012**, *6*, 47–56. [CrossRef]
14. Ahmad, A.; Javaid, N.; Alrajeh, N.; Khan, Z.A.; Qasim, U.; Khan, A. A Modified Feature Selection and Artificial Neural Network-Based Day-Ahead Load Forecasting Model for a Smart Grid. *Appl. Sci.* **2015**, *5*, 1756–1772. [CrossRef]
15. Che, J.X.; Wang, J.Z.; Tang, Y.J. Optimal training subset in a support vector regression electric load forecasting model. *Appl. Soft Comput.* **2012**, *12*, 1523–1531. [CrossRef]
16. Dudek, G. Short-Term Load Forecasting Using Random Forests. *Intell. Syst.* **2015**, *323*, 821–828.
17. Lloyd, R.J. GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes. *Int. J. Forecast.* **2014**, *30*, 369–374. [CrossRef]
18. Che, J.X.; Wang, J.Z. Short-term load forecasting using a kernel-based support vector regression combination model. *Appl. Energy* **2014**, *132*, 602–609. [CrossRef]
19. Božić, M.; Stojanović, M.; Stajić, Z.; Stajić, N. Mutual Information-Based Inputs Selection for Electric Load Time Series Forecasting. *Entropy* **2013**, *15*, 926–942. [CrossRef]
20. Rong, G.; Liu, X. Support vector machine with PSO algorithm in short-term load forecasting. In Proceedings of the 2008 Chinese Control and Decision Conference, Yantai, China, 2–4 July 2008; pp. 1140–1142.
21. Ma, L.H.; Zhou, S.; Lin, M. Support Vector Machine Optimized with Genetic Algorithm for Short-Term Load Forecasting. In Proceedings of the International Symposium on Knowledge Acquisition and Modeling IEEE, Wuhan, China, 21–22 December 2008; pp. 654–657.
22. Zhang, Y.J.; Peng, X.Y.; Peng, Y.; Pang, J.Y.; Liu, D.T. Weighted bagging gaussion process regression to predict remaining useful life of electro-mechanical actuator. In Proceedings of the Prognostics and System Health Management Conference, Chengdu, China, 19–21 October 2016; pp. 1–6.
23. Lahouar, A.; Slama, J.B.H. Day-ahead load forecast using random forest and expert input selection. *Energy Convers. Manag.* **2015**, *103*, 1040–1051. [CrossRef]

24. Ghofrani, M.; West, K.; Ghayekhloo, M. Hybrid time series-bayesian neural network short-term load forecasting with a new input selection method. In Proceedings of the 2015 IEEE Power & Energy Society General Meeting, Denver, CO, USA, 26–30 July 2015; pp. 1–5.
25. Chandrashekar, G.; Sahin, F. *A Survey on Feature Selection Methods*; Pergamon Press, Inc.: New York, NY, USA, 2014.
26. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1996**, *97*, 273–324. [CrossRef]
27. Hu, Z.Y.; Bao, Y.K.; Chiong, R.; Xiong, T. Mid-term interval load forecasting using multi-output support vector regression with a memetic algorithm for feature selection. *Energy* **2015**, *84*, 419–431. [CrossRef]
28. Goldberg, D.E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley Longman Publishing Co., Inc.: Boston, MA, USA, 1990; pp. 2104–2116.
29. Hyojoo, S.; Kim, C. Forecasting Short-term Electricity Demand in Residential Sector Based on Support Vector Regression and Fuzzy-rough Feature Selection with Particle Swarm Optimization. *Procedia Eng.* **2015**, *118*, 1162–1168.
30. Isabelle, G.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
31. Koprinska, I.; Rana, M.; Agelidis, V.G. Correlation and instance based feature selection for electricity load forecasting. *Knowl.-Based Syst.* **2015**, *82*, 29–40. [CrossRef]
32. Hu, Z.Y.; Bao, Y.K.; Xiong, T.; Chiong, R. Hybrid filter–wrapper feature selection for short-term load forecasting. *Eng. Appl. Artif. Intell.* **2015**, *40*, 17–27. [CrossRef]
33. Abedinia, O.; Amjady, N.; Zareipour, H. A New Feature Selection Technique for Load and Price Forecast of Electrical Power Systems. *IEEE Trans. Power Syst.* **2016**, *32*, 62–74. [CrossRef]
34. Raza, M.Q.; Khosravi, A. A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. *Renew. Sustain. Energy Rev.* **2015**, *50*, 1352–1372. [CrossRef]
35. Li, S.; Goel, L.; Wang, P. An ensemble approach for short-term load forecasting by extreme learning machine. *Appl. Energy* **2016**, *170*, 22–29. [CrossRef]
36. Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Mach. Learn. J.* **2003**, *53*, 23–69.
37. Breiman, L.I.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. Classification and Regression Trees (CART). *Biometrics* **1984**, *40*, 17–23.
38. Breiman, L. Random Forest. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
39. He, Y.Y.; Liu, R.; Li, H.Y.; Wang, S.; Lu, X.F. Short-term power load probability density forecasting method using kernel-based support vector quantile regression and Copula theory. *Appl. Energy* **2107**, *185*, 254–266. [CrossRef]
40. Yu, F.; Xu, X.Z. A short-term load forecasting model of natural gas based on optimized genetic algorithm and improved BP neural network. *Appl. Energy* **2014**, *134*, 102–113. [CrossRef]
41. Seeger, M. Gaussian processes for machine learning. *J. Neural Syst.* **2011**, *14*, 69–106. [CrossRef] [PubMed]
42. ISO New England Load Data. Available online: https://www.iso-ne.com/isoexpress/web/reports/pricing/-/tree/zone-info (accessed on 11 November 2014).
43. Singapore Load Data. Available online: https://www.emcsg.com/PriceInformation#download (accessed on 19 December 2016).
44. Sheela, K.G.; Deepa, S.N. Review on Methods to Fix Number of Hidden Neurons in Neural Networks. *Math. Prob. Eng.* **2013**, *6*, 389–405. [CrossRef]