Energy Reports 7 (2021) 1246–1255

Contents lists available at ScienceDirect

Energy Reports

journal homepage: www.elsevier.com/locate/egyr

Prediction of home energy consumption based on gradient boosting regression tree

Peng Nie [a], Michele Roccotelli [b,*], Maria Pia Fanti [b,1], Zhengfeng Ming [a], Zhiwu Li [a,c,1]

[a] School of Electro-Mechanical Engineering, Xidian University, Xi'an 710071, PR China
[b] Polytechnic University of Bari, Bari 70125, Italy
[c] Institute of Systems Engineering, Macau University of Science and Technology, Taipa 999078, Macao Special Administrative Region of China

ARTICLE INFO

Article history:
Received 18 November 2020
Received in revised form 26 January 2021
Accepted 2 February 2021
Available online 20 February 2021

Keywords:
Energy management
Energy consumption
Gradient boosting regression tree
Data prediction

ABSTRACT

Energy consumption prediction of buildings has drawn attention in the related literature since it is very complex and affected by various factors. Hence, a challenging work is accurately estimating the energy consumption of buildings and improving its efficiency. Therefore, effective energy management and energy consumption forecasting are now becoming very important in advocating energy conservation. Many researchers work on saving energy and increasing the utilization rate of energy. Prior works about the energy consumption prediction combine software and hardware to provide reasonable suggestions for users based on the analyzed results. In this paper, an innovative energy consumption prediction model is established to simulate and predict the electrical energy consumption of buildings. In the proposed model, the energy consumption data is more accurately predicted by using the gradient boosting regression tree algorithm. By comparing the performance index Root Mean Square Error of different prediction models through experiments it is shown that the proposed model obtains lower values on different testing data. More detailed comparison with other existing models through experiments show that the proposed prediction model is superior to other models in energy consumption prediction.

1. Introduction

With the development of society and industry, the consumption of energy is increasing. In 2008, Yang et al. (2008) reported that the offices consumed about 70–300 kWh/m^2 every year in China. It is estimated that buildings in Europe consume 40% of the total energy each year (Rosa et al., 2014). In Hong Kong, the energy consumption of buildings represents around over 60% of the total energy consumption (Hong Kong energy end-use data 2012, 2012). The above data shows that the energy consumed by the building is very huge, and the large amount of energy consumption will also affect the surrounding environment. Therefore, using different technologies to reduce energy consumption in buildings has become one of the topics of many researchers. Effective prediction of building energy consumption demand is one of the ways to avoid energy waste. In fact, it is difficult to predict the energy consumption of the building due to the different materials and complicated structure of the buildings. To overcome the influencing factors of traditional energy modeling, prior researchers tried to translate these problems by using sensor-based machine learning method to statistically model energy consumption. Accurate prediction models can improve buildings energy performance and optimize buildings' heating, ventilation and air conditioning (HVAC) systems (Kusiak and Xu, 2012). In the past, people used energy management system to analyze the information of energy consumption (Fanti et al., 2014, 2015). In this process, many sensors will be used to collect data and build management system and the management system uses its own network system to analyze the collected data and provide users with scientific suggestions based on the analysis results (Eder and Nemov, 2017).

Subsequently, in order to predict energy consumption more accurately, some other predictive models of energy consumption have been used by other researchers. Support Vector Machine (SVM) is one of the algorithms in data mining. SVM is also categorized as a new machine learning algorithm for forecasting (Dong et al., 2005). It is used in research and industry due to its highly effective model in solving non-linear problems. Besides that, since it can be used to solve nonlinear regression estimation problems, SVM can be used to forecast time series. SVM so far has been widely used in various analyses such as regression, classification and nonlinear function approximation. The higher accuracy

* Corresponding author.
E-mail addresses: niepeng2847322793@163.com (P. Nie), michele.roccotelli@poliba.it (M. Roccotelli), mariapia.fanti@poliba.it (M.P. Fanti), mingzf@xidian.edu.cn (Z. Ming), zhwli@xidian.edu.cn (Z. Li).
[1] Fellow IEEE.

https://doi.org/10.1016/j.egyr.2021.02.006

predictive results will be obtained with the advantages of the SVM algorithm. Hou and Lian (2009), in 2009, applied SVM to prediction cooling load for HVAC system. In addition, Paudel et al. (2017) use the SVM model to predict the energy consumption of the low energy buildings (LEBs). The researchers apply the SVM model to obtain better prediction results by selecting the relevant data and all the data of the energy consumption. Therefore, the SVM model is still widely used in energy consumption research and its predictive performance is still quite good. Edwards et al. (2012) use a Least Squares Support Vector Machine approach to predict energy consumption. This prediction model had the best performance compared with other machine learning models.

Recurrent neural network (RNN) is another widely applied to predict energy consumption of buildings. Rahman et al. (2018) use RNN to predict energy consumption in commercial and residential buildings. The predictive model had lower error when compared with the conventional multi-layered perceptron neural network. Ugurlu et al. (2018) propose a method to estimate and analysis the electrical prices using RNN. Zagrebina et al. (2019) use RNN model to forecast the Russian energy market and had a good performance. Subsequently, more research results are presented in (Fan et al., 2019; Kim and Cho, 2019; Kong et al., 2019). Hybrid models are that combining advantages of different predictive models, and sometimes it can obtain better prediction results than a single predictive model. Ullah et al. (2020) propose a hybrid model to forecast energy consumption. The results showed that the hybrid model that combines a convolutional neural network with a multi-layer bi-directional long–short term memory had improved comparing with other predictive models. The autoregressive integrated moving average (ARIMA) is established on the basis of a stationary time series and it is often used in time series forecasting (Zhang, 2003). Some researchers try to use some machine learning models or neural network models combined with ARIMA models to build new hybrid models for time series forecasting. Kumar et al. (2018) propose two hybrid models ARIMA-SVR and ARIMA-RNN to predict wind and the hybrid models had better performance.

In recent years, the application of ensemble learning models in prediction has attracted the attention of researchers and obtains great success. Gradient Boosting (GB) (Friedman, 2001) is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models. Touzani et al. (2018) applied gradient boosting model to energy consumption forecasting and achieved good results. More research results are presented in (Zhang and Haghani, 2015; Ayaru et al., 2015; Chen et al., 2013) by using the gradient boosting model. Gradient boosting decision trees (GBDT) models is another prediction model composed of gradient boosting model and decision trees that used in different fields. It also can be named gradient boosting regression trees (GBRT) when using the models for regression prediction. From some current literature (Xie and Coggeshall, 2010; Wang et al., 2016; Ma et al., 2017; Friedman and Meulman, 2003; Ding et al., 2016a), we can find that this model is very helpful for improving the prediction performance.

In this study, we propose a hybrid ARIMA-GBRT model and a GBRT model to predict energy consumption. The GBRT builds the model in a stage-wise fashion and updates it by minimizing the loss function. It may reduce training errors and improve the accuracy by fitting the trees and residuals. The ARIMA-GBRT is a hybrid model that is a combination of ARIMA (Ediger and Akar, 2007; Calheiros et al., 2015; Lee and Ko, 2011) and GBRT model. Both models here are presented and applied for the first time to the prediction of energy consumption.

This paper is organized as follows. In Section 2, we introduce the building energy consumption simulation system that we build and use for accurate data generation. A detailed description of the GBRT model and hybrid model is provided in Section 3. Some experiments are done to compare the proposed models with existing predictive models in Section 4. In addition, conclusions and future perspectives are drawn in Section 5.

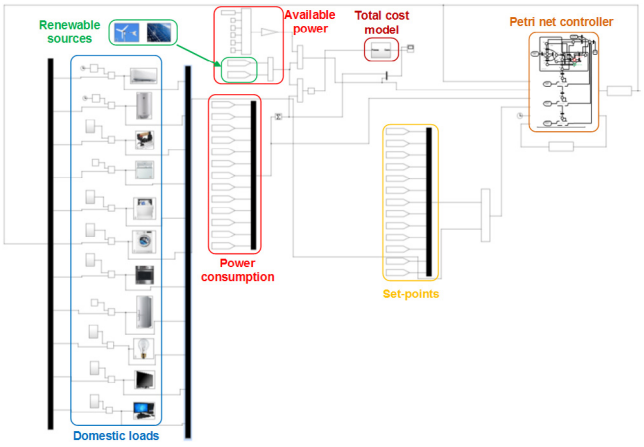

Fig. 1. The energy simulation and control architecture (Simulink).

2. Energy consumption modeling and simulation

In this section, we recall the models to simulate home and building electrical energy consumption presented in Fanti et al. (2018). Such a simulation model is used to generate data set and energy consumption in order to set up the use case for showing the effectiveness of the methodologies developed in the following Section 3. The energy models are implemented by Matlab/Simulink software to simulate and control the electrical energy consumption of the most common domestic appliances such as, HVAC system, water heater, washing machine, dishwashers, computer, TV, refrigerator, oven and lights. Moreover, wind and solar renewable sources are also modeled and integrated in the system.

In Fig. 1, the high-level architecture of the energy simulation and control models is depicted. Here, different models can be distinguished: the domestic loads models; the power consumption model; the available power model that integrates the power from the grid, the wind and solar sources; the energy cost model; the set-point model and the control system model.

Each appliance is modeled based on the technical datasheets to accurately simulate the power consumption. For instance, the HVAC system is modeled to simulate both the heating and cooling functioning modes.

By selecting the HVAC system, it is evident that it includes the heating and cooling models as well as the building thermodynamics and energy cost models (see Fig. 2). Furthermore, the washing machine and dishwasher are modeled to simulate different working programs with different time duration and consumption. For more details on the appliance models and on the other components of the architecture refer to Fanti et al. (2018).

In this paper, the architecture of Fig. 1 is recalled in order to generate datasets of energy consumption. In the proposed architecture, there is a GUI Panel where it is possible to set the available power and to schedule the functioning of each appliance during 24 h by setting all the parameters and the on–off intervals, as it is shown in Fig. 3. After the settings are done, the 24 h simulation can be run and energy data can be plotted and recorded in Matlab. In particular, the aggregated energy consumption by

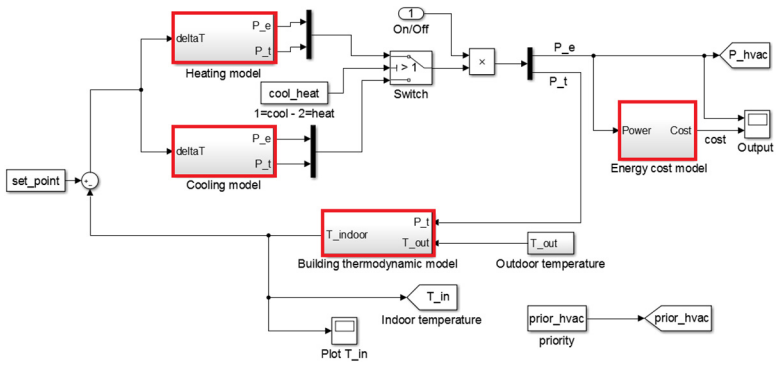

Fig. 2. The Simulink model of the HVAC system.

the appliances is needed. Moreover, several daily scenarios can be reproduced in order to generate the requested datasets to be used for energy profile prediction.

3. Methodology

In this section, the proposed methodology to efficiently predict electrical energy consumption is presented. The proposed gradient boosting regression tree (GBRT) is a modification of the gradient boosting method by using a regression tree of fixed size as the weak learners. The modified version improves the quality of the model. It is an additive regression model consisting of an ensemble of regression trees. In this section, we mainly introduce the GBRT model and ARIMA-GBRT model that we use to forecast energy consumption. The GBRT algorithm is an iterative regression tree algorithm composed of multiple regression trees. The conclusions of all trees are accumulated as the final output. In the following, we first recall the gradient boosting algorithm and then provide the definition of the GBRT and ARIMA-GBRT algorithms.

3.1. Gradient boosting algorithm

Boosting methods combine weak learners by iteratively focusing in the errors resulting at each step until a suitable strong learner is obtained as a sum of the successive weak ones.

Let us consider a response variable y and a set of random input variables $\boldsymbol{x} = \{x_1, x_2, \ldots, x_3\}$. Using a training data in the form of $\{(\boldsymbol{x}_i, y_i)\}$ for $i = 1, 2, \ldots, N$ with $\boldsymbol{x}_i \in R^n$ and $y_i \in R$, the goal is finding an approximation $\widetilde{F}(x)$ of the function $F(x)$ mapping $\boldsymbol{x}$ to y, to minimize loss function $L(y, F(x))$. Errors are inevitable when we expect to seek function $\widetilde{F}(x)$. In the process, the gradient boosting algorithm fits weak learners to loss function and each weak learner model aims to correct errors made by previous weak learner models. This can strengthen the prediction performance and reduce the prediction error of the model.

$$\widetilde{F}(x) = \arg\min_{F(x)} L_{y,x}(y, F(x)) \tag{1}$$

The squared error function is applied as the loss function to estimate the approximation function as $L(y, F(x)) = (y - F(x))^2$. The gradient boosting algorithm starts by setting an initial base learner $F_0(x)$ that usually is a constant function (step 1), and then applies a steepest descent step for the minimization of the loss function. The steepest descent takes steps proportional to the negative gradient of the loss function in order to find the local minimum.

In particular, the gradient of loss function $L(y, F(x))$ can be calculated by using the following equation (step 3):

$$\widetilde{y}_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, \quad i = 1, \ldots, N. \tag{2}$$

It can generalize the calculation range of the gradient when we use regression trees $h(x_i; \boldsymbol{a})$ with parameter $\boldsymbol{a}$ as weak learners. It is usually a parameterized function of the input variables $\boldsymbol{x}$, characterized by parameters $\boldsymbol{a}$ (Friedman, 2001). The tree can be obtained by solving the following equation (step 4):

$$\boldsymbol{a}_m = \arg\min_{\boldsymbol{a},\beta} \sum_{i=1}^{N} [\widetilde{y}_i - \beta h(x_i; \boldsymbol{a})]^2 \tag{3}$$

where $\boldsymbol{a}_m$ is the parameters obtained at iteration m, β is the weight value, also called expansion coefficient, of each weak learner. Each regression tree is fitted to the current negative gradient. Subsequently, the optimal length ρ_m is determined at step 5 and the model $F_m(x)$ is updated at step 6, at each iteration m, with $m = 1, \ldots, M$. The Gradient boosting algorithm is formalized by Algorithm 1 proposed in Friedman (2001).

Algorithm 1: **Gradient boosting**

1. $F_0(x) = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i, \rho)$
2. For m =1 to M do;
3. $\tilde{y}_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, \quad i = 1, \ldots, N$
4. $\boldsymbol{a}_m = \arg\min_{\boldsymbol{a},\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(x_i; \boldsymbol{a})]^2$
5. $\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i - F_{m-1}(x_i) + \rho h(x_i; \boldsymbol{a}_m))$
6. $F_m(x) = F_{m-1}(x) + \rho_m h(x; \boldsymbol{a}_m)$
7. End for

End algorithm

3.2. Gradient boosting regression tree algorithm

Classification and regression trees (CARTs) are proposed by Breiman et al. in 1984 (Breiman et al., 1984). CARTs can be used for both classification and regression models (Prasad et al., 2006; Ding et al., 2016b; Li et al., 2010). The trees used in these two models are called decision trees and decision trees generation is the use of recursive methods to generate binary trees. Since we are studying energy consumption forecasting, we mainly review

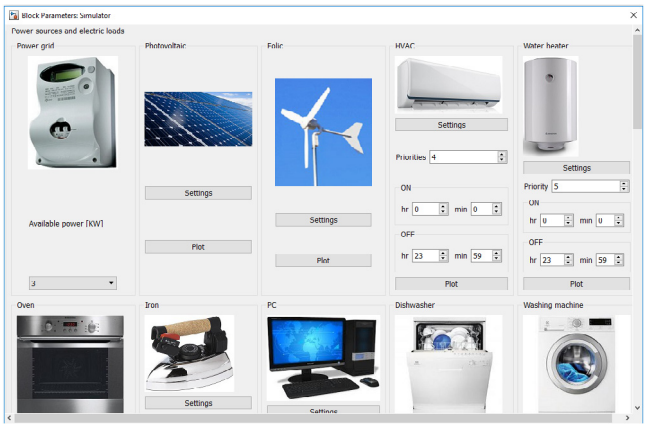

Fig. 3. The GUI panel of the Simulink model.

the algorithm that uses the square error minimization criterion to generate regression trees.

The GBRT algorithm that is the combination of the CART algorithm and the GB algorithm proposed by He et al. (2013) is recalled. It is remarked that, the CART has better performance in prediction compared with most artificial intelligence model (He et al., 2013) because it can model nonlinear relationships without requiring prior information about the probability distribution of variables. As we previously stated, the gradient boosting algorithm integrates weak learners into strong learners. In this study, we use regression trees as weak learners that are generated by CART algorithm. The weak learners are added to the model to correct the prediction errors made by previous models in order to further reduce the prediction error and improve the accuracy of the model.

The GBRT Algorithm 2 is formalized as follows.

Algorithm 2: GBRT

1. $F_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c)$
2. For m =1 to M do;
3. $r_{m,i} = -\left[\dfrac{\partial L(y_i, F(x_i))}{\partial F(x)}\right]_{F(x)=F_{m-1}(x)}, \quad i = 1,\ldots,N$
4. $c_{m,j} = \arg\min_c \sum_{x_i \in R_{m,j}} L(y_i, F_{m-1}(x) + c)$
5. $F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} c_{m,j} I(x \in R_{m,j})$
6. $F_M(x) = F_0(x) + \sum_{m=1}^{M} \sum_{j=1}^{J_m} c_{m,j} I(x \in R_{m,j})$
7. End for

End algorithm

The GBRT algorithm starts by setting the initial value of $F_0(x)$ according to the following equation (step 1):

$$F_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c) \tag{4}$$

where $L(.)$ is a loss function. To calculate the value of the negative gradient of the loss function in the current model as the residual approximation at iteration m, with $m = 1, \ldots, M$, the following equation is introduced (step 3):

$$r_{m,i} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x)}\right]_{F(x)=F_{m-1}(x)}, \quad i = 1, \ldots, N. \tag{5}$$

The number of splits is assumed to be J_m for each regression tree and, therefore, each tree partitions the input space into J_m disjoint regions $R_{m,1}, \ldots, R_{m,Jm}$ and predicts a value $c_{m,j}$ for region $R_{m,j}$. The value of $c_{m,j}$ can be obtained by minimizing the following equation (step 4):

$$c_{m,j} = \arg\min_c \sum_{x_i \in R_{m,j}} L(y_i, F_{m-1}(x) + c). \tag{6}$$

The mth regression tree $F_m(x)$, i.e, the updated model, whose corresponding leaf node area is $R_{m,j}$, $j = 1, 2, \ldots, J_m$, can be obtained as follows (step 5):

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} c_{m,j} I(x \in R_{m,j}). \tag{7}$$

Where $I = 1$ if $x \in R_{m,j}$ and $I = 0$ otherwise (Bevilacqua et al., 2003). Moreover, J_m represents the number of leaf nodes of the mth regression tree. Finally, the model is updated at step 6.

3.3. The hybrid model ARIMA-GBRT

In this section, the proposed hybrid ARIMA-GBRT algorithm model is introduced, which is mainly compared with the hybrid ARIMA-RNN algorithm model presented in Madan and Mangipudi (2018) to verify the performance. Auto regressive (AR) is a model that describes the relationship between the current value and the historical value proposed by Yule (1926), and it can be represented by Eq. (8). Moving average (MA) model focuses on the accumulation of error terms in the autoregressive model which proposed by Slutzky (1937) and it can effectively eliminate random fluctuations in prediction and express in Eq. (9). Subsequently Box and Jenkins combined the AR model and MA model and introduced integrated method to propose ARIMA in Box and Jenkins (1976), where the letter 'I', between AR and MA, stood for the 'Integrated' and reflected the need for differencing to make the series stationary. The model is described by the following equations:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \omega_i \tag{8}$$

$$y_t = \mu + \omega_t + \sum_{j=1}^{q} \theta_j \omega_{t-j} \tag{9}$$

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \omega_t + \sum_{j=1}^{q} \theta_j \omega_{t-j} \tag{10}$$

The meaning of the variables is specified as follows:
y_t current value
ω_t random error term
μ constant
γ_i auto regressive parameters for $i = 1, 2, \ldots, p$
θ_j moving average parameters for $j = 1, 2, \ldots, q$
p order for the differenced series
q order for the white noise series.

The ARIMA model is a linear regression, which uses its own historical data to perform regression. It is suitable for the internal and stable correlation between the data itself. It is widely used in time series forecasting problems and has achieved good forecast results. In the data processing, some data may have a certain linear relationship, and some may have a nonlinear relationship.

It may obtain better results to analyze the data by using a hybrid algorithm. Hybrid algorithms mainly combine the advantages of different algorithms to process corresponding data.

When performing hybrid model to predict the task, first it is necessary to divide the data into two parts: linear data and nonlinear data. Using the ARIMA model to predict linear data and for the nonlinear data can be predicted by GBRT model. Moreover, the prediction results of the two models are extracted as the prediction result. We mainly use the following three indicators, namely Root Mean Square Error (*RMSE*), Mean Absolute Error (*MAE*) and Mean Absolute Percentage Error (*MAPE*), to compare the performance of the models:

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

$$MAE = \frac{\sum_{i=1}^{n}|Y_i - \hat{Y}_i|}{n} \tag{12}$$

$$MAPE = \frac{1}{n}\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{Y_i} \tag{13}$$

where Y_i is the actual measurement, $\hat{Y}_i$ is the predicted value; n is the number of measurements.

4. Experimental study

In this section the proposed prediction model is simulated and validated for electrical energy consumption. From the literature analysis it is known that some researchers use the SVM algorithm model to predict electrical energy consumption and have achieved good prediction results. In other contributions, some researchers use a hybrid algorithm model to predict electrical energy consumption. As we introduce previous, the researchers mainly want to combine the advantages of different algorithms to achieve good prediction results. To this purpose, we first compare the prediction performance between the GBRT model used with the SVM model that are presented in Dong et al. (2005). Then, we compare the prediction performance of the ARIMA-GBRT hybrid model and the ARIMA-RNN hybrid model. The data set of energy consumption is generated by simulating the use of HVAC, water heater, iron, electric oven, PC, dishwasher, washing machine, hair dryer, TV, dimmable and fluorescent lamps. The technical parameters of the appliances are described in Appendix. In addition, 20 daily operating schedules are applied to each day of the week in order to get the average aggregate consumption per day.

4.1. Validation for energy consumption prediction

Before conducting the experiment, we normalize the experimental data. First, electrical energy consumption data are selected for one day every 15 s for experimental verification. Fig. 4 shows the results of the electrical energy consumption predicted by the SVM algorithm model and Fig. 5 shows the electrical energy consumption results predicted by the GBRT algorithm model. We can intuitively find that the model prediction effect of the GBRT algorithm is better than the model prediction performance of the SVM algorithm. Through the specific prediction numerical calculation, the performance index *RMSE* predicted by the GBRT algorithm model is 45.58% lower than that of the SVM algorithm model.

Then, the prediction performance of the hybrid algorithm model ARIMA-GBRT is compared with ARIMA-RNN. Figs. 6 and 7 are the electrical energy consumption prediction results of the ARIMA-GBRT algorithm model and the ARIMA-RNN algorithm model, respectively. We clearly see that the ARIMA-GBRT algorithm model has better electrical energy consumption prediction performance than the ARIMA-RNN algorithm model. The performance index *RMSE* predicted by the ARIMA-GBRT algorithm model is 12.22% lower than that of the ARIMA-RNN algorithm model.

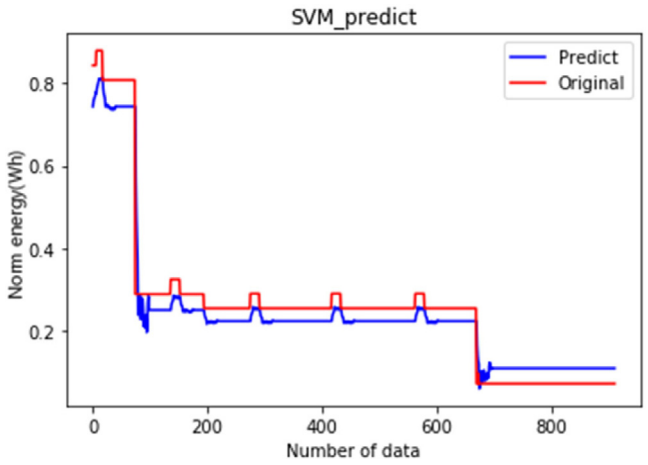

Fig. 4. The electrical energy consumption prediction with the SVM algorithm.

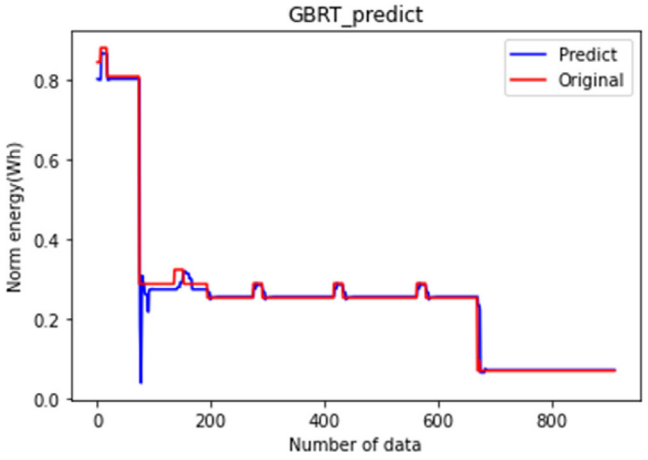

Fig. 5. The electrical energy consumption prediction with the GBRT algorithm.

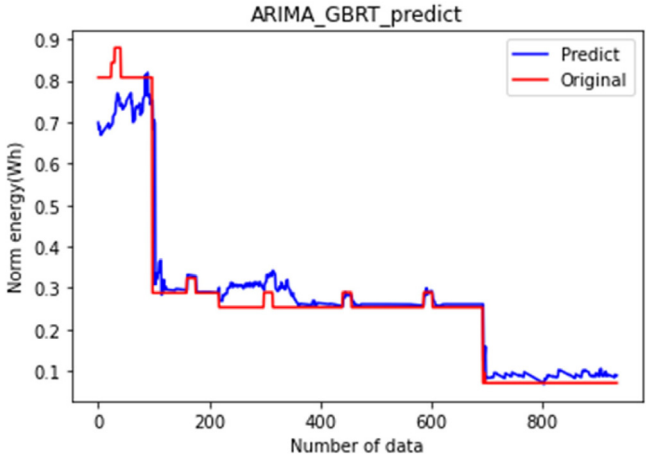

Fig. 6. The electrical energy consumption prediction with the ARIMA-GBRT algorithm.

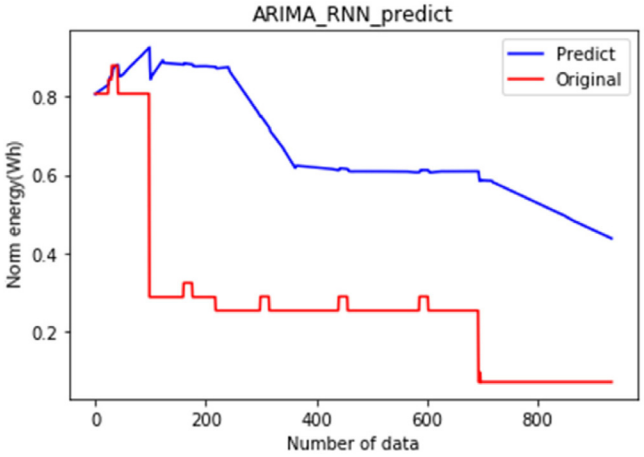

Fig. 7. The electrical energy consumption prediction with the ARIMA-RNN algorithm.

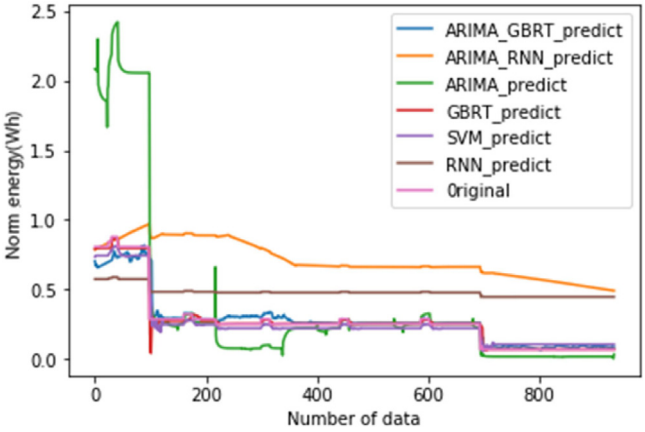

Fig. 8. Electrical energy consumption prediction results of different algorithm models.

4.2. Comparison with prediction models

In order to further illustrate the performance of the proposed algorithm, we perform a comparison with other prediction models. We use the electrical energy consumption data for the first 20 h as the training set, and the data for the next 4 h as the test set to verify the prediction performance of different algorithm models. Fig. 8 shows the electrical energy consumption prediction results of different algorithm models. To use more data to compare the performance of different algorithm models, then use the same experimental method for the electrical energy consumption data from the first day to the seventh day. We mainly selected three performance indexes of *RMSE*, *MAE* and *MAPE* to compare the performance of different algorithm models. Fig. 9 shows the comparison results of three performance indicators *RMSE*, *MAE* and *MAPE* on the training dataset by different algorithm models. Fig. 10 shows the comparison results of three performance indicators *RMSE*, *MAE* and *MAPE* on the testing set by different algorithm models. From Fig. 9 and Fig. 10, we can find that the three performance indicators achieved by the GBRT algorithm model are the best. Table 1 shows that the percentage of the *RMSE* values obtained by the GBRT prediction model are lower than the corresponding values obtained by other prediction models in the range of 1.45% to 94.60% on the training data and in the range of 1.42% to 96.30% on the testing data, respectively. In addition, Table 2 reports the *MAE* values showing that the performance of the GBRT prediction model is worse than the hybrid model ARIMA-GBRT model on few training data and testing data. The hybrid model ARIMA-GBRT has better performance than ARIMA-RNN. On some data sets, the hybrid model ARIMA-GBRT has a slightly worse performance than the SVM model.

Through a set of experimental verifications, we find that the GBRT algorithm model and the ARIMA-GBRT model are better than the other commonly used algorithm models in the electrical energy consumption prediction. Next, we study to predict the electrical energy consumption data of the seventh day based on the electrical energy consumption data of the previous six days. We also use the data from the first six days as the training set and the data on the seventh day as the test set. Fig. 11 shows the performance index values of *RMSE*, *MAE* and *MAPE* obtained by using different algorithm models on the training data and testing data. The experimental results show that the performance index parameters *RMSE* obtained by the GBRT algorithm model lower is smaller than those of other prediction models. The RNN algorithm has lower value of *MAE* than that of GBRT algorithm. Experimental results also show that the prediction performance of the ARIMA-GBRT hybrid model is better compared with the ARIMA-RNN hybrid model. Finally, Table 3 shows that the percentage of the *RMSE* and *MAE* values obtained by the GBRT model are lower

Table 1
The percentages (%) of *RMSE* obtained by the GBRT prediction model are lower than ones obtained by other prediction models on the training data and testing data.

	RNN	SVM	ARIMA	ARIMA-GBRT	ARIMA-RNN
		Training data			
day1	91.76	39.76	84.51	28.47	88.28
day2	23.29	33.73	83.16	25.33	81.76
day3	93.05	33.43	83.08	24.75	94.09
day4	59.26	45.58	39.80	1.01	11.25
day5	94.36	41.06	75.17	50.00	93.98
day6	94.60	42.61	58.08	49.57	94.14
day7	79.48	6.77	1.45	1.56	68.81
		Testing Data			
day1	91.67	37.57	85.42	31.27	90.56
day2	5.74	18.42	80.72	65.43	79.78
day3	91.33	16.96	87.36	78.27	92.97
day4	57.76	44.66	96.30	67.53	71.28
day5	94.12	38.91	93.55	83.26	94.61
day6	94.49	44.36	64.49	69.40	93.95
day7	79.73	7.01	1.75	1.42	84.71

Table 2
The percentages (%) of *MAE* obtained by the GBRT prediction model are lower than ones obtained by prediction models on the training data and testing data.

	RNN	SVM	ARIMA	ARIMA-GBRT	ARIMA-RNN
		Training data			
day1	88.24	10.58	5.76	−32.44	83.42
day2	6.01	10.88	38.12	4.97	81.02
day3	93.85	3.68	34.17	−1.10	95.06
day4	80.39	69.52	−63.75	25.23	36.25
day5	98.18	79.21	76.49	71.58	98.03
day6	97.97	74.11	82.30	64.76	97.86
day7	83.75	35.34	20.39	7.69	72.62
		Testing Data			
day1	88.74	−26.07	45.78	62.26	82.92
day2	9.73	12.11	82.85	75.48	83.88
day3	93.51	−1.53	81.75	71.22	94.27
day4	78.68	67.59	81.62	69.88	75.46
day5	98.20	79.53	95.60	91.24	98.18
day6	97.79	72.76	80.53	76.05	97.46
day7	83.28	32.48	30.26	−38.87	68.79

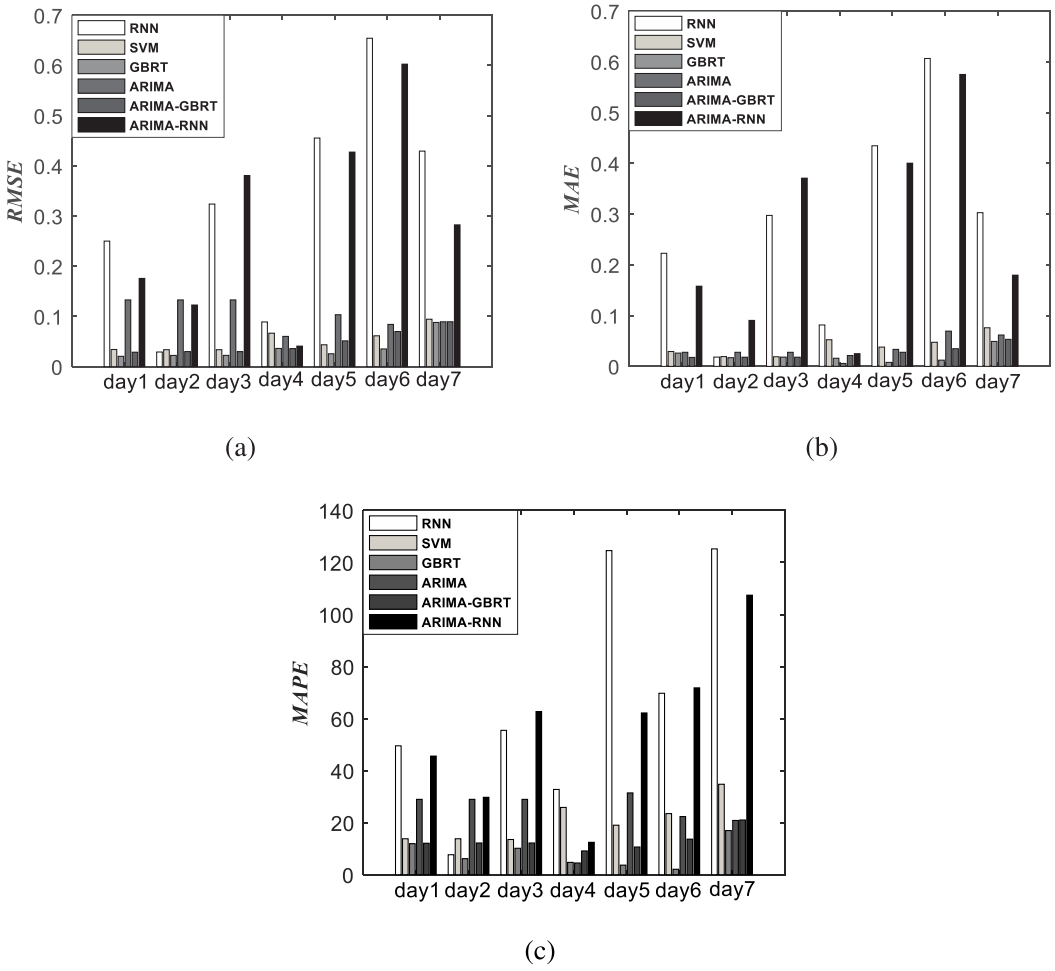

Fig. 9. Performance indexes obtained by different algorithm models on the training set, (a) *RMSE*, (b) *MAE*, (c) *MAPE*.

Table 3
The percentage (%) of the *RMSE* and *MAE* values of the GBRT prediction model are lower than ones of other prediction models on the training data and testing data.

	Training data		Testing data	
	RMSE	*MAE*	*RMSE*	*MAE*
RNN	1.51	−27.33	1.48	−29.47
SVM	71.18	89.27	70.54	89.49
ARIMA	92.72	86.57	92.90	87.40
ARIMA-GBRT	35.56	50.75	91.41	96.18
ARIMA-RNN	40.58	69.19	92.12	96.51

than ones obtained by other prediction models, after 20 repeated experiments.

5. Conclusions

This paper simulates and predicts the electrical energy consumption of buildings and analyzes the electrical energy consumption data by using new algorithm models. Moreover, the proposed hybrid ARIMA-GBRT model and GBRT model are compared with other prediction models presented in the related literature. The experimental results show that the prediction model we use has a better performance than others. Indeed, the analysis of the results shows the lower values of the indices *RMSE* and *MAE*, by indicating that the forecasting performance of the proposed models is more accurate. Moreover, the presented prediction model has higher accuracy and computational speed.

The proposed models to forecast the electrical energy consumption is useful for designing the building HVAC systems by accurately estimating the electric energy consumption and suitably allocating energy in an optimal way. In addition, the predicting models can be applied to optimize the cost of electrical energy consumption in the buildings.

Further research study will deal with a long-term forecasting strategy by using the GBRT algorithm.

CRediT authorship contribution statement

Peng Nie: Conception and design of study, Analysis and/or interpretation of data, Drafting the manuscript. **Michele Roccotelli:** Conception and design of study, Acquisition of data, Drafting the manuscript. **Maria Pia Fanti:** Conception and design of study, Revising the manuscript critically for important intellectual content. **Zhengfeng Ming:** Conception and design of study, Revising the manuscript critically for important intellectual content. **Zhiwu Li:** Conception and design of study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

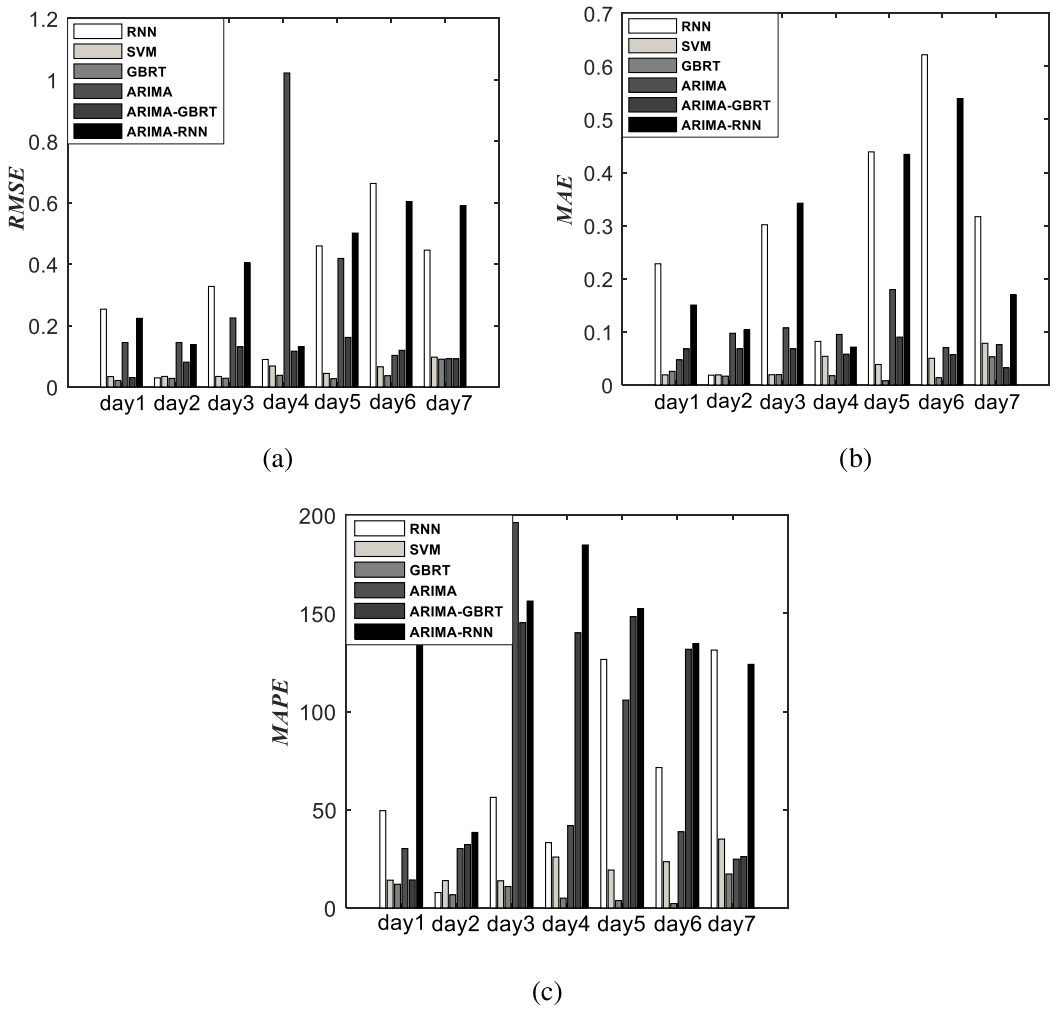

Fig. 10. Performance indexes obtained by different algorithm models on the testing set:(a) *RMSE*, (b) *MAE*, (c) *MAPE*.

Acknowledgment

All authors approved the version of the manuscript to be published. This work has been financed by the Italian Project AMSARA under the program Smart Cities and Communities and Social Innovation, D.D. 5 luglio 2012 n. 391/Ric.

Appendix

The parameters of the building appliances set up for the experimentations of Section 4 are reported in the following list

1. HVAC
   - Cooling power: 7000W
   - Heating power: 8000W
   - Energy Efficiency Ratio: 3.6
   - Coefficient of performance: 3.2
   - Set-point temperature: 22 °C
   - Outdoor mean temperature: 10 °C
2. Water heater
   - Electric power: 1500W
   - Set-point temperature: 50 °C
   - Cold water temperature: 20 °C
   - Thermostat tolerance: 2 °C
3. Iron
   - Electric power: 1300W
   - Set-point temperature: 140 °C
   - Thermostat tolerance: 10 °C
4. Electric oven
   - Electric power: 2000W
   - Set-point temperature: 180 °C
   - Thermostat tolerance: 10 °C
5. PC
   - Electric power: 100W
6. Dishwasher
   - Electric power: 1950W
   - Working programs: eco, light, classic, intense (max consumption)
7. Washing machine
   - Electric power: 1950W
   - Working programs: white, eco, synthetics, delicates, wool, centrifuge
8. Hair dryer

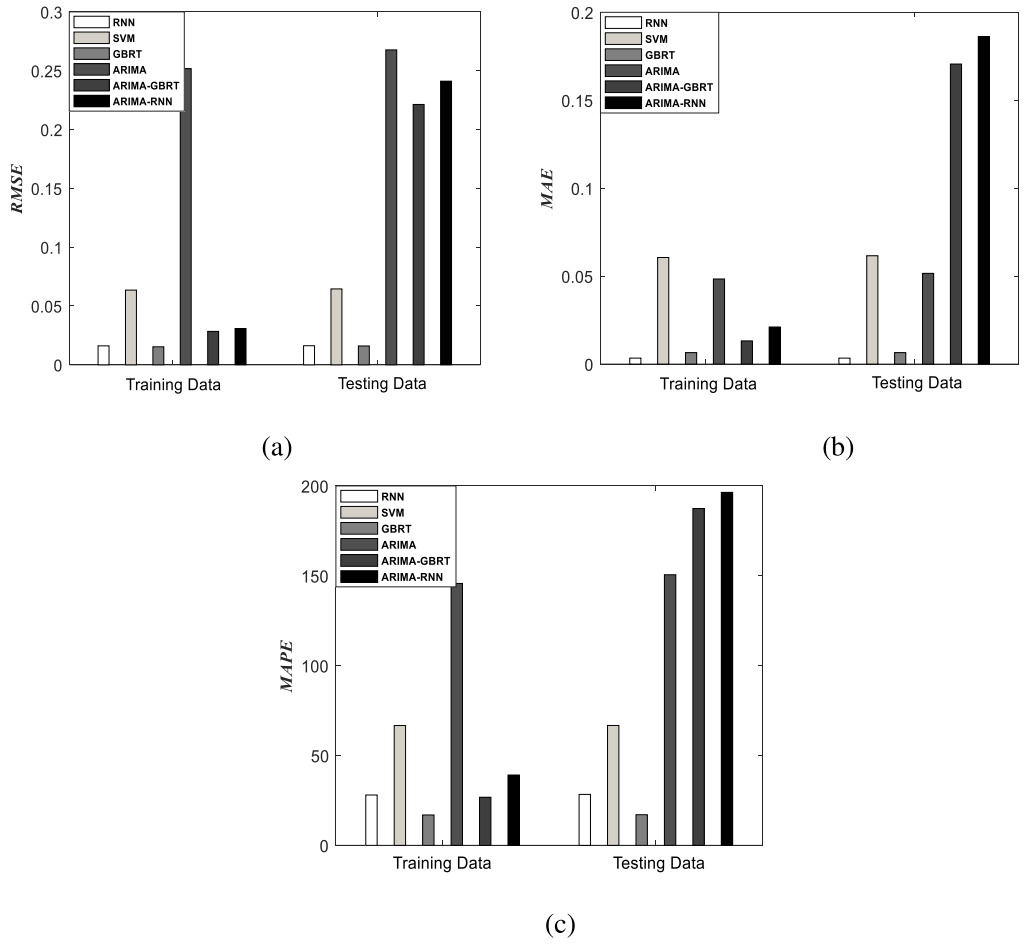

Fig. 11. The performance index obtained on training data and testing data by using different prediction models:(a) *RMSE*, (b) *MAE*, (c) *MAPE*.

- Electric power: 1950W
- Air velocity: 0, 1, 2 (max)
- Hot intensity: 0, 1, 2 (max)

9. Dimmable lamps

   - Source voltage: 0-220V (220V default)

10. Fluorescent lamps

    - Number of lamps: 3
    - Mean electric power: 30.6W
    - Mean electric current: 0.39A

11. TV

    - Electric power: 120W

References

Ayaru, L., Ypsilantis, P.P., Nanapragasam, A., Choi, R.C., Thillanathan, A., Min-Ho, L., Montana, G., 2015. Prediction of outcome in acute lower gastrointestinal bleeding using gradient boosting. PloS One 10 (7).

Bevilacqua, M., Braglia, M., Montanari, R., 2003. The classification and regression tree approach to pump failure rate analysis. Reliab. Eng. Syst. Saf. 79 (1), 59–67.

Box, G.E.P., Jenkins, G., 1976. Time Series Analysis, Forecasting and Control, Holden-Day, San Francisco, CA.

Breiman, L., Friedman, J.H., Olshen, R., Stone, C.J., 1984. Classification and regression trees, Wadworth.

Calheiros, R.N., Masoumi, E., Ranjan, R., Buyya, R., 2015. Workload prediction using ARIMA model and its impact on cloud applications' QoS. IEEE Trans. Cloud Comput. 3 (4), 449–458.

Chen, Y., Jia, Z., Mercola, D., Xie, X., 2013. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. Comput. Math. Methods Med. 2013, http://dx.doi.org/10.1155/2013/873595.

Ding, C., Wang, D., Ma, X., Li, H., 2016a. Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. Sustainability 8 (11).

Ding, C., Wu, X., Yu, G., Wang, Y., 2016b. A gradient boosting logit model to investigate driver's stop-or-run behavior at signalized intersections using high-resolution traffic data. Transp. Res. 72, 225–238.

Dong, B., Cao, C., Lee, S.E., 2005. Applying support vector machines to predict building energy consumption in tropical region. Energy Build. 37 (5), 545–553.

Eder, L.V., Nemov, V.Y., 2017. Forecast of energy consumption of vehicles. Stud. Russian Econ. Develop. 28 (4), 423–430.

Ediger, V.S., Akar, S., 2007. ARIMA forecasting of primary energy demand by fuel in Turkey. Energy Policy 35 (3), 1701–1708.

Edwards, R.E., New, J., Parker, L.E., 2012. Predicting future hourly residential electrical consumption: a machine learning case study. Energy Build. 49, 59–603.

Fan, C., Wang, J., Gang, W., Li, S., 2019. Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. Appl. Energy 236 (15), 700–710.

Fanti, M.P., Mangini, A.M., Roccotelli, M., 2014. A Petri Net model for a building energy management system based on a demand response approach. In: 2014 IEEE 22nd Mediterranean Conference on Control and Automation (MED), pp. 816–821.

Fanti, M.P., Mangini, A.M., Roccotelli, M., 2018. A simulation and control model for building energy management. Control Eng. Pract. 72, 192–205.

Fanti, M.P., Mangini, A.M., Roccotelli, M., Ukovich, W., Pizzuti, S., 2015. A control strategy for district energy management. In: 2015 IEEE International Conference on Automation Science and Engineering (CASE), pp. 432–437.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann. Statist. 29 (5), 1189–1232.

Friedman, J.H., Meulman, J.J., 2003. Multiple additive regression trees with application in epidemiology. Stat. Med. 22 (9), 1365–1381.

He, Q., Kamarianakis, Y., Jintanakul, K., Wynter, L., 2013. Incident duration prediction with hybrid tree-based quantile regression. In: Advances in Dynamic Network Modeling in Complex Transportation Systems. Springer, New York, NY, USA, pp. 287–305.

Hong Kong energy end-use data 2012, 2012. Hong kong electrical & mechanical services department.

Hou, Z., Lian, Z., 2009. An application of support vector machines in cooling load prediction. Int. Syst. Appl. 1–4.

Kim, T., Cho, S., 2019. Predicting residential energy consumption using CNN-LSTM neural networks. Energy 182 (1), 72–81.

Kong, W., Dong, Z., Jia, Y., Hill, D.J., Xu, Y., Zhang, Y., 2019. Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Trans. Smart Grid 10 (1), 841–851.

Kumar, H., Arora, P., Panigrahi, B.K., 2018. Wind forecasting: Hybrid statistical and deep neural network approaches. In: 2018 3rd International Conference on Contemporary Computing and Informatics (IC3I), Gurgaon, India, pp. 62–67.

Kusiak, A., Xu, G., 2012. Modeling and optimization of HVAC systems using a dynamic neural network. Energy 42 (1), 241–250.

Lee, C.M., Ko, C.N., 2011. Short-term load forecasting using lifting scheme and ARIMA models. Expert Syst. Appl. 38 (5), 5902–5911.

Li, H., Sun, J., Wu, J., 2010. Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. Expert Syst. Appl. 37 (8), 5895–5904.

Ma, X., Ding, C., Luan, S., Wang, Y., Wang, Y., 2017. Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method. IEEE Trans. Intell. Transp. Syst. 18 (9), 2303–2310.

Madan, R., Mangipudi, P.S., 2018. Predicting computer network traffic: A time series forecasting approach using DWT, ARIMA and RNN. In: 2018 Eleventh International Conference on Contemporary Computing (IC3), 2018, pp. 1–5.

Paudel, S., Elmitri, M., Couturier, S., Nguyen, P.H., Kamphuis, R., Lacarrière, B., Corre, O.L., 2017. A relevant data selection method for energy consumption prediction of low energy building based on support vector machine. Energy Build. 138 (1), 240–256.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9, 181–199.

Rahman, A., Srikumar, V., Smith, A.D., 2018. Predicting electrical consumption for commercial and residential buildings using deep recurrent neural networks. Appl. Energy 212 (15), 372–385.

Rosa, M.D., Bianco, V., Scarpa, F., Tagliafico, L.A., 2014. Heating and cooling building energy demand evaluation: a simplified model and a modified degree days approach. Appl. Energy 128 (1), 217–229.

Slutzky, E., 1937. The summation of random causes as the source of cyclic processes. Econometrica 5 (2), 105–146.

Touzani, S., Granderson, J., Fernandes, S., 2018. Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy Build. 158 (1), 1533–1543.

Ugurlu, U., Oksuz, I., Tas, O., 2018. Electrical price forecasting using recurrent neural networks. Energy 11 (5), 1–23.

Ullah, F.U.M., Ullah, A., Haq, I.U., Rho, S., Baik, S.W., 2020. Short-term prediction of residential power energy consumption via CNN and multi-layer bi-directional LSTM networks. IEEE Access 8, 123369-123380.

Wang, Y., Feng, D., Li, D., Chen, X., Zhao, Y., Niu, X., 2016. A mobile recommendation system based on logistic regression and gradient boosting decision trees. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 1896–1902.

Xie, J., Coggeshall, S., 2010. Prediction of transfers to tertiary care and hospital mortality: A gradient boosting decision tree approach. Stat. Anal. Data Min. 3 (4), 253–258.

Yang, L., Lam, J.C., Tsang, C.L., 2008. Energy performance of building envelopes in different climate zones in China. Appl. Energy 85 (9), 800–817.

Yule, G.U., 1926. Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series. J. R Stat. Soc. 89 (1), 1–63.

Zagrebina, S.A., Mokhov, V.G., Tsimbol, V.I., 2019. Electrical energy consumption prediction is based on the recurrent neural network. Procedia Comput. Sci. 150, 340–346.

Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50, 159–175.

Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. Transp. Res. 58, 308–324.