buildings

Article

Data-Driven Based Prediction of the Energy Consumption of Residential Buildings in Oshawa

Yaolin Lin [1,*], Jingye Liu [1], Kamiel Gabriel [2,*], Wei Yang [3] and Chun-Qing Li [4]

[1] School of Environment and Architecture, University of Shanghai for Science and Technology, Shanghai 200093, China
[2] Faculty of Engineering and Applied Science, Ontario Tech University, Oshawa, ON L1G 0C5, Canada
[3] Faculty of Architecture, Building and Planning, The University of Melbourne, Melbourne 3010, Australia
[4] School of Engineering, RMIT University, Melbourne 3000, Australia
[*] Correspondence: ylin@usst.edu.cn (Y.L.); kamiel.gabriel@ontariotechu.ca (K.G.)

Abstract: Buildings consume about 40% of the global energy. Building energy consumption is affected by multiple factors, including building physical properties, performance of the mechanical system, and occupants' activities. The prediction of building energy consumption is very complicated in actual practice. Accurate and fast prediction of the building energy consumption is very important in building design optimization and sustainable energy development. This paper evaluates 24 energy consumption models for 83 houses in Oshawa, Canada. The energy consumption, social and demographic information of the occupants, and the physical properties of the houses were collected through smart metering, a phone survey, and an energy audit. A total of 63 variables were determined, and based on the variable importance, three groups with different numbers of variables were selected, i.e., 26, 12, and 6 for electricity consumption; and 26, 13, and 6 for gas consumption. A total of eight data-driven algorithms, namely Multiple Linear Regression (MLR), Stepwise Regression (SR), Support Vector Machine (SVM), Backpropagation Neural Network (BPNN), Radial Basis Function Neural Network (RBFN), Classification and Regression Tree (CART), Chi-Square Automatic Interaction Detector (CHAID), and Exhaustive CHAID (ECHAID), were used to develop energy prediction models. The results show that the BPNN model has the best accuracies in predicting both the annual electricity consumption and gas consumption, with mean absolute percentage errors (MAPEs) of 0.94% and 0.94% for training and validation data for electricity consumption, and 2.63% and 0.16% for gas consumption, respectively.

Keywords: data-driven; electricity consumption; prediction model; gas consumption

check for updates

Citation: Lin, Y.; Liu, J.; Gabriel, K.; Yang, W.; Li, C.-Q. Data-Driven Based Prediction of the Energy Consumption of Residential Buildings in Oshawa. *Buildings* **2022**, *12*, 2039. https://doi.org/10.3390/buildings12112039

Academic Editor: Francesco Asdrubali

Received: 13 October 2022
Accepted: 18 November 2022
Published: 21 November 2022

1. Introduction

Globally, buildings consume about 30% of end energy usage and over 55% of electricity [1]. Building energy consumption is increasing with the growth of the global population. It is affected by a large number of physical and sociological factors. Accurate energy prediction can help quantify and compare the energy-saving potentials of different conservation measures, as well as assist design optimization [2,3].

There are two approaches to predict building energy consumption. One is based on a physical model, and the other is data driven. The physical modeling approach is also called the forward modeling approach. The forward modeling approach is usually conducted with commercial software, e.g., DOE-2, DesignBuilder, etc., with given inputs to estimate the building energy consumption through simulation. The differences of the outcomes among different software are typically small with the same/identical input values of the variables [4]. Fumo et al. [5] used EnergyPlus Benchmark Models to generate the determining factors based on the monthly electrical and fuel utility bills to estimate the hourly electricity consumption and fuel energy consumption for a hypothetical building in Atlanta, GA, and in Meridian, MS, with estimated errors within 10%. Amiri et al. [6]

developed a Stepwise Regression (SR) model, based on the simulation results from DOE-2, to predict the building energy consumption at the early design phase. The physical modeling approach requires detailed information about the building, mechanical systems, and occupants' activities to develop a mathematical model to estimate the building energy consumption, which might not be readily available. Meanwhile, the physical model could not take into account the sociological factors that potentially affect the energy-usage patterns of the occupants.

The data-driven approach uses data analysis through known data sets to overcome the limitations of physical models to predict the energy consumption. Typically, an energy-usage database is created through the simulation of building samples or data collection. Examples of data-driven approaches include Multiple Linear Regression (MLR), Classification and Regression Tree (CART), artificial neural network (ANN), etc.

MLR models have been developed to replace the outcomes from building simulation software. Chen et al. [7] developed a physical-based MLR model to predict the building cooling load based on the data set created through building energy simulation using EnergyPlus. It was demonstrated to have a stronger generalization ability than the BP-ANN and MLR models. By using this method, the space cooling load can be predicted based on the total cooling load. Ciulla et al. [8] used TRNSYS to run 1560 simulations of a non-residential building with different configurations across Italy to create an energy database and developed MLR models to estimate the building energy consumption with determination coefficients (R^2) higher than 0.9 and mean absolute error (MAE) lower than 10 kWh/m^2 year.

Stepwise Regression (SR) can help overcome the multicollinearity problem that could exist in the multiple regression problem and reduce the number of input variables. Tso and Yau [9] developed the SR analysis of the household electricity consumption in winter and summer in Hongkong. Zhao and Lin [10] proposed SR models to predict the energy consumption and visual discomfort of a passive house, compared with the simulated outcomes from DesignBuilder. R-squares of 0.9808 and 0.8487 were found, respectively, which demonstrate the potential of SR in predicting the building energy consumption.

The Support Vector Machine (SVM) helps to solve high-dimensional difficulty and local minima problems. Ma et al. [11] applied support vector regression (SVR) models to estimate the provincial building energy consumption in four provinces in Southern China. Seven parameters, including yearly mean outdoor dry-bulb air temperature, relative humidity, total solar radiation, urbanization ratio, gross domestic product, household consumption level, and total construction area of were used as inputs. Good agreements were found between the predicted and actual energy consumptions, with the mean square errors (MSEs) and correlation coefficients found to be less than 0.001 and greater than 0.99, respectively. Li et al. [12] developed a SVM model to estimate the office hourly cooling load with outdoor air temperature, relative humidity, and solar radiation intensity as the input variables. The SVM model outperforms the Backpropagation Neural Network (BPNN) model in terms of accuracy and generalization. Paudel et al. [13] developed a SVM model for a low-energy residential building in France, using a small representative day data set. The outdoor air temperature, horizontal solar radiation, solar gain transmitted through windows, solar energy absorbed by walls, occupancy profile, and time moving average of outdoor air temperature were used as input variables for the model. It was found that the model achieves higher prediction accuracy (R^2 = 0.98; RMSE = 3.4), compared to the one developed with all the data sets (R^2 = 0.93; RMSE = 7.1).

BPNN is the most widely used neural network. Ahmad et al. [14] developed feed-forward BPNN and random forest (RF) models to estimate the energy demand of the HVAC system in a commercial building in Madrid, Spain. The input variables include outdoor air temperature, dew point temperature, relative humidity, wind speed, duration time, number of guests on the day, and number of rooms booked. The results show that the RMSEs of the prediction results of the BPNN and RF models were 4.97 and 6.10, respectively. The BPNN model achieves a slightly better performance than the RF model in terms of accuracy.

Radial Basis Function Neural Networks (RBFNs) have been used to predict the energy consumption of university buildings. Han et al. [15] proposed an RBFN model to evaluate the energy performance of the buildings, using the University of California Irvine data sets. The predicted values agree well with the simulation outcome from Ecotech. Zhao et al. [16] developed an RBFN model to predict the energy consumption of colleague buildings in Fujian Province in China, with a maximum error of 13.3%.

Classification and Regression Tree (CART) is also one of the machine learning approaches favored by the researchers. Zekić-Sušac et al. [17] developed a CART model to predict the energy cost of public buildings in the Republic of Croatia. Capozzoli et al. [18] developed a CART model to predict the heating energy consumption in schools with an R-square of 0.86.

The Chi-Square Automatic Interaction Detector (CHAID) can be used to generate a multi-branched decision tree and determine the branch variables' values based on statistical significance. Yang and Wu [19] applied CHAID to find the energy-saving strategies for central air-conditioning system operation in Shenzhen, China. Kusiak et al. [20] developed a CHAID model to predict the building steam load with a mean absolute error (MAE) of 405 for training and 578 for testing.

Exhaustive CHAID (ECHAID) is another decision tree algorithm that ensures the same degree of freedom for all the inputs. Kusiak et al. [20] compared the outcomes from ECHAID model with the CHAID model in predicting the building steam load. The ECHAID achieved a mean absolute error (MAE) of 398 for training and 570 for testing. Yan et al. [21] developed an ECHAID model to predict the system coefficient of performance (COP) of a ground-source heat pump with an MAE of 0.098 for training and 0.105 for testing.

Researchers have also investigated other data-driven approaches; for example, Li et al. [22] developed a hybrid teaching–learning artificial neural network model (TL-ANN) to predict the hourly electrical energy consumption for two educational buildings located in USA and China, using weather conditions, calendar date, occupancy pattern, and historical energy usage data. Moayedi [23] compared the performances of three cooling load prediction models for a residential building. The elephant herding optimization (EHO), ant colony optimization (ACO), and Harris Hawks optimization (HHO), were combined with a multilayer perceptron neural network (MLP) model. The relative compactness of the building, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution are used as inputs for the model. The results show that the EHO–MLP has the highest prediction accuracy, followed by HHO–MLP and ACO–MLP. Aruta et al. [24] developed an artificial neural networks (ANNs) model, using NARX (nonlinear autoregressive model with exogenous inputs) networks for training based on simulated heating load of a building in Rome from EnergyPlus. The outdoor air temperature and solar radiation were used as inputs and demonstrated satisfactory prediction performance. Ndiaye and Gabriel [25] used the latent root regression technique to reduce the number of input variables from 59 to 9, while achieving an R-square of 0.79 in predicting the housing unit electricity consumption in Oshawa. Still, they performed studies only on a few data-driven algorithms.

From the literature survey, it can be found that very few studies were conducted to predict the yearly residential building energy consumption based on actual energy consumption data. Many studies focus on monthly [26], daily [27–29], or hourly [13,27–30] energy consumption, based on the simulation outcomes from commercial software [26,31–35]. Short-term energy predictions are easily affected by seasonal variation and the outcomes from the simulation often deviate from actual energy consumption. In addition, the effects of occupants' behaviors on the energy usage are often neglected in the prediction model, and most of the parameters focus on weather data [26–29,31] or design parameters of the building envelope [26,31,33–35], thus causing deviations in energy consumption predictions for different households; social and demographic information are often neglected, as well. Moreover, many of the studies used fixed number of input variables and training/validation ratio, without seeking for the least number of inputs needed and the models with the best performance. Therefore, it is important to develop a residential building energy prediction model based on the collected data from actual annual energy consumption, taking into account the social and demographic information and evaluate the impact of the number of input variables, as well as the training/validation ratio for the performance of the prediction model.

This paper attempts to develop yearly energy consumption prediction models for residential buildings in Oshawa. Data related to electricity consumption, gas consumption, physical information of the buildings, and social and demographic information of the residents were collected through smart metering, a phone survey, and energy auditing of a total of 83 households. A total of eight data-driven algorithms, namely Multiple Linear Regression (MLR), Stepwise Regression (SR), Support Vector Machine (SVM), Backpropagation Neural Network (BPNN), Radial Basis Function Neural Network (RBFN), Classification and Regression Tree (CART), Chi-Square Automatic Interaction Detector (CHAID), and Exhaustive CHAID (ECHAID), were used to develop energy prediction models to select the most suitable models for electricity consumption and gas consumption predictions. Different numbers of input variables and training/validation ratios were employed to find the models with the best prediction performance with the least number of inputs. The outcomes from this paper can provide references for residential-building energy prediction.

2. Method

The actual electricity and gas consumption data, physical properties, mechanical system information, and consumer information of 227 houses in Oshawa—which has a humid continental climate with large seasonal temperature variations, with warm summers and cold winters—were collected and analyzed. The energy consumption is for a full year. Firstly, smart meters were installed on 227 houses in Oshawa to obtain the electricity readings, and a phone survey on the social and demographic information of the occupants, as well as information on the electrical appliances, was conducted on the houses with installed smart meters. Energy audits were conducted according to the willingness of the house owner/renter. A total of 65 input and output parameters were identified after an analysis of the gathered information. During the data preprocessing, it was found that, due to the reluctance of some house owners/renters to disclose certain information, or that they were unclear about certain information, there were 144 samples with missing data for annual electricity consumption and 154 samples with missing data for gas consumption. Therefore, the predictions of electricity consumption and gas consumption are based on 83 and 73 residential buildings, respectively. Then three groups of input parameters are selected based on variable importance (VI) through statistical analysis. Finally, eight data-driven modeling approaches were used to develop electricity and gas consumption prediction models based on different groups of input parameters. The performances of different models were evaluated, and the best prediction models for electricity and gas consumption were identified. The IBM SPSS Statistics 26.0 and Clementine 12.0 were used to apply the algorithm [36]. A flowchart of the research strategy is presented in Figure 1.

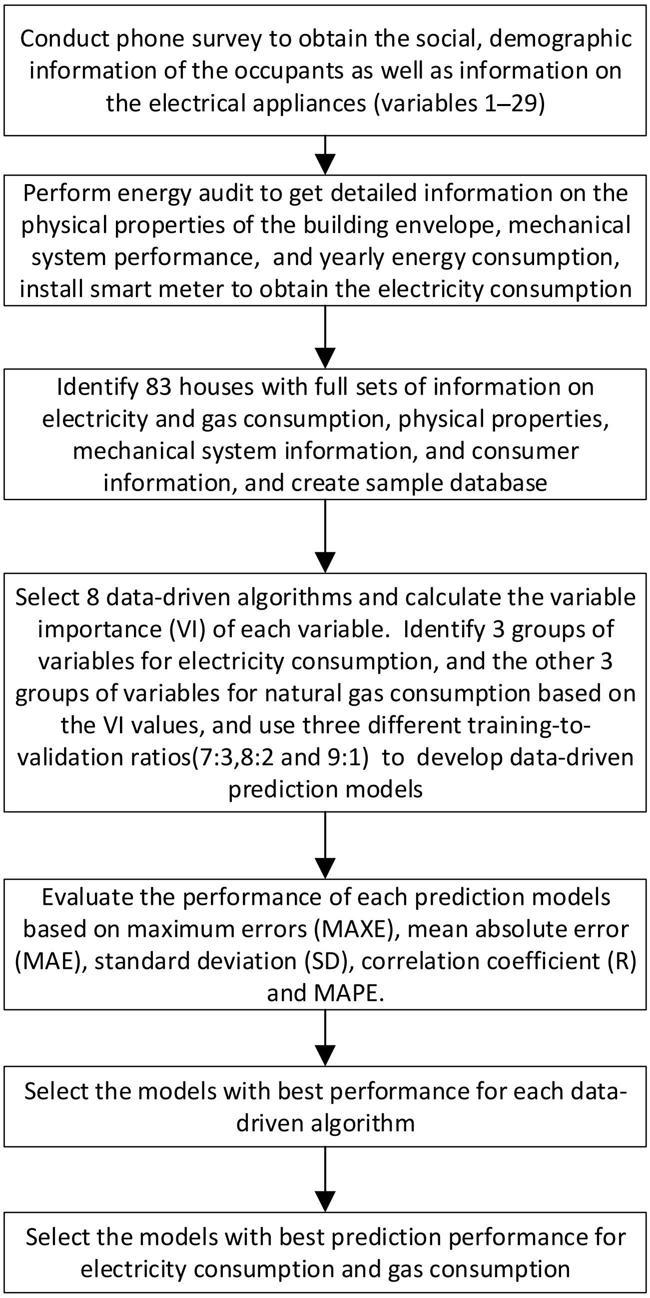


Figure 1. Flowchart of the research strategy.

2.1. Independent and Dependent Variables

Table 1 lists the variable names and their value ranges, where the independent variables 1–29 and 30–63 and dependent variables 64–65 were collected through a phone survey, energy audit, and smart metering. The range of values is formed based on the outcomes from the collected data.

Table 1. Variable names and value ranges.

No.	Information of the Variable	Variable Name	Collecting Method	Value Range
1	Number of halogen bulbs used outdoors	Halogen	Phone survey	0–5
2	Number of compact fluorescent lamp (CFL) bulbs used outdoors	CFL	Phone survey	0–4
3	Number of fluorescent bulbs used outdoors	Fluor	Phone survey	0–4
4	Number of incandescent lamps used outdoors	Incand	Phone survey	0–5
5	Awareness of the importance of reducing energy consumption	RedEnerg	Phone survey	1–5
6	Awareness of the importance of spending less on energy bill	SpenLess	Phone survey	1–5
7	Perceptions of government involvement in energy conservation	GvInvolv	Phone survey	1–5
8	Interested in learning more about ways to save energy indoors	LearnMor	Phone survey	1–5
9	Interest in using computer software to control indoor energy consumption	CompSoft	Phone survey	1–5
10	Number of occupants	NbOccup	Phone survey	1–6
11	Number of residents working full-time	FullTime	Phone survey	0–5
12	Number of residents working part-time	ParTime	Phone survey	0–1
13	Number of residents working in shifts	SiftWork	Phone survey	0–1
14	Number of people working or staying at home	FromHome	Phone survey	0–3
15	Housing situation	HomState	Phone survey	Owned (1), Rent (2)
16	Lights turned on when empty for a short period of time	LOnEmpty	Phone survey	1–3 Occurs more and more frequently
17	The moment when the outdoor lights in front of the house are turned on	TOnOutLt	Phone survey	1–3 Occurs more and more frequently
18	Feeling safe between neighbors	Safety	Phone survey	1–5 Increased sense of security
19	Worry about crime	Crime	Phone survey	1–5 Increased sense of security
20	Age of the homeowner	AgeRange	Phone survey	18–24 (1), 25–35 (2), 36–45 (3), 46–55 (4), 56–65 (5), over 65 (6)
21	Number of energy-saving electrical appliances purchased in the past 5 years	NbNewApp	Phone survey	0–7
22	Fuel type of the oven	OvenFuel	Phone survey	Natural gas (1), electricity (2)
23	Fuel type of the dryer	DryerFl	Phone survey	Natural gas (1), electricity (2)
24	Fuel type of the pool heaters	PHeatrFl	Phone survey	Unused (0), Solar (1), Natural Gas (2), Electricity (3)
25	Upgrade or renovation of the house in the past five to ten years	RecUpgd	Phone survey	Renovated (1), Not renovated (2)
26	Amount willing to spend on energy-efficient equipment (CAD)	WlgSpend	Phone survey	<$100 (1), $100–250 (2), $250–500 (3), >$1000 (4)
27	Highest level of education	LevelEdu	Phone survey	High School (1), College (2), University (3)
28	Gross household income before taxes (CAD/year)	HsIncome	Phone survey	<$20,000 (1), $20,000–$39,999 (2), $40,000–$59,999 (3), $60,000–$79,999 (4), $80,000–$99,999 (5), >$100,000 (6)
29	Born in Canada	BornCan	Phone survey	Yes (1), No (2)
30	Fuel type for heating system	HeatType	Energy audit	Electricity (1), Natural gas (2), Oil (3)

Table 1. *Cont.*

No.	Information of the Variable	Variable Name	Collecting Method	Value Range
31	House type	HsType	Energy audit	Single detached (1), Row end (2)
32	Number of floors	NbStoris	Energy audit	1–2
33	Heating system type	HSysType	Energy audit	Baseboard (1), medium-efficiency furnace (2), heat pump (3), high-efficiency boiler (4)
34	Fuel type for domestic water heaters	DHWFuel	Energy audit	Natural gas (1), Electricity (2)
35	Types of domestic hot water heater	DHWType	Energy audit	Condensing unit (1), Induced draft fan boiler (2), conventional tank heater (3)
36	Existing air-conditioning system	ACSyst	Energy audit	No (0), Yes (1)
37	Air-conditioning system type	ACType	Energy audit	central system (1), heat pump (2), Not applicable (3)
38	Year built	ConstYr	Energy audit	Pre 76 (1),1976–1987 (2), 1988–2002 (3)
39	Heating system efficiency	HSysEffi	Energy audit	76–100%
40	Service length of the heating system (years)	HSysAge	Energy audit	0–35
41	Service length of the air-conditioning system (years)	ACAge	Energy audit	0–33
42	thermal resistance of the window ($m^2{\cdot}K/W$)	TherReWind	Energy audit	0.99–2.64
43	thermal resistance of the external wall ($m^2{\cdot}K/W$)	TherReWal	Energy audit	0.64–3.12
44	thermal resistance of the ceiling ($m^2{\cdot}K/W$)	TherReCei	Energy audit	0.53–7.05
45	Area of the ceiling (m^2)	CeilArea	Energy audit	45.2–227.4
46	Area of the external wall (m^2)	TWlArea	Energy audit	52.8–317.6
47	Area of the window (m^2)	TWdArea	Energy audit	6.7–49.2
48	U-value of foundation wall ($W/(m^2{\cdot}K)$)	FwUvalue	Energy audit	0.23–3.17
49	U-value of the basement ceiling ($W/(m^2{\cdot}K)$)	BhUvalue	Energy audit	0.48–3.87
50	Air change rate per hour at 50 Pa	NbACH	Energy audit	1.49–14.88
51	Residential floor area (m^2)	ReFlArea	Energy audit	49–166
52	Building orientation	OriBuild	Energy audit	1 East 2 West 3 South 4 North 5 Northeast 6 Southeast 7 Northwest 8 Southwest
53	Building width (m)	WidBuild	Energy audit	5.18–16.46
54	Building depth (m)	DepBuild	Energy audit	7.01–16.46
55	Building perimeter (m)	PerBuild	Energy audit	28.65–52.43
56	Window type	TypWind	Energy audit	Single-layer (1), Double-layer (2), Double-layer Low-E (3)
57	Window frame type	TypWindFra	Energy audit	Wood (1), Vinyl (2), Metal (3)
58	Door type	TypDoor	Energy audit	Wood (1), Steel (2)
59	Door area (m^2)	AreDoor	Energy audit	0.94–6.8
60	Cooling system COP	COPRefSys	Energy audit	2–10
61	Ventilation system exhaust volume (m^3/h)	ExVolVenti	Energy audit	1–15
62	Floor area (m^2)	AreFloor	Energy audit	97.8–374.6
63	Total basement wall area (m^2)	AreBaseWal	Energy audit	43.4–117.5
64	Annual electricity consumption (kWh)	AnnPowConsu	Energy audit+smart metering	8944–50,415
65	Annual natural gas consumption (m^3)	AnnNaGEnConsu	Energy audit	0–5937

2.2. Prediction Model Development

The MLR, SR, SVM, BPNN, RBFN, CART, CHAID, and ECHAID were employed to develop electricity consumption and gas consumption prediction models.

2.2.1. Multiple Linear Regression

MLR has been widely used in building energy consumption prediction and can be used in the early design stage to improve the building performance [37] and hourly cooling load prediction [7]. In this paper, MLR is used to develop the relationship between the independent variables (variables 1–63), and dependent variables (variables 64 and 65). The MLR model can be presented as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_p + \varepsilon \tag{1}$$

where β_0 denotes the regression constant; $\beta_1, \beta_2,$ and β_p denote the regression coefficients; x_i refers to the input variables; ε is the random error; and p denotes the number of independent variables involved in the regression.

The regression coefficients are determined based on the least square method, which minimizes the residual sum of squares (RSS). The RSS is calculated by the following equation:

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p\right)^2 \tag{2}$$

where n is the number of samples.

2.2.2. Stepwise Regression

The SR uses a step-by-step iterative approach to develop a regression model by selecting only the important independent variables. It is also widely used in building simulation [38]. In this paper, 63 independent variables were introduced into the regression model one-by-one and sorted according to their importance. Each dependent variable goes through an F-test and T-test and remains in the model if it is statistically significant.

2.2.3. Support Vector Machine

The SVM introduces the principle of structural risk minimization, which effectively solves the high-dimensional difficulty and local minima problem. Gao [39] developed an SVM model to predict building energy consumption based on historical data with good prediction performance. By studying the output/input variables relationship, the SWM predicts the output variable values of new samples with the same distribution as the training sample set. A loss function is introduced to correct the distance to the decision boundary, so as to determine the regression function. Thus, a prediction model is developed to predict the outputs for new samples with the same distribution [40].

2.2.4. Backpropagation Neural Network

The BPNN is the most widely used neural network. As a multilayer feed-forward neural network, it is trained according to an error backpropagation algorithm [41]. BPNN features arbitrarily complex pattern classification ability and demonstrates excellent multi-dimensional function mapping ability. It includes an input, a hidden, and an output layer. The least square error of the network is obtained by using the gradient descent method to for minimization.

2.2.5. Radial Basis Function Neural Network

RBFN utilizes radial basis functions (RBFs) as activation functions. The RBF network consists only of a single hidden layer that has its own way of computing the output. The input layer receives the input data and feeds them into the special hidden layer. The computations in the hidden layers are based on comparisons with prototype vectors from the training set. Each neuron computes the similarity between the input vector and its prototype vector. RBFN has been proven to have a good prediction performance for the building cooling load [13].

2.2.6. CART

The CART is a classification algorithm that builds a decision tree based on Gini's impurity index [42]. It applies the binary segmentation method to recursively construct the binary decision tree process and uses the square error minimization criterion for feature selection for the regression tree. CART has been proven to achieve good performance in heating energy prediction [18].

2.2.7. CHAID

CHAID is based on adjusted significance testing, which was proposed by Kass et al. [43]. In this method, multi-branch decision trees can be generated. First, the F-test is carried out, and variables statistically similar to the target variable are combined; then *p*-values for the remaining variables are calculated, and the variable with the best predictor (lowest *p*-value) is selected as the first variable in the decision tree branches. The process repeats until the tree is fully grown. It has been successfully used to predict the steam load [20].

2.2.8. Exhaustive CHAID

As an improved algorithm based on CHAID, ECHAID is different from CHAID on the merging step [44]. The latter stops when all remaining categories are found to be statistically different. The former continues grouping, leaving only two super categories. In this way, all input variables are ensured to have the same degree of freedom. It has been successfully employed to predict the performance of heat pumps [21].

2.3. Choice of Input Variables

In order to eliminate the variables that are unimportant to the prediction of building energy consumption, the variable importance (VI) is employed to assist in the selection of the input variables to develop prediction models; detailed information in the calculation can be found in Ref. [45]. At the same time, the ratios of samples for training and validation are set as 7:3, 8:2, and 9:1, respectively. The data are split randomly.

2.4. Prediction Model Evaluation

The prediction model performance is evaluated through maximum errors (MAXEs), mean absolute error (MAE), standard deviation (SD), correlation coefficient (R), and MAPE. The MAE, SD, and R can be calculated as shown below:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}(|\hat{y}_i - y_i|) \tag{3}$$

$$\mathrm{SD} = \sqrt{\frac{\sum_{i=1}^{n}(|\hat{y}_i - y_i| - \mathrm{MAE})^2}{n}} \tag{4}$$

$$\mathrm{R} = \sqrt{1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - y)^2}} \tag{5}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}(|\hat{y}_i - y_i|) \times 100\% \tag{6}$$

where $\hat{y}_i$ denotes the prediction value, y_i denotes the targeted value, $\overline{y}$ denotes the average value of the targeted values, and n is the number of samples.

Evaluation on the validation of the performance of the prediction model based on MAXE, MAE, SD, R, and MAPE under different training-to-validation ratios (7:3, 8:2, and 9:1) to ensure the best performance and the least amount of data for training.

3. Results and Discussion

3.1. Results of Variable Selection

Depending on the variable importance (VI) of each variable, totals of 26, 12, and 6 variables were selected to develop the prediction models for electricity consumption (Table 2), and totals of 26, 13, and 6 variables were selected to develop the prediction models for natural gas consumption (Table 3).

Table 2. Variable selected for predicting electricity consumption.

Number of Variables	Variable Set
26 (importance of variable (IV) $\geq$ 0.01)	HeatType, DHWFuel, AreFloor, HSysEffi, HSysAge, HSysType, Halogen, NbOccup, TherReCei, FromHome, ACSyst, OriBuild, LOnEmpty, TherReWal, SpenLess, Incand, NbACH, PHeatrFl, AgeRange, LearnMor, ExVolVenti, FullTime, TWdArea, ConstYr, COPRefSys, CFL
12 (IV $\geq$ 0.016)	HeatType, DHWFuel, AreFloor, HSysEffi, HSysAge, HSysType, Halogen, NbOccup, TherReCei, FromHome, ACSyst, OriBuild
6 (IV $\geq$ 0.05)	HeatType, DHWFuel, AreFloor, HSysEffi, HSysAge, HSysType

Table 3. Variable selected for predicting natural gas consumption.

Number of Variables	Variable Set
26 (IV $\geq$ 0.015)	HeatType, NbACH, HSysEffi, TWlArea, Fluor, DHWFuel, Halogen, TherReWind, TherReWal, PerBuild, RedEnerg, NbOccup, PHeatrFl, SpenLess, TypWindFra, CeilArea, OvenFuel, BhUvalue, DHWType, ReFlArea, TherReCei, WidBuild, HomState, FwUvalue, AreBaseWal, AreFloor
13 (IV $\geq$ 0.022)	HeatType, NbACH, HSysEffi, TWlArea, Fluor, DHWFuel, Halogen, TherReWind, TherReWal, PerBuild, RedEnerg, NbOccup, PHeatrFl
6 (IV $\geq$ 0.032)	HeatType, NbACH, HSysEffi, TWlArea, Fluor, DHWFuel

3.2. Performance of Electricity Consumption Prediction Model

Analyses of the results of the prediction models for electricity consumption are presented in Tables A1–A8 in Appendix A. The regressions between predicted and simulated electricity consumption for the best models of each data-driven approach are presented in Figure 2a–h.

The outcomes of the MLR models on the prediction of electricity consumption are listed in Appendix A Table A1. It can be found that when the number of variables is 6 and the ratio of training sample vs. validation samples is 9:1, the MLR model has the best performance, with MAPEs of 15.05% for training and 11.71% for validation, respectively. Figure 2a presents the regression between predicted and simulated electricity consumption for the best MLR model. The model predicts pretty well when the electricity consumption is less than 35,000 kWh (93% of all the samples), and it underpredicts the electricity consumption when it exceeds 35,000 kWh.

The outcomes of the SR models on the prediction of electricity consumption are listed in Appendix A Table A2; they are similar to those of the MLR models. When the number of variables is 6 and the ratio of training sample vs. validation samples is 9:1, the SR model has the best performance with MAPEs of 14.79% for training and 14.18% for validation, respectively. Figure 2b presents the regression between predicted and simulated electricity consumption for the best SR model. The model also predicts pretty well when the electricity consumption is less than 35,000 kWh, and it underpredicts the electricity consumption when it exceeds 35,000 kWh.

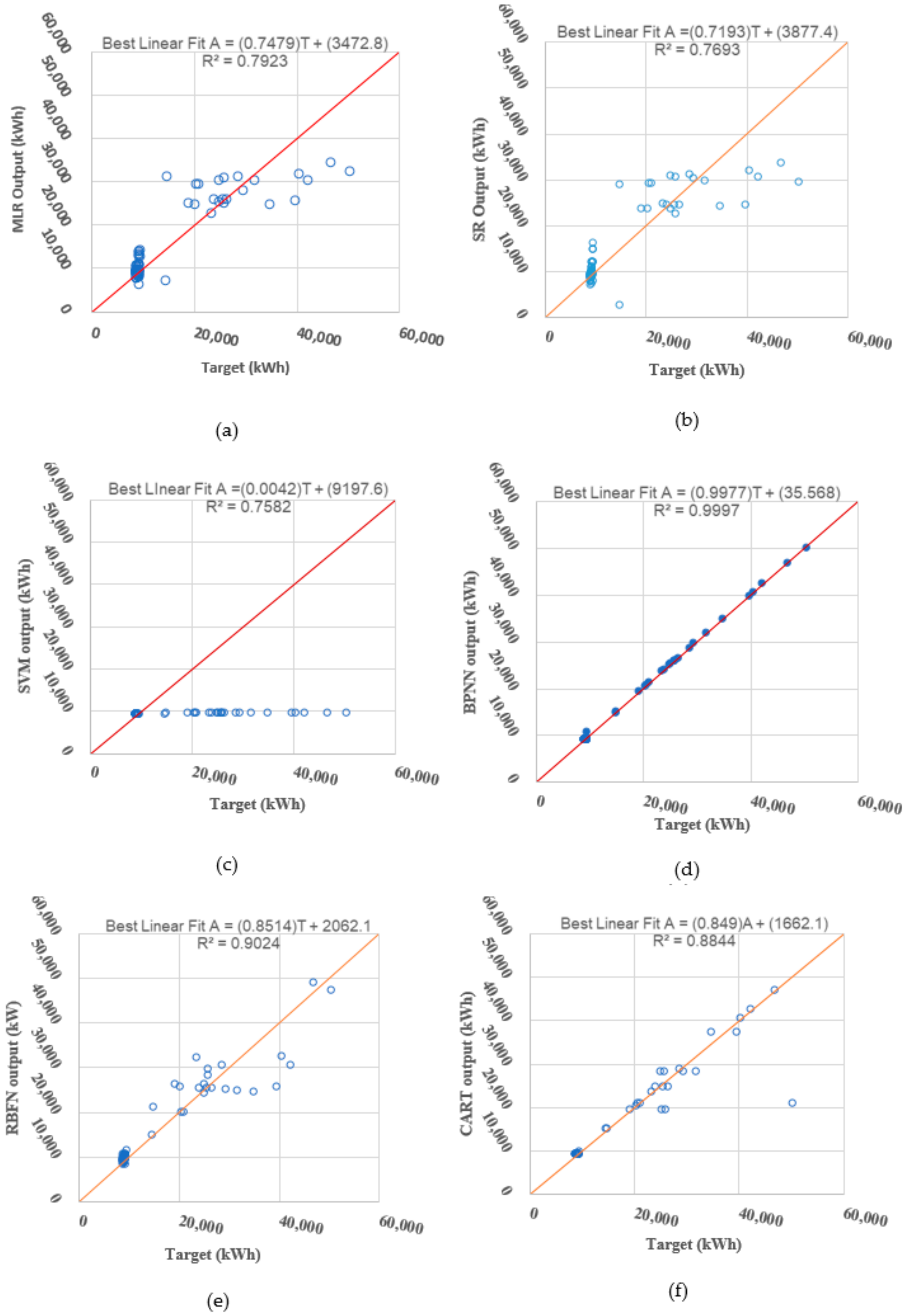
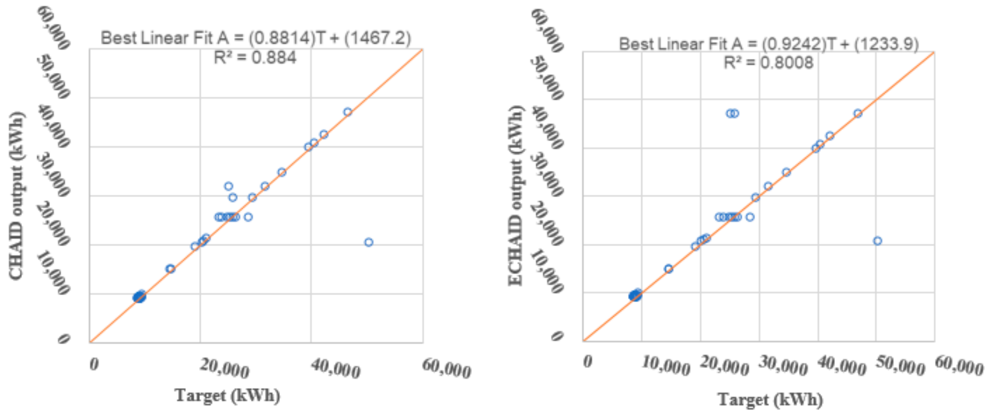


Figure 2. *Cont.*



(g) (h)

Figure 2. Regression between predicted and simulated electricity consumption: (**a**) MLR model vs. (**b**) BPNN model vs. (**c**) SVM vs. (**d**) BPNN model vs. (**e**) RNFN model vs. (**f**) CART model vs. (**g**) CHAID vs. (**h**) ECHAID model.

The outcomes of the SVM models on the prediction of electricity consumption are listed in Appendix A Table A3. It can be found that when the number of variables is 6 and ratio of training sample vs. validation samples is 7:3, the SVM model has the best performance, with MAPEs of 21.89% for training and 11.50% for validation, respectively. Figure 2c presents the regression between predicted and simulated electricity consumption for the best SVM model. The model predicts pretty well when the electricity consumption is around 10,000 kWh, and it underpredicts the electricity consumption when it exceeds 15,000 kWh.

The outcomes of the BPNN models on the prediction of electricity consumption are listed in Appendix A Table A4. It can be found that when the number of variables is 26 and the ratio of training sample vs. validation samples is 9:1, the BPNN model has the best performance, with MAPEs of 0.94% for training and 0.94% for validation, respectively. The number of inputs can be reduced to 12, with a correlation coefficient almost equal to 1.0 and MAPE less than 1.18%. Figure 2d presents the regression between predicted and simulated electricity consumption for the best BPNN model. Compared with the results from Ndiaye and Gabriel (2011), the R-square value is significantly improved from 0.79 to 0.9997. The model predicts pretty well for all the samples.

The outcomes of the RBFN models on the prediction of electricity consumption are listed in Appendix A Table A5. It can be found that when the number of variables is 6 and the ratio of training sample vs. validation samples is 8:2, the RBFN model has the best performance, with MAPEs of 8.82% for training and 5.62% for validation, respectively. Figure 2e presents the regression between predicted and simulated electricity consumption for the best RBFN model. The model predicts pretty well when the electricity consumption is less than 35,000 kWh, and it tends to underpredict the electricity consumption when it is in the range of 35,000–40,000 kWh.

The outcomes of the CART models on the prediction of electricity consumption are listed in Appendix A Table A6. It can be found that when the number of variables is 6 and the ratio of training sample vs. validation samples is 7:3, the CART model has the best performance, with MAPEs of 1.41% for training and 5.50% for validation, respectively. Figure 2f presents the regression between predicted and simulated electricity consumption for the best CART model. The model predicts pretty well for almost all the samples, with the exception that it underpredicts one sample with actual consumption at around 50,000 kWh.

The outcomes of the CHAID models on the prediction of electricity consumption are listed in Appendix A Table A7. It can be found that when the number of variables is 26 and the ratio of training sample vs. validation samples is 7:3, the CHAID model has the best performance, with MAPEs of 0.87% for training and 5.03% for validation, respectively.

Figure 2g presents the regression between predicted and simulated electricity consumption for the best CHAID model. Similar to the CART model, it predicts pretty well for almost all the samples, with the exception that it underpredicts one sample with actual consumption at around 50,000 kWh.

The outcomes of the ECHAID models on the prediction of electricity consumption are listed in Appendix A Table A8. It can be found that when the number of variables is 26 and the ratio of training sample vs. validation samples is 7:3, the CHAID model has the best performance, with MAPEs of 0.92% for training and 9.89% for validation, respectively. Figure 2h presents the regression between predicted and simulated electricity consumption for the best ECHAID model. It predicts pretty well for almost all the samples, except that it overpredicts two samples with actual consumption at around 26,000 kWh and underpredicts one sample with actual consumption at around 50,000 kWh.

Table 4 presents the range of relative errors for the eight best prediction models for each data-driven approach. It can be found that the BPNN model has the best prediction performance, followed by the CHAID model, ECHAID model, CART model, and RBFN model. The performances of the SVM, SR, and MRL models are not as good as the other ones.

Table 4. Range of relative errors for the eight electricity consumption prediction models.

Method	≤5%	≤15%	≤25%	≤50%
MLR	38%	64%	79%	98%
SR	43%	68%	81%	94%
SVM	73%	73%	73%	75%
BPNN	99%	100%	100%	100%
RBFN	57%	85%	92%	100%
CART	89%	97%	98%	99%
CHAID	93%	98%	98%	99%
ECHAID	93%	97%	97%	97%

3.3. Performance of Natural Gas Consumption Prediction Model

The outcomes of the natural gas consumption prediction models are listed in Tables A9–A16 in Appendix A. The regressions between predicted and simulated natural gas consumption for the best models of each data-driven approach are presented in Figure 3a–h.

The outcomes of the MLR models on the prediction of natural gas consumption are listed in Appendix A Table A9. It can be found that when the number of variables is 13 and the ratio of training sample vs. validation samples is 7:3, the MLR model has the best performance, with MAPEs of 13.98% for training and 24.67% for validation, respectively. Figure 3a presents the regression between predicted and simulated natural gas consumption for the best MLR model. Good agreements are found between the predicted and actual energy consumption.

The outcomes of the SR models on the prediction of natural gas consumption are listed in Appendix A Table A10. Similar to the MLR model, when the number of variables is 13 and the ratio of training sample vs. validation samples is 7:3, the SR model has the best performance, with MAPEs of 14.03% for training and 24.89% for validation, respectively. Figure 3b presents the regression between predicted and simulated natural gas consumption for the best SR model. Good agreements are found between the predicted and actual energy consumption.

The outcomes of the SVM models on the prediction of natural gas consumption are listed in Appendix A Table A11. It can be found that when the number of variables is 26 and the ratio of training sample vs. validation samples is 7:3, the SVM model has the best performance, with MAPEs of 59.47% for training and 53.23% for validation, respectively. Figure 3c presents the regression between predicted and simulated natural gas consumption for the best SVM model. Large deviations between the predicted value and actual energy consumption are found.

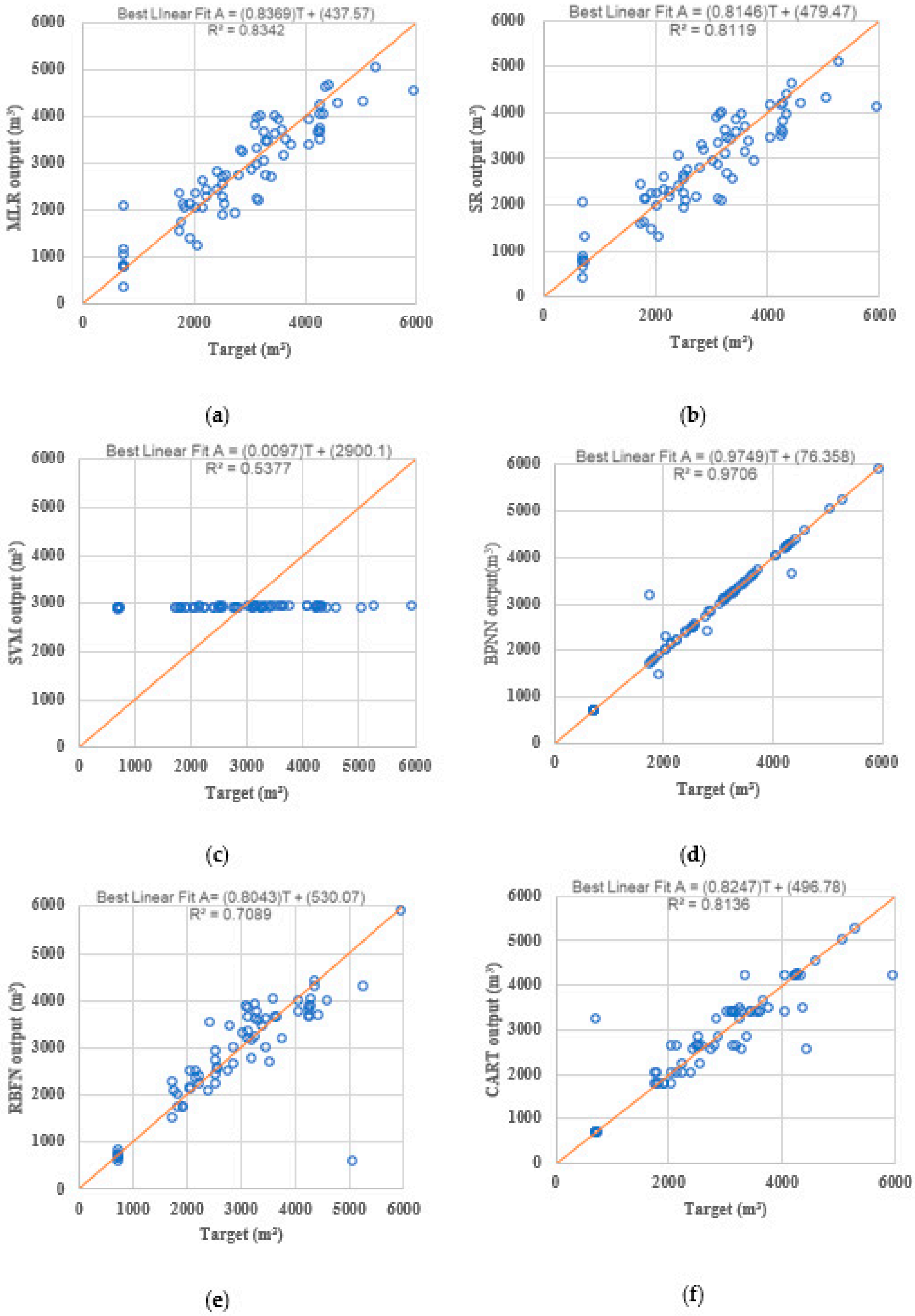


Figure 3. *Cont*.

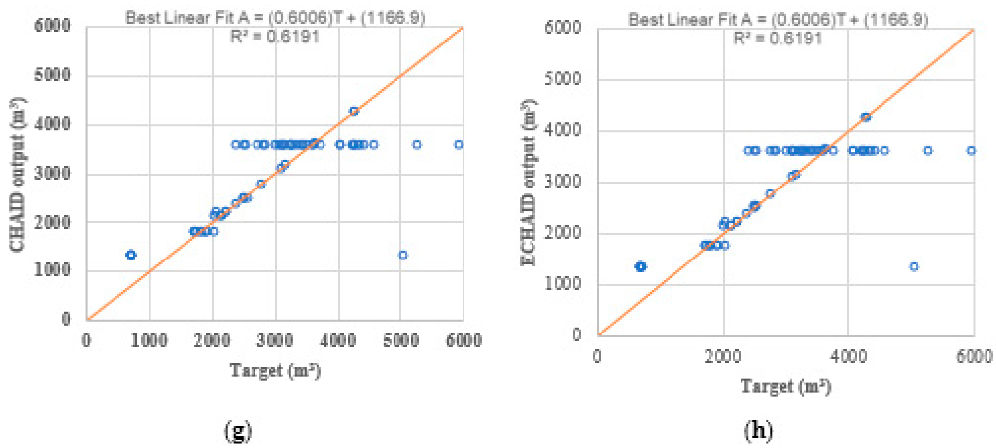


Figure 3. Regression between predicted and simulated natural gas consumption: (**a**) MLR model vs. (**b**) BPNN model vs. (**c**) SVM vs. (**d**) BPNN model vs. (**e**) RNFN model vs. (**f**) CART model vs. (**g**) CHAID vs. (**h**) ECHAID model.

The outcomes of the BPNN models on the prediction of natural gas consumption are listed in Appendix A Table A12. It can be found that when the number of variables is 26 and the ratio of training sample vs. validation samples is 9:1, the BPNN model has the best performance, with MAPEs of 2.63% for training and 0.16% for validation, respectively. The number of inputs can be reduced to 13, with a correlation coefficient higher than 0.979 and MAPEs less than 7.03%. When the number of inputs is reduced to 6, the correlation coefficient is still higher than 0.927, with MAPEs less than 11.63%. Figure 3d presents the regression between predicted and simulated natural gas consumption for the best BPNN model. The model predicts pretty well for almost all the samples.

The outcomes of the RBFN models on the prediction of natural gas consumption are listed in Appendix A Table A13. It can be found that when the number of variables is 6 and ratio of training sample vs. validation samples is 8:2, the RBFN model has the best performance, with MAPEs of 12.85% for training and 7.57% for validation, respectively. Figure 3e presents the regression between predicted and simulated natural gas consumption for the best RNFN model. The model predicts pretty well for all the samples, except under predicting one sample with natural gas consumption of 5049 m^3.

The outcomes of the CART models on the prediction of natural consumption are listed in Appendix A Table A14. It can be found that when the number of variables is 13 and the ratio of training sample vs. validation samples is 7:3, the CART model has the best performance with MAPEs of 5.08% for training and 31.56% for validation, respectively. Figure 3f presents the regression between predicted and simulated natural gas consumption for the best CART model. The model predicts generally well for most of the samples, with big deviations for only a few samples.

The outcomes of the CHAID models on the prediction of natural consumption are listed in Appendix A Table A15. It can be found that when the number of variables is 6 and the ratio of training sample vs. validation samples is 7:3, the CHAID model has the best performance, with MAPEs of 18.74% for training and 24.72% for validation, respectively. Figure 3g presents the regression between predicted and simulated natural gas consumption for the best CHAID model. It can be observed that the model predicts generally well for some of the samples; however, for some of the samples, the natural gas consumption is predicted to be about 3600 m^3 regardless of their actual consumption.

The outcomes of the ECHAID models on the prediction of natural consumption are listed in Appendix A Table A16. Similar to the CHAID model, when the number of variables is 6 and the ratio of training sample vs. validation samples is 7:3, the ECHAID model has the best performance, with MAPEs of 18.74% for training and 24.72% for validation, respectively. Figure 3h presents the regression between predicted and simulated natural gas consumption for the best ECHAID model, which is similar to the CHAID model.

Table 5 presents the ranges of relative errors for the eight best prediction models. It can be found that the BPNN model has the best prediction performance, followed by the CART model and RBFN model. The performance of other models is much poorer, with the SVM model being the worst case.

Table 5. Range of relative errors for the eight natural gas consumption prediction models.

Method	≤5%	≤15%	≤25%	≤50%
MLR	22%	62%	83%	98%
SR	25%	60%	82%	98%
SVM	6%	32%	48%	78%
BPNN	93%	96%	99%	99%
RBFN	30%	75%	93%	99%
CART	49%	83%	93%	99%
CHAID	38%	60%	76%	87%
ECHAID	38%	60%	76%	87%

4. Conclusions and Limitations

In this paper, eight data-driven methods were employed to develop energy prediction models for residential buildings in Oshawa with different numbers of input variables and training to validation ratios. The following conclusions can be made:

(1) The performance of the prediction model can be improved through careful selections of variables based on VI and training to validation ratios. As only a small number of input variables are used, it can also help reduce the efforts of data collection.
(2) With 26 input variables, the BPNN models have the best performance in predicting both the electricity consumption and gas consumption because their maximum error, mean absolute error, standard deviation, and MAPE are smaller than those of other models, and their correlation coefficient is larger than that of other models.
(3) The MLR model has the worst performance in predicting the electricity consumption, and the SVM model has the worst performance in natural gas consumption prediction.
(4) The number of inputs can be reduced to 12 in the BPNN model to predict the electricity consumption, with a correlation coefficient almost equal to 1.0 and MAPE $\leq$ 1.18%. By using the CART model, the number of inputs can be further reduced to 6, with a correlation coefficient $\geq$0.95 and MAPE $\leq$ 5.50%.
(5) The number of inputs can be reduced to 13 in the BPNN model for natural gas consumption prediction with a correlation coefficient $\geq$0.979 and MAPE $\leq$ 7.03%. When it is further reduced to 6, the correlation coefficient of the BPNN model is still $\geq$0.927, with the MAPE $\leq$ 11.63%.
(6) Based on the performance of the prediction models, when the human factor, e.g., SpenLess (awareness of the importance of spending less on energy bills), FromHome (number of people working or staying at home), and HomState (housing situation), are introduced, the performance of the prediction model can be improved. Those variables are often very difficult to introduce to develop physical models in traditional methods.

The limitations of the prediction models are as follows:

(1) They can only be applied to residential buildings (houses) in Oshawa and cannot be applied to commercial buildings.
(2) More data collection is needed, including weather data, to develop prediction models that are applicable throughout Canada.

Author Contributions: Y.L. and K.G. contributed to the conception of the study and the development of the methodology; Y.L., J.L., K.G., W.Y. and C.-Q.L. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by NATURAL SCIENCE FOUNDATION OF HUBEI PROVINCE, grant number 2017CFB602.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to their containing information that could compromise the privacy of the research participants.

Acknowledgments: The authors acknowledge the support from Natural Resources Canada, Oshawa Public Utility Corporation, Ontario Center for Excellence, and the Faculty of Engineering and Applied Science of University of Ontario Institute of Technology.

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

Appendix A

Table A1. Analysis of the results of the MLR model for electricity consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	9217	2759	3657	0.94	20.8%
		Validation	18,347	4300	5789	0.79	36.8%
	8:2	Training	10,174	2751	3691	0.93	20.5%
		Validation	17,416	3841	5310	0.85	32.1%
	9:1	Training	10,686	2568	3597	0.93	19.1%
		Validation	15,901	3571	5221	0.91	26.8%
12	7:3	Training	13,489	3040	4542	0.90	20.0%
		Validation	13,655	2242	3501	0.95	18.5%
	8:2	Training	13,496	2905	4428	0.90	19.2%
		Validation	13,830	2444	3733	0.95	19.8%
	9:1	Training	14,043	2712	4205	0.90	18.4%
		Validation	13,415	2748	4244	0.96	20.5%
6	7:3	Training	14,332	2864	4892	0.88	16.1%
		Validation	18,652	2207	4268	0.92	16.6%
	8:2	Training	14,339	2780	4795	0.88	15.8%
		Validation	19,260	2215	4560	0.93	15.5%
	9:1	Training	14,231	2584	4563	0.89	15.0%
		Validation	18,420	2179	4971	0.95	11.7%

Table A2. Analysis of the results of the SR model for electricity consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	12,178	3189	4520	0.90	21.9%
		Validation	17,646	2683	4364	0.91	21.9%
	8:2	Training	12,116	3080	4428	0.90	21.3%
		Validation	17,879	2728	4593	0.91	21.4%
	9:1	Training	12,450	2840	4209	0.90	19.8%
		Validation	17,765	3196	5387	0.92	22.9%
12	7:3	Training	13,208	3228	4722	0.89	21.2%
		Validation	17,633	2751	4371	0.91	22.5%
	8:2	Training	13,126	3109	4621	0.89	20.6%
		Validation	17,894	2811	4616	0.91	22.1%
	9:1	Training	13,636	2880	4402	0.89	19.2%
		Validation	17,612	3151	5314	0.94	22.3%
6	7:3	Training	15,664	2898	5033	0.87	16.4%
		Validation	21,563	2565	4918	0.90	18.2%
	8:2	Training	15,638	2814	4916	0.88	16.2%
		Validation	21,503	2681	5174	0.91	18.2%
	9:1	Training	15,443	2583	4694	0.88	14.8%
		Validation	21,016	2688	5740	0.95	14.2%

Table A3. Analysis of the results of the SVM model for electricity consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	37,611	6290	10,341	0.81	21.9%
		Validation	41,168	3595	9408	0.85	11.5%
	8:2	Training	37,611	5980	10,166	0.81	20.9%
		Validation	41,171	4051	9934	0.83	12.9%
	9:1	Training	37,612	5521	9791	0.82	19.5%
		Validation	41,171	5096	11,658	0.86	15.0%
12	7:3	Training	37,567	6278	10,325	0.84	21.9%
		Validation	41,129	3588	9396	0.86	11.5%
	8:2	Training	37,564	5969	10,150	0.84	20.9%
		Validation	41,127	4043	9920	0.86	12.9%
	9:1	Training	37,567	5511	9775	0.85	19.4%
		Validation	41,130	5086	11,643	0.89	14.9%
6	7:3	Training	37,514	6268	10,311	0.86	21.9%
		Validation	41,063	3582	9380	0.92	11.5%
	8:2	Training	37,519	5960	10,137	0.86	20.8%
		Validation	41,068	4036	9904	0.92	12.8%
	9:1	Training	37,515	5502	9761	0.87	19.4%
		Validation	41,064	5078	11,624	0.93	14.9%

Table A4. Analysis of the results of the BPNN model for electricity consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	16,131	2806	4381	0.91	16.5%
		Validation	13,618	2024	3386	0.95	14.4%
	8:2	Training	2554	422	833	1.00	1.9%
		Validation	156	237	411	1.00	1.5%
	9:1	Training	345	87	171	1.00	0.9%
		Validation	435	110	155	1.00	0.9%
12	7:3	Training	7112	376	1002	1.00	1.8%
		Validation	2735	300	549	1.00	1.9%
	8:2	Training	4586	743	1329	0.99	3.5%
		Validation	1803	427	566	1.00	2.7%
	9:1	Training	564	81	133	1.00	0.8%
		Validation	236	136	188	1.00	1.1%
6	7:3	Training	11,857	872	2110	0.98	3.9%
		Validation	2443	364	800	1.00	2.3%
	8:2	Training	13,089	1697	3586	0.94	7.7%
		Validation	3652	345	865	1.00	1.7%
	9:1	Training	17,032	2187	4537	0.89	10.3%
		Validation	20,134	1723	5297	0.94	6.5%

Table A5. Analysis of the results of the RBFN model for electricity consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	19,346	4214	5336	0.86	28.2%
		Validation	6519	2216	2641	0.96	20.1%
	8:2	Training	14,505	2846	4444	0.90	16.8%
		Validation	15,093	2274	4082	0.91	13.7%
	9:1	Training	13,076	2774	4252	0.90	19.1%
		Validation	8920	1942	2715	0.99	12.9%

Table A5. *Cont.*

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
12	7:3	Training	15,797	2482	4227	0.91	14.3%
		Validation	3274	1135	1440	0.99	9.5%
	8:2	Training	17,058	3167	4966	0.87	19.5%
		Validation	7338	1788	2498	0.98	15.1%
	9:1	Training	15,795	2094	3855	0.92	12.2%
		Validation	2710	1154	1459	0.99	8.8%
6	7:3	Training	15,105	2100	3925	0.93	10.5%
		Validation	2989	902	1268	0.99	7.8%
	8:2	Training	14,315	1878	3708	0.93	8.8%
		Validation	3392	764	1095	1.00	5.6%
	9:1	Training	13,931	1428	2855	0.96	8.6%
		Validation	895	628	1142	1.00	6.0%

Table A6. Analysis of the results of the CART model for electricity consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	5224	460	1207	0.99	2.2%
		Validation	10,680	1420	3846	0.92	5.5%
	8:2	Training	5224	444	1176	0.99	2.1%
		Validation	10,680	1586	4097	0.92	5.9%
	9:1	Training	5224	618	2086	0.98	2.9%
		Validation	10,680	1408	3319	0.97	4.8%
12	7:3	Training	3850	275	717	1.00	1.2%
		Validation	18,575	1965	5466	0.83	7.0%
	8:2	Training	3850	268	700	1.00	1.2%
		Validation	18,575	2203	5825	0.83	7.7%
	9:1	Training	3850	462	1888	0.98	2.1%
		Validation	18,575	2354	6174	0.85	7.4%
6	7:3	Training	3745	338	881	1.00	1.4%
		Validation	29,551	1790	5937	0.87	5.5%
	8:2	Training	3745	327	859	1.00	1.4%
		Validation	29,551	2006	6298	0.87	6.1%
	9:1	Training	5915	387	1079	0.99	1.7%
		Validation	29,551	2629	7685	0.85	7.1%

Table A7. Analysis of the results of the CHAID model for electricity consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	3175	167	547	1.00	0.9%
		Validation	29,983	1684	6132	0.76	5.0%
	8:2	Training	3175	169	534	1.00	0.9%
		Validation	29,346	1846	6396	0.77	5.3%
	9:1	Training	10,496	833	2403	0.97	0.9%
		Validation	29,346	2831	8204	0.71	5.3%
12	7:3	Training	18,988	2538	5279	0.86	10.2%
		Validation	22,535	1191	4515	0.89	3.7%
	8:2	Training	18,988	2415	5143	0.86	9.8%
		Validation	22,535	1329	4807	0.89	3.9%
	9:1	Training	19,124	2166	4837	0.87	8.7%
		Validation	22,671	1798	5932	0.89	4.7%

Table A7. *Cont.*

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
6	7:3	Training	18,988	2547	5279	0.86	10.3%
		Validation	22,535	1193	4515	0.89	3.7%
	8:2	Training	18,988	2420	5143	0.86	9.8%
		Validation	22,535	1332	4808	0.89	4.0%
	9:1	Training	19,124	2168	4837	0.87	8.8%
		Validation	22,671	1801	5932	0.89	4.76%

Table A8. Analysis of the results of the ECHAID model for electricity consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	3175	171	547	1.00	0.9%
		Validation	29,983	2928	8492	0.65	9.9%
	8:2	Training	3175	144	530	1.00	0.7%
		Validation	29,346	3272	8953	0.65	11.0%
	9:1	Training	18,441	1987	4555	0.89	7.5%
		Validation	21,988	1858	5803	0.89	5.2%
12	7:3	Training	18,259	2338	4962	0.88	9.0%
		Validation	21,806	1246	4432	0.89	3.9%
	8:2	Training	18,259	2216	4834	0.88	8.5%
		Validation	21,806	1382	4720	0.89	4.1%
	9:1	Training	18,441	2006	4555	0.89	7.7%
		Validation	21,988	1841	5808	0.89	5.0%
6	7:3	Training	18,259	2343	4962	0.88	9.1%
		Validation	21,806	1249	4432	0.89	3.9%
	8:2	Training	18,259	2221	4834	0.88	8.6%
		Validation	21,806	1377	4721	0.89	4.1%
	9:1	Training	18,441	2010	4555	0.89	7.8%
		Validation	21,988	1846	5808	0.89	5.0%

Table A9. Analysis of the results of the MLR model for natural gas consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	768	271	340	0.96	11.9%
		Validation	1452	603	835	0.77	32.6%
	8:2	Training	771	261	334	0.96	11.4%
		Validation	1460	662	876	0.76	35.8%
	9:1	Training	763	277	343	0.96	12.0%
		Validation	2172	734	1052	0.69	43.0%
13	7:3	Training	964	326	409	0.94	14.0%
		Validation	1381	526	649	0.86	24.7%
	8:2	Training	969	316	402	0.94	13.5%
		Validation	1394	577	684	0.86	27.1%
	9:1	Training	902	315	402	0.94	13.2%
		Validation	1831	666	855	0.82	33.5%
6	7:3	Training	2892	512	729	0.79	22.3%
		Validation	2494	469	757	0.81	21.0%
	8:2	Training	2882	506	720	0.78	21.9%
		Validation	2515	480	785	0.82	21.5%
	9:1	Training	2878	472	686	0.81	20.1%
		Validation	1523	458	714	0.88	26.2%

Table A10. Analysis of the results of the SR model for natural gas consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	1152	317	403	0.94	13.2%
		Validation	1723	625	806	0.78	30.7%
	8:2	Training	1153	305	395	0.94	12.7%
		Validation	1723	693	850	0.78	34.1%
	9:1	Training	989	327	415	0.93	13.5%
		Validation	1850	700	870	0.81	33.6%
13	7:3	Training	1091	331	426	0.93	14.0%
		Validation	1790	554	696	0.84	24.9%
	8:2	Training	1085	321	417	0.93	13.6%
		Validation	1797	608	733	0.83	27.4%
	9:1	Training	989	327	415	0.93	13.5%
		Validation	1850	700	870	0.81	33.6%
6	7:3	Training	2568	564	755	0.77	28.4%
		Validation	2482	585	866	0.73	28.0%
	8:2	Training	2559	559	744	0.77	27.9%
		Validation	2503	612	893	0.74	29.4%
	9:1	Training	2982	485	694	0.80	19.5%
		Validation	2493	533	908	0.79	27.2%

Table A11. Analysis of the results of the SVM model for natural gas consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	2313	940	1164	0.75	59.5%
		Validation	2991	958	1262	0.74	53.2%
	8:2	Training	2312	928	1148	0.75	58.3%
		Validation	2990	993	1311	0.78	56.0%
	9:1	Training	2192	926	1142	0.77	57.1%
		Validation	2872	1019	1427	0.74	73.4%
13	7:3	Training	2319	945	1168	0.80	59.9%
		Validation	2989	958	1265	0.78	53.4%
	8:2	Training	2317	933	1152	0.78	58.6%
		Validation	2986	993	1313	0.78	56.2%
	9:1	Training	2206	930	1146	0.77	57.3%
		Validation	2875	1019	1428	0.81	73.5%
6	7:3	Training	2325	947	1170	0.69	59.8%
		Validation	3000	959	1266	0.71	53.3%
	8:2	Training	2325	935	1154	0.69	58.6%
		Validation	3000	994	1314	0.74	56.1%
	9:1	Training	2215	933	1148	0.70	57.3%
		Validation	2887	1017	1430	0.83	73.3%

Table A12. Analysis of the results of the BPNN model for natural gas consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	1334	263	309	0.97	11.0%
		Validation	551	272	322	0.97	13.2%
	8:2	Training	1467	145	276	0.97	6.3%
		Validation	272	102	125	1.00	5.2%
	9:1	Training	663	55	226	0.98	2.6%
		Validation	13	2	5	1.00	0.2%

Table A12. *Cont.*

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
13	7:3	Training	262	118	148	0.99	5.1%
		Validation	487	173	233	0.98	6.0%
	8:2	Training	809	191	259	0.98	8.2%
		Validation	186	138	168	0.99	6.3%
	9:1	Training	848	192	243	0.98	7.0%
		Validation	533	239	286	0.98	9.2%
6	7:3	Training	1463	374	460	0.92	11.9%
		Validation	1033	373	545	0.91	11.3%
	8:2	Training	2650	427	617	0.85	14.3%
		Validation	975	342	435	0.96	14.2%
	9:1	Training	913	338	435	0.93	11.6%
		Validation	512	282	376	0.97	10.3%

Table A13. Analysis of the results of RBFN model for natural gas consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	1320	470	607	0.86	18.7%
		Validation	973	458	587	0.89	23.2%
	8:2	Training	2848	525	717	0.79	21.2%
		Validation	1031	470	596	0.89	20.0%
	9:1	Training	2896	476	691	0.80	16.6%
		Validation	804	477	618	0.90	21.4%
13	7:3	Training	1171	381	469	0.92	16.5%
		Validation	789	319	407	0.95	15.6%
	8:2	Training	1424	441	539	0.89	18.2%
		Validation	706	346	420	0.95	17.0%
	9:1	Training	1816	447	562	0.87	17.4%
		Validation	666	419	496	0.94	20.4%
6	7:3	Training	2928	633	740	0.81	26.3%
		Validation	1008	500	712	0.84	27.7%
	8:2	Training	4432	394	744	0.79	12.9%
		Validation	695	230	324	0.97	7.6%
	9:1	Training	4555	395	731	0.79	13.2%
		Validation	461	222	246	0.99	9.3%

Table A14. Analysis of the results of the CART model for natural gas consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	634	133	209	0.98	5.0%
		Validation	1840	689	994	0.64	34.3%
	8:2	Training	660	164	247	0.98	5.7%
		Validation	2569	817	1155	0.55	39.2%
	9:1	Training	834	154	252	0.98	5.4%
		Validation	2440	723	1135	0.61	43.5%
13	7:3	Training	634	139	212	0.98	5.1%
		Validation	1840	605	924	0.69	31.6%
	8:2	Training	660	173	261	0.97	5.9%
		Validation	2569	705	1076	0.60	35.2%
	9:1	Training	834	162	264	0.97	5.6%
		Validation	2440	680	1124	0.63	41.5%

Table A14. *Cont.*

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
6	7:3	Training	494	117	210	0.98	3.8%
		Validation	2569	806	1335	0.40	47.6%
	8:2	Training	660	143	222	0.98	4.5%
		Validation	2569	891	1406	0.34	51.9%
	9:1	Training	979	172	299	0.97	5.5%
		Validation	2569	998	1681	0.28	68.5%

Table A15. Analysis of the results of the CHAID model for natural gas consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	1366	271	438	0.93	7.7%
		Validation	2038	665	1012	0.64	37.4%
	8:2	Training	1411	280	442	0.93	8.1%
		Validation	2083	638	1015	0.65	37.0%
	9:1	Training	1589	242	441	0.92	6.6%
		Validation	2261	1007	1386	0.43	60.5%
13	7:3	Training	1246	254	421	0.93	7.9%
		Validation	672	708	988	0.66	41.4%
	8:2	Training	1915	430	647	0.83	14.5%
		Validation	2587	794	1184	0.46	43.5%
	9:1	Training	1390	230	407	0.94	6.1%
		Validation	2062	994	1392	0.40	57.4%
6	7:3	Training	3714	385	722	0.79	18.7%
		Validation	2339	612	799	0.78	24.7%
	8:2	Training	3714	377	709	0.79	18.2%
		Validation	2351	656	843	0.78	26.6%
	9:1	Training	1640	305	478	0.91	9.7%
		Validation	2312	861	1515	0.34	64.0%

Table A16. Analysis of the results of the ECHAID model for natural gas consumption.

Number of Variables	Training: Validation	Data Set	MAX Error	MAE	SD	R	MAPE
26	7:3	Training	1246	288	482	0.91	8.1%
		Validation	4164	920	1439	0.28	45.9%
	8:2	Training	1246	168	366	0.95	4.4%
		Validation	4164	1065	1551	0.19	51.4%
	9:1	Training	1589	242	441	0.92	6.6%
		Validation	2261	1007	1386	0.43	60.5%
13	7:3	Training	1913	397	643	0.84	13.1%
		Validation	2585	754	1136	0.47	40.6%
	8:2	Training	1915	382	631	0.84	12.6%
		Validation	2587	830	1201	0.45	44.8%
	9:1	Training	1150	243	427	0.93	6.9%
		Validation	1692	873	1294	0.52	58.2%
6	7:3	Training	3714	385	722	0.79	18.7%
		Validation	2339	612	799	0.78	24.7%
	8:2	Training	3714	377	709	0.79	18.2%
		Validation	2351	656	843	0.78	26.6%
	9:1	Training	1640	306	478	0.91	9.7%
		Validation	2312	861	1515	0.34	64.0%

References

1. International Energy Agency (IEA). Available online: https://www.iea.org/reports/the-critical-role-of-buildings (accessed on 19 September 2022).
2. Lin, Y.; Zhou, S.; Yang, W.; Li, C.-Q. Design optimization considering variable thermal mass, insulation, absorptance of solar radiation and glazing ratio using prediction model and Genetic Algorithm. *Sustainability* **2018**, *10*, 336. [CrossRef]
3. Lin, Y.; Yang, W. Application of Multi-objective Genetic Algorithm based Simulation for Cost-effective Building Energy Efficiency Design and Thermal Comfort Improvement. *Front. Energy Res.* **2018**, *6*, 25. [CrossRef]
4. Zhu, D.; Yan, D.; Wang, C.; Hong, T. Comparison of building energy consumption simulation software: DeST, EnergyPlus and DOE-2. *Build. Sci.* **2012**, *28*, 213–222.
5. Fumo, N.; Mago, P.; Luck, R. Methodology to estimate building energy consumption using EnergyPlus Benchmark Models. *Energy Build.* **2010**, *42*, 2331–2337. [CrossRef]
6. Amiri, S.S.; Mottahedi, M.; Asadi, S. Using multiple regression analysis to develop energy consumption indicators for commercial buildings in the U.S. *Energy Build.* **2015**, *109*, 209–216. [CrossRef]
7. Chen, S.; Zhou, X.; Zhou, G.; Fan, C.; Ding, P.; Chen, Q. An online physical-based multiple linear regression model for building's hourly cooling load prediction. *Energy Build.* **2022**, *254*, 11574. [CrossRef]
8. Ciulla, G.; D'Amico, A. Building energy performance forecasting: A multiple linear regression approach. *Appl. Energy* **2019**, *253*, 113500. [CrossRef]
9. Tso, G.K.F.; Yau, K.K.W. Predicting electricity energy consumption-A comparison of regression analysis, decision tree and networks. *Energy* **2007**, *32*, 1761–1768. [CrossRef]
10. Zhao, L.; Lin, Y.; Huang, X. Prediction Model for Energy Consumption and Visual Discomfort of Passive House Based on Stepwise Regression Analysis. *Build. Energy Effic.* **2021**, *49*, 50–55, 69.
11. Ma, Z.; Ye, C.; Ma, W. Support vector regression for predicting building energy consumption in southern China. *Energy Procedia* **2019**, *158*, 3433–3438. [CrossRef]
12. Li, Q.; Meng, Q.; Mochida, A. Applying support vector machine to predict hourly cooling load in the building. *Appl. Energy* **2009**, *86*, 2249–2256. [CrossRef]
13. Paudel, S.; Elmitri, M.; Le Corre, O. A relevant data selection method for energy consumption prediction of low energy building based on support vector machine. *Energy Build.* **2017**, *138*, 240–256. [CrossRef]
14. Ahmad, M.W.; Mourshed, M.; Rezgui, Y. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy Build.* **2017**, *147*, 77–89. [CrossRef]
15. Han, Y.; Fan, C.; Yu, B. Energy efficient building envelope using novel RBF neural network integrated affinity propagation. *Energy* **2020**, *209*, 118414. [CrossRef]
16. Zhao, C.; Lin, S.; Xu, Q. Prediction of building energy consumption in collegue buildings based on GM-RBF neural network. *J. Nanjing Univ. Sci. Technol.* **2014**, *38*, 48–53.
17. Zekić-Sušac, M.; Has, A.; Knežević, M. Predicting energy cost of public buildings by artificial neural networks, CART, and random forest. *Neurocomputing* **2021**, *439*, 223–233. [CrossRef]
18. Capozzoli, A.; Grassi, D.; Causone, F. Estimation models of heating energy consumption in schools for local authorities planning. *Energy Build.* **2015**, *105*, 302–313. [CrossRef]
19. Yang, J.; Wu, J. Research on energy-saving optimization of commercial central air-conditioning based on data mining algorithm. *Energy Build.* **2022**, *272*, 112326. [CrossRef]
20. Kusiak, A.; Li, M.; Zhang, Z. A data-driven approach for steam load prediction in buildings. *Appl. Energy* **2010**, *87*, 925–933. [CrossRef]
21. Yan, L.; Hu, P.; Li, C.; Yao, Y.; Xing, L.; Lei, F.; Zhu, N. The performance prediction of ground source heat pump system based on monitoring data and data mining technology. *Energy Build.* **2016**, *127*, 1085–1095. [CrossRef]
22. Li, K.; Xie, X.; Yang, X. A hybrid teaching-learning artificial neural network for building electrical energy consumption prediction. *Energy Build.* **2018**, *174*, 323–334. [CrossRef]
23. Moayedi, H.; Mu'azu, M.A.; Foong, K.K. Novel swarm-based approach for predicting the cooling load of residential buildings based on social behavior of elephant herds. *Energy Build.* **2020**, *206*, 109579. [CrossRef]
24. Aruta, G.; Ascione, F.; Boettcher, O.; De Masi, R.F.; Mauro, G.M.; Vanoli, G.P. Machine learning to predict building energy performance in different climates. *IOP Conf. Ser. Earth Environ. Sci.* **2022**, *1078*, 012137. [CrossRef]
25. Ndiaye, D.; Gabriel, K. Principal component analysis of the electricity consumption in residential dwellings. *Energy Build.* **2011**, *43*, 446–453. [CrossRef]
26. Shen, M.; Sun, H.; Lu, Y. Household Electricity Consumption Prediction Under Multiple Behavioural Intervention Strategies Using Support Vector Regression. *Energy Procedia* **2017**, *142*, 2734–2739. [CrossRef]
27. Jain, R.K.; Smith, K.M.; Taylor, J.E. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. *Appl. Energy* **2014**, *123*, 168–178. [CrossRef]
28. Rahman, A.; Srikumar, V.; Smith, A.D. Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. *Appl. Energy* **2017**, *212*, 372–385. [CrossRef]

29. Wang, Z.; Liu, X.; Li, H. Energy performance prediction of vapor-injection air source heat pumps in residential buildings using a neural network model. *Energy Build.* **2020**, *228*, 110499. [CrossRef]
30. Kim, T.-Y.; Cho, S.-B. Predicting residential energy consumption using CNN-LSTM neural networks. *Energy* **2019**, *182*, 72–81. [CrossRef]
31. Bui, D.-K.; Nguyen, T.N.; Nguyen-Xuan, H. An artificial neural network (ANN) expert system enhanced with the electromagnetism-based firefly algorithm (EFA) for predicting the energy consumption in buildings. *Energy* **2020**, *190*, 116370. [CrossRef]
32. Farzana, S.; Liu, M.; Hossain, M.U. Multi-model prediction and simulation of residential building energy in urban areas of Chongqing, South West China. *Energy Build.* **2014**, *81*, 161–169. [CrossRef]
33. Guo, Z.; Moayedi, H.; Bahiraei, M. Optimal modification of heating, ventilation, and air conditioning system performances in residential buildings using the integration of metaheuristic optimization and neural computing. *Energy Build.* **2020**, *214*, 109866. [CrossRef]
34. Kerdan, I.G.; Gálvez, D.M. Artificial neural network structure optimisation for accurately prediction of exergy, comfort and life cycle cost performance of a low energy building. *Appl. Energy* **2020**, *280*, 115862. [CrossRef]
35. Chegari, B.; Tabaa, M.; Medromi, H. Multi-objective optimization of building energy performance and indoor thermal comfort by combining artificial neural networks and metaheuristic algorithms. *Energy Build.* **2021**, *239*, 110839. [CrossRef]
36. IBM. Available online: https://www.ibm.com/spss (accessed on 1 November 2022).
37. Asadi, S.; Amiri, S.S.; Mottahedi, M. On the development of multi-linear regression analysis to assess energy consumption in the early stages of building design. *Energy Build.* **2014**, *85*, 246–255. [CrossRef]
38. Wang, M.; Wright, J.; Brownlee, A.; Buswell, R. A comparison of approaches to stepwise regression on variables sensitivities in building simulation and analysis. *Energy Build.* **2016**, *127*, 313–326. [CrossRef]
39. Gao, Y. Research on Building Energy Consumption Prediction Method Based on machine Learning. PhD. Thesis, Beijing University of Civil Engineering and Architecture, Beijing, China, 2020.
40. Xue, W. *SPSS Modeler Data Mining Methods and Applications, 3rd. ed.*; Electronics Industry Press: Beijing, China, 2020.
41. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
42. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wadsworth Inc.: New York, NY, USA, 1984.
43. Kass, G.V. An exploratory technique for investigating large quantities of categorical data. *J. R. Stat. Soc.* **1980**, *29*, 119–127. [CrossRef]
44. Biggs, D.; de Ville, B.; Suen, E. A method of choosing multiway partitions for classification and decision trees. *J. Appl. Stat.* **1991**, *18*, 49–62. [CrossRef]
45. Zhang, X.; Wada, T.; Fujiwara, K.; Kano, M. Regression and independence based variable importance measure. *Comput. Chem. Eng.* **2020**, *135*, 106757. [CrossRef]