energies

MDPI

Article

Development of an Improved Model to Predict Building Thermal Energy Consumption by Utilizing Feature Selection

Jihoon Jang [1], Joosang Lee [1], Eunjo Son [1], Kyungyong Park [1], Gahee Kim [1], Jee Hang Lee [2] and Seung-Bok Leigh [1,*]

[1] Department of Architectural Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea; eands777@yonsei.ac.kr (J.J.); batthoman@naver.com (J.L.); archicho88@gmail.com (E.S.); kypark193@yonsei.ac.kr (K.P.); kazzang@yonsei.ac.kr (G.K.)

[2] Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea; jeehang@kaist.ac.kr

* Correspondence: sbleigh@yonsei.ac.kr; Tel.: +82-2-2123-7830

Received: 4 October 2019; Accepted: 1 November 2019; Published: 2 November 2019

Abstract: Humans spend approximately 90% of the daytime in buildings, and greenhouse gases (GHGs) emitted by buildings account for approximately 20% of total GHG emissions. As the energy consumed during building operation from a building life-cycle perspective amounts to approximately 70–90% of the total energy, it is essential to accurately predict the energy consumption of buildings for their efficient operation. This study aims to optimize a model for predicting the thermal energy consumption of buildings by (i) first extracting major variables through feature selection and deriving significant variables in addition to the collected data and (ii) predicting the thermal energy consumption using a machine learning model. Feature selection using random forest was performed, and 11 out of 17 available data were selected. The accuracy of the prediction model was significantly improved when the hour of day variable was added. The prediction model was constructed using an artificial neural network (ANN), and the improvement in the prediction accuracy was analyzed by comparing different cases of variable combinations. The ANN prediction accuracy was improved by 15% using the feature selection process compared to when all data were used as input data, and 25% coefficient of variation of the root mean square error (CVRMSE) accuracy was achieved.

Keywords: thermal energy; artificial neural network; feature selection; building operation; building energy conservation; building energy consumption

1. Introduction

1.1. Background

The emission of greenhouse gas (GHG) has steeply increased by approximately 82.5% since 1979 [1], which has been a major contributor to climate change globally. This gives rise to considerable global efforts to reduce its emissions. As a part of it, South Korea announced a target of reducing GHG emissions by 37% compared to Business As Usual (BAU) by 2030 in the Paris Agreement of the Intergovernmental Panel on Climate Change (IPCC) in 2016 [2].

Humans, who are mostly responsible for GHG emissions, spend approximately 90% of the daytime in buildings [3]. The estimated GHG emissions from buildings are as high as 20% of the total emissions [4]. Therefore, many studies have been conducted to reduce energy consumption in buildings in an effort to reduce GHG emissions from buildings.

The application of renewable energy to buildings has increased recently in an effort to reduce the energy consumption of buildings. To properly distribute the supply of renewable energy sources, it is necessary to accurately predict the energy consumption of each building. In particular, from a building life-cycle perspective, the energy consumed in the operation stage of a building amounts to approximately 70–90% of the total energy. Therefore, it is essential to accurately predict the energy consumption of buildings for their efficient operation [5,6].

1.2. Literature Review (Analysis of Previous Studies)

In this section, previous studies were reviewed to introduce (i) the necessity of predicting building energy consumption, (ii) a comparative analysis of energy prediction methods, and (iii) the advantages of feature selection. Based on these, the optimal modeling strategy for predicting building energy consumption was developed.

Linda et al. showed that energy distribution systems can be planned by predicting heat and electricity demand. They created regression analysis models for buildings using the heat and electricity consumptions measured every hour and predicted the heat and electricity loads as well as the annual energy demand in a specific area. As a result, they optimized a system for energy distribution to buildings [7]. Yudong Ma et al. conducted research on energy reduction in buildings using an energy demand prediction. They constructed a system optimized for storing the heat source generated from a cooling system and a prediction control system for preventing the overproduction of heat sources. The simulation results showed that an energy-saving effect of up to 24.5% can be achieved [8]. The district heating system efficiently provides heat sources using a centralized supply of cold and hot water used in local facilities. Therefore, it is possible to determine an optimal operation plan before using chillers and boilers if their loads can be accurately predicted. M. Sakawa et al. predicted the load using an artificial neural network (ANN) for efficient district cooling and heating operation [9]. Kody M. Powell et al. conducted research on the energy reduction of heating, ventilation and air conditioning (HVAC) using a building energy consumption prediction model. They presented an efficient method of coping with the peak load by controlling the thermal storage system [10]. Samuel et al. conducted a study to predict the heating load in a district heating system by applying four types of machine learning. They used district heating data collected from 10 residential and commercial buildings located in Skellefteå, Sweden, and using machine learning techniques, namely, support vector machine (SVM), forward backpropagation neural network (FFNN), multilinear regression (MLR), regression. Prediction of energy consumption shows that it is a key step towards the realization of optimized energy production, distribution, and consumption [11]. To improve present and future energy supplies, predicting energy demands is an essential stage. However, the lack of accurate and comprehensive data sets to predict future demand is one of the big problems in developing economies. Therefore, Sasan et al. proposed an ensemble hybrid forecasting model as a solution for predicting energy consumption while dealing with the shortage of data sets [12]. These studies controlled the production and operation of heat sources by predicting the building energy consumption and the induced energy reduction. They also prevented the overproduction of heat sources through the energy consumption prediction.

For the prediction of building energy consumption, both traditional methods and methods using artificial intelligence (AI) have benefits and drawbacks. Methods using AI, however, attract much attention due to their reliability and high prediction accuracy [13]. The most active research over the past several years is on load prediction based on neural networks. In particular, ANN requires no load model for calculation [14]. Alberto et al. used an ANN model and a simulation (EnergyPlus) model to predict the energy demand of a university building. When the prediction accuracy of ANN was compared with that of EnergyPlus, ANN exhibited higher accuracy. For the ANN model, when a prediction model using only the external dry-bulb temperature as a variable was compared with a prediction model using temperature, humidity, and solar radiation as variables, the latter exhibited higher accuracy [15]. Tso et al. compared and analyzed the prediction accuracy of regression analysis, decision tree, and neural network models for predicting the electricity energy consumption of Hong Kong. Among them, the prediction results of the decision tree and neural network models for summer and winter times were more accurate than those of the regression model [16]. Nonlinear factors, such as weather and the indoor condition of a building, significantly affect the energy consumption prediction accuracy [17]. The traditional prediction methods have limitations processing such nonlinear factors. AI is the most suitable method for dealing with nonlinear factors and provides better prediction performance [18]. Almonacid, F. et al. showed that a method using ANN exhibited a higher accuracy than the traditional method for estimating the amount of energy using existing formulas when the annual energy produced by photovoltaic (PV) power generation was estimated [19]. Kialashaki conducted research on an ANN model and a multilinear regression (MLR) model to estimate the energy consumption of industries in the United States. Several independent variables, such as GDP and energy prices, were used for the analysis, and a comparison of the two models revealed that the ANN model had a higher prediction accuracy than the MLR model [20]. L. Ekonomou et al. used a multilayer perceptron (MLP) model based on four factors (climate condition, GDP, energy demand, and power capacity) to predict the energy consumption of Greece. As a result, it was found that the MLP model was more accurate and reliable than the regression model and the support vector machine (SVM) model [21]. For the prediction of heating energy consumption of a university campus, Radisa et al. used various artificial neural networks. Of the three ANN networks, feed-forward backpropagation neural network (FFNN) showed the highest accuracy, and the ensemble model of the three networks showed more accurate predictions [22]. Muhammad et al. compared the performance of the widely-used feed-forward back-propagation artificial neural network with random forest, an ensemble-based method for predicting the hourly HVAC energy consumption of a hotel in Madrid, Spain. Overall, in this study, ANN performed marginally better than random forest for prediction with root mean squared error (RMSE) of 4.97 and 6.10, respectively [23]. Therefore, as shown by previous studies, AI-based ANN models can replicate the characteristics of various variables more accurately than conventional statistical regression and simulation models.

A large number of candidate input data exist for building energy consumption prediction. Feature selection is a method for increasing the prediction accuracy of a model when there are many candidate input data. Better learning and inference can be possible for a prediction model by removing ineffective candidates and reducing the size of the input set, leading to improved accuracy. The purpose of feature selection is to find the minimal subset from the entire set while preserving the major information to facilitate future analysis. Oveis Abedinia et al. significantly reduced the training time, forecast time, and daily mean absolute percentage error (DMAPE) by selecting seven out of 41 features [24]. Ping Jiang et al. revealed that high-quality feature selection improves the prediction accuracy and training speed, and minimizes modeling complexity. In other words, feature selection is a major element for successful prediction and plays an important role in predicting the load [25]. Liu et al. performed feature selection based on Pearson's correlation analysis to predict the load of a building and created a prediction model using an improved Elman neural network (IENN). In the study, feature selection optimized the weight of the model and brought better prediction results [26]. Zhao et al. used a support vector regression (SVR) model for building energy consumption prediction. They reduced the data set using feature selection because the building energy consumption was complicated and was affected by many factors. When SVR was trained by selecting subsets from three data sets, the accuracy of the model was improved, and the runtime was reduced [27]. Sooyoun et al. selected electric energy as a subset of total building energy consumption and identified the variables that contribute to electric energy use. The K-means and density-based spatial clustering of applications with noise (DBSCAN) clustering techniques were used to select the features directly affecting the consumption among 16 features, and they compared the prediction results using only the main variables with the results using all variables. As a result of this study, they explained that selecting and using sensors also make it possible to find the most significant measurement points because the data can be used to obtain clustering results for correlation analysis [28]. Aurora et al. dealt with multivariate time-dependent series of data points for energy forecasting in smart buildings. They applied different types of feature selection methods for regression tasks. The results of the experiments carried out show that the proposed methodology effectively reduces both the complexity of the forecast model and their RMSE and mean absolute error (MAE) [29]. Therefore, the use of feature selection in predicting energy consumption can improve the prediction accuracy and reduce the runtime by extracting the most important input data set.

1.3. Study Objectives

Despite the development of technologies for predicting energy consumption due to advances in AI technology, many of the previous studies performed prediction using the collected raw data as they were or did not systematically consider the variable selection process when constructing prediction models. Dynamic data have been accumulated in large quantities due to the development of the internet of things (IoT) technology and information and communication technology (ICT), and methods for processing such data have been established [30]. When an energy consumption prediction model is implemented, it is necessary to secure the prediction accuracy by removing data with low importance that interfere with prediction and select only major variables. Moreover, it is possible to improve the prediction accuracy of a model if variables with significant relevance are added to the model in addition to the data collected by conventional sensors [31]. Therefore, this study was conducted to optimize a model for predicting the thermal energy consumed for heating and domestic hot water (DHW) in buildings during winter. In this study, a model for accurately predicting thermal energy consumption was developed by (i) collecting sensor data to determine the indoor and outdoor conditions of the chosen building (e.g., indoor temperature, indoor humidity, outdoor temperature, irradiation, etc.), (ii) extracting important variables through feature selection, (iii) deriving significant variables in addition to the collected sensor data, and iv) constructing the ANN model using the selected input data. The performance of the final optimized model was then analyzed and compared with previous prediction models.

2. Methodology

2.1. Building and Sensor Descriptions

The Jincheon eco-friendly energy town is located in the Chungbuk Innovation City of South Korea. The center of the town is 36.9 °N, 127.5 °E, and Köppen's climate characteristic is humid subtropical (Cwa) [32]. The experimental period, from 1 December 2017 to 30 April 2018, has a temperature range of −16.2 to 34.2 °C and a humidity range of 15.4% to 99.3%. For this town, located in the central inland basin area, the wind speed is relatively weak, the weather is usually clear, and the sunshine time is long. The average annual temperature is 12.5 °C.

There are 6 types of public buildings in town: central machine room, high school, youth center, library, health care center, daycare center. A high school building located in the Jincheon town was specifically used for this study. The town has a large-capacity thermal storage tank for storing solar heat, which supplies the stored thermal energy to public buildings in the town for heating and DHW. Eight hundred square meters of two types of solar collectors (i.e., flat-plate type and evacuated type) are connected in series, and the collected heat is stored in the large-capacity thermal storage tank throughout the year. The large-capacity thermal storage tank is a solar heat storage system, where a storage tank of 25.2m × 17.2m × 9.6m (L × W × H) is installed on the ground. If the stored solar heat is not sufficient, it is supplemented by a heat pump in the central machine room of the town. Therefore, the large-capacity thermal storage tank with solar collectors acts as solar district heating for a net-zero energy community in the town [33].

Table 1 shows the details of the building, and Figure 1 shows the front view and the floor plan of the building. There is also a machine room with a distribution system that receives heat from the thermal storage tank and delivers it to each room of the building. After the thermal energy is supplied from the central large-capacity thermal storage tank to the machine room of the subject building, the thermal energy is supplied to each room by fan coil units (FCU). Technical information on the chosen building is presented in Table 2. As this study focused on winter, the cooling energy in summer was not considered.

Table 1. Building details.

Principal Use	Maximum Height	Total Floor Area	Building Area
High school	14.7m	10,431.85 m^2	3871.92 m^2

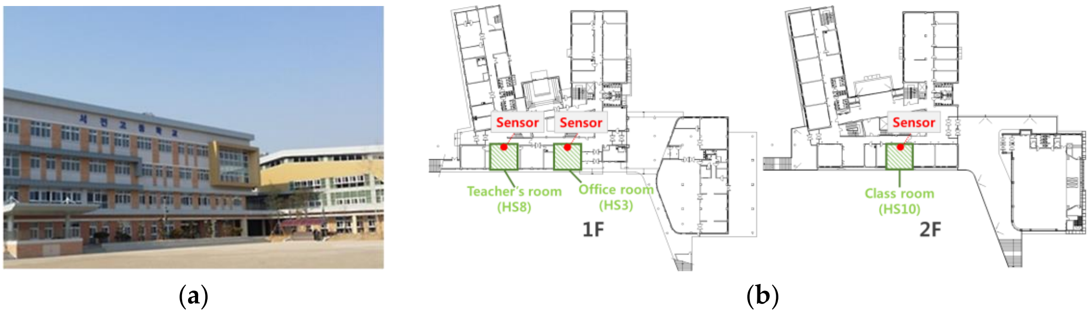


Figure 1. High school building used for this study: (**a**) View of High school; (**b**) Floor plan and sensor locations. (Name of the data points: Office room is HS3; Teacher's room is HS8; Classroom HS10.)

Table 2. Technical information of the building.

Type	Nominal Operating Conditions	Value
Water Circulation Pump	Flow Rate(l/min) Head (m)	2200 10
Heat Exchanger	Capacity for heating (kcal/h) Capacity for DHW (kcal/h)	800,000 100,000
FCU	Capacity for heating (kcal/h) Flow Rate (l/min)	8450 18

Table 3 exhibits the information of the three rooms by a field survey conducted in the high school. The chosen building has three types of occupants: support staff, teachers, and students. Each occupant differs from its daily pattern of the activity, which is rather regulative. Support staff tend to stay continuously in the office room (HS3) during working hours except for lunch time and break times. Teachers spend most of their working hours teaching in classrooms, not in the teachers' room (HS8). The occupancy rate of the teachers' room, therefore, is relatively lower than in other rooms. Contrary to this, students rarely leave the classroom (HS10), hence, a high occupancy rate. According to this, the office room, the teachers' room, and a classroom that can be representative of three main occupants were selected for the study. Particularly, a classroom with the most similar conditions to the office room and teachers' room was selected in the contexts of sizes and orientations. Consequently, indoor environment data were collected from the office room, the teachers' room, and the classroom considering the occupancy rate and occupancy density.

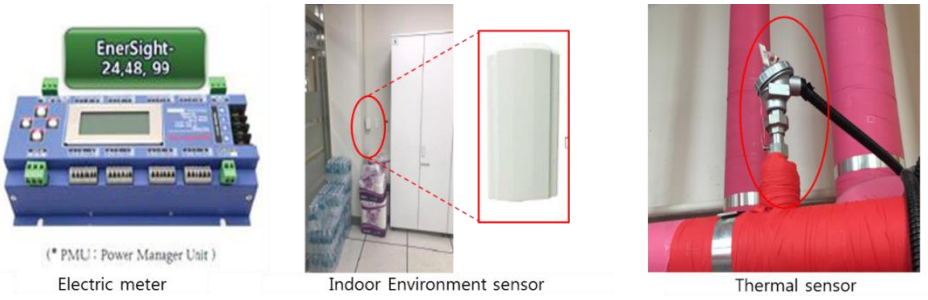
Table 4 shows a list of the data collected to create a thermal energy consumption prediction model. Figure 1b shows the locations of the three rooms where the indoor environment sensors were installed. Figure 2 shows the electricity energy sensor, indoor environment sensor, and central heating sensor installed in the building. The data from the sensors were stored in chronological order at one-hour intervals. As shown in Table 5, there were three types of sensors used for the monitoring of this experiment: watt-hour sensor, indoor environment sensor, and calorimeter sensor. All sensors could communicate over Wi-Fi networks. The watt-hour sensor showed a voltage and current measurement error of up to ±0.2%, and the power measurement error was ±0.1%. The indoor environment sensor could measure room temperature, humidity, and CO_2 concentration. The room temperature showed a measuring range of 0–50 °C, and the measuring range of humidity was 0–95%. CO_2 could be measured from 0 to 10,000 ppm, with a measurement error of ± 5%. The calorimeter measures temperature and flow rate, the temperature showed the measuring range of 0–135 °C, and the flow rate measured 10.0–250 m^3/h. Each error range is ±5% and ±2 m^3/h, respectively.

Table 3. Room information.

	Office Room (HS3)	Teachers' Room (HS8)	Classroom (HS10)
The type of occupants	Support staff	Teachers	Students
The number of occupants	6	6	23
Area (m^2)	56	63	57
Occupancy rate	0.88	0.13	0.90
Occupancy density (1·person/m^2)	0.11	0.10	0.40
Installed FCU (EA)	1	1	1

Table 4. List of collected sensor data.

No.	Category		Feature List
1	Indoor environment data	Office (HS3)	Temperature (°C) Humidity (%) Concentration of CO_2 (ppm)
		Teachers' room (HS8)	Temperature (°C) Humidity (%) Concentration of CO_2 (ppm)
		Classroom (HS10)	Temperature (°C) Humidity (%) Concentration of CO_2 (ppm)
2	Outdoor environment data		Outdoor temperature (°C) Outdoor humidity (%) Solar radiation (W/m^2)
3	Central heating data		Return water temperature (°C) Supply water temperature (°C) Flow meter (L/min)
4	Electricity energy data		Lighting (kWh) Plug load (kWh)
5	Thermal energy consumption (kWh)		



Electric meter

Indoor Environment sensor

Thermal sensor

Figure 2. Sensor types in the building.

Table 5. Information on each sensor.

Experimental Sensor	Measuring Range	Accuracy	Resolution
Watt-hour sensor	30–120 A	±0.2% A, ±0.1% P	0.01 kW
Temperature sensor	0–50 °C	±2 °C	0.01 °C
Humidity sensor	0–95% RH	±2% RH	0.01% RH
CO_2 sensor	0–10,000 ppm	±5% measurement Value	0.01 ppm
Calorimeter (Temperature)	0–135 °C	±5% measurement Value, ±1 °C	2.5–50 °C
Calorimeter (Flow)	10.0–250 m^3/h	±2 m^3/h	0.1 m^3/h

For the indoor environment, the temperature, humidity, and CO_2 concentration were measured. The data were used as variables to identify the behavioral pattern of the occupants as well as the use schedules of the rooms. Since building energy is significantly affected by the outdoor environment [34], outdoor environment data were used to secure the accuracy of the prediction model. For the outdoor environment data, the temperature, humidity, and solar radiation data provided by the Korea Meteorological Administration were used. The central thermal data included the supply water temperature from the center, return water temperature, and flow rate data. The corresponding sensor data were measured to represent how much of the hot water supplied from the central machine room was consumed in the building. The electricity energy data were composed of the lighting and plug loads used in the building. Based on these data, the thermal energy consumption of the chosen building was predicted.

Table 6 shows the operation plan of the studied building. In winter, heating was operated from 0700 to 2200 hours. Heating was operated until the late hour of 2200 for the students' after-school self-study. There were 590 users of the building, including 510 students and 80 employees and teachers. Heating was operated from November to April, and the indoor set-point temperature was 20 °C in winter.

Table 6. Operation plan of the building.

Heating Hour	Occupants	Heating Duration	Set-Point Temperature
07–22	590 people	Winter season (Nov.–Apr.)	20 °C

2.2. Prediction Process

For the construction of the machine learning model for predicting the building thermal energy consumption, the steps shown in Figure 3 were taken.

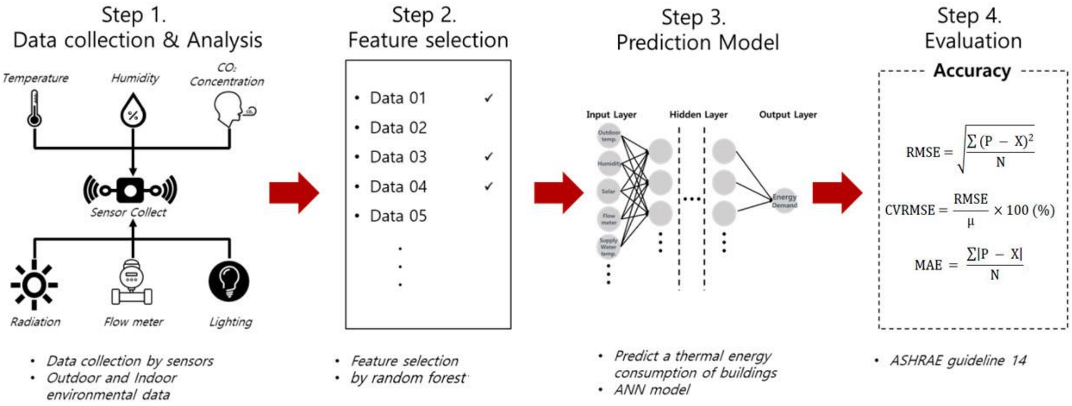


Figure 3. Process steps for constructing the machine learning model.

In Step 1, data were collected and analyzed. In this step, the characteristics and the environment of the building were analyzed using the collected data to create the optimal energy prediction model. In addition, similar types of data were grouped.

Step 2 is the feature selection step. In this step, unnecessary data or data that may act as noise were identified and removed so that only major variables were used. Random forest was used as a method for selection, and variables that significantly affect the thermal energy consumption to be predicted were extracted. In this instance, variables with high importance for the output value were considered as major variables.

In Step 3, a model for predicting the thermal energy consumption was constructed. For the prediction model, ANN was used. The model for predicting the building's thermal energy consumption was constructed using the variables selected in Step 2 as input data. The ANN model was constructed according to three cases. Case 1 was a model that utilized all data as input data without feature selection. Case 2 was a model that applied feature selection and used only major variables as input data. Case 3 was a model that added variables determined to be significant from an analysis of the major variables selected in Case 2.

In Step 4, the machine learning models were evaluated. In this step, the accuracy of the Case 1, Case 2, and Case 3 models was compared to derive an improved model for thermal energy consumption prediction

2.3. Feature Selection

Random forest is a data-driven method based on the basic properties of a decision tree. It is an ensemble learning methodology and relies on the combination of several decision trees via a voting scheme. The particularity of random forest is that their tree-based components are grown from a certain amount of randomness [35]. Based on this idea, random forest is defined as a generic principle of randomized ensembles of decision trees. Using a random selection of features to split each node is more robust with respect to noise [36]. The training procedure of a randomly generated forest can be summarized as follows [23]: first, build a bootstrap sample from the training dataset, second, grow a tree for each bootstrap sample and select the best split among a randomly selected subset of input variables, third, the tree is fully grown until no further splits are possible and repeat above procedure until all trees are grown. Random forest is a high-dimensional nonparametric method that works well on large numbers of variables [37]. It has been shown that the method is extremely accurate in a variety of applications [38].

Random forests can be used to rank the importance of each variable in a data analysis or prediction. Once random forest has created a lot of trees, the importance of variables that affect the output value is calculated. In general, the feature importance provided by random forest consisting of a large number of trees is more reliable than a simple decision tree method provided by one tree. In a random forest model, Gini importance is used as a measure for quantifying the importance of a feature [36]. Gini importance is derived from the Gini impurity value [39]. Gini importance is the averaged value of the total decrease in impurity over all individual decision trees in the random forest. As the value of Gini importance increases, the feature is considered more important. Gini importance is represented by a number between 0 and 1. A value closer to 1 means that a variable is more important among the data.

In this study, feature selection using random forest was used to improve the accuracy of the model and to reduce the run time. If there is extremely low or high Gini importance in its own group, variables are selected based on the 70th percentile as a major data, which have a significant effect on the prediction of energy consumption.

2.4. ANN

In this study, an ANN-based model was used to predict the thermal energy used in a building. ANN is a type of supervised learning developed by Warren S. McCulloch and Walter H. Pitt in 1943 [40]. ANN is constructed using the human central nervous system as a motif so that complex operations and calculations are possible. In particular, it has specialized features for the analysis and prediction of variables with nonlinear relationships. The ANN model for this study was constructed using the Python Scikit-learn library.

For the data used for machine learning, their ranges were matched, and preprocessing was performed so that they could be fairly reflected by the model. The normalized values ranged from zero to 1. In this instance, the equation used was as follows:

$$x_{\mathrm{normalization}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

where $x_{\max}$ means the maximum value of data, $x_{\min}$ is the minimum value of data.

The data was split into training data for the ANN model and test data for validation. Raw data were divided based on a 7:3 ratio; 70% was used as training data and the remaining 30% as test data. To compare each case on the same condition, the ANN model, in every case, has the same structure with the same hyper-parameter. The structure of the final ANN model is shown in Table 7.

Table 7. Artificial neural network (ANN) model structure.

Category	Parameter
Model	MLP (Multilayer Perceptron) Regressor
Activation function	ReLU
Learning rate	0.01
Momentum	0.4
Iterations	1000
The number of hidden layers	5
The number of hidden nodes	128

2.5. Evaluation

The coefficient of variation of the root mean square error (CVRMSE) and mean absolute error (MAE) were used to evaluate the prediction results using ANN. Both the CVRMSE and MAE verify the accuracy of a model by comparing the measured values with the predicted values. For both CVRMSE and MAE, the accuracy can be said to be higher if the result is closer to zero. According to the criteria specified in the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) guideline, the criterion for the hourly prediction is considered accurate and allowed if CVRMSE is less than 30% [41].

$$\mathrm{RMSE} = \sqrt{\frac{\sum (P - X)^2}{N}}, \tag{2}$$

$$\mathrm{CVRMSE} = \frac{\mathrm{RMSE}}{\mu} \times 100\ (\%), \tag{3}$$

$$\mathrm{MAE} = \frac{\sum |P - X|}{N}, \tag{4}$$

where P means the predicted value of ANN model, X is actual data, N is the number of actual data, μ is the mean of actual data.

3. Results

3.1. Overview of Data

Figure 4 shows histograms of the data used to construct the thermal energy consumption prediction model. The outdoor environment, indoor environment, central heating, and electricity energy data were collected from the building. The data were collected from 1 December 2017 to 30 April 2018. The data were collected on an hourly basis by the sensors, and a total of 3576 hours of data were collected.

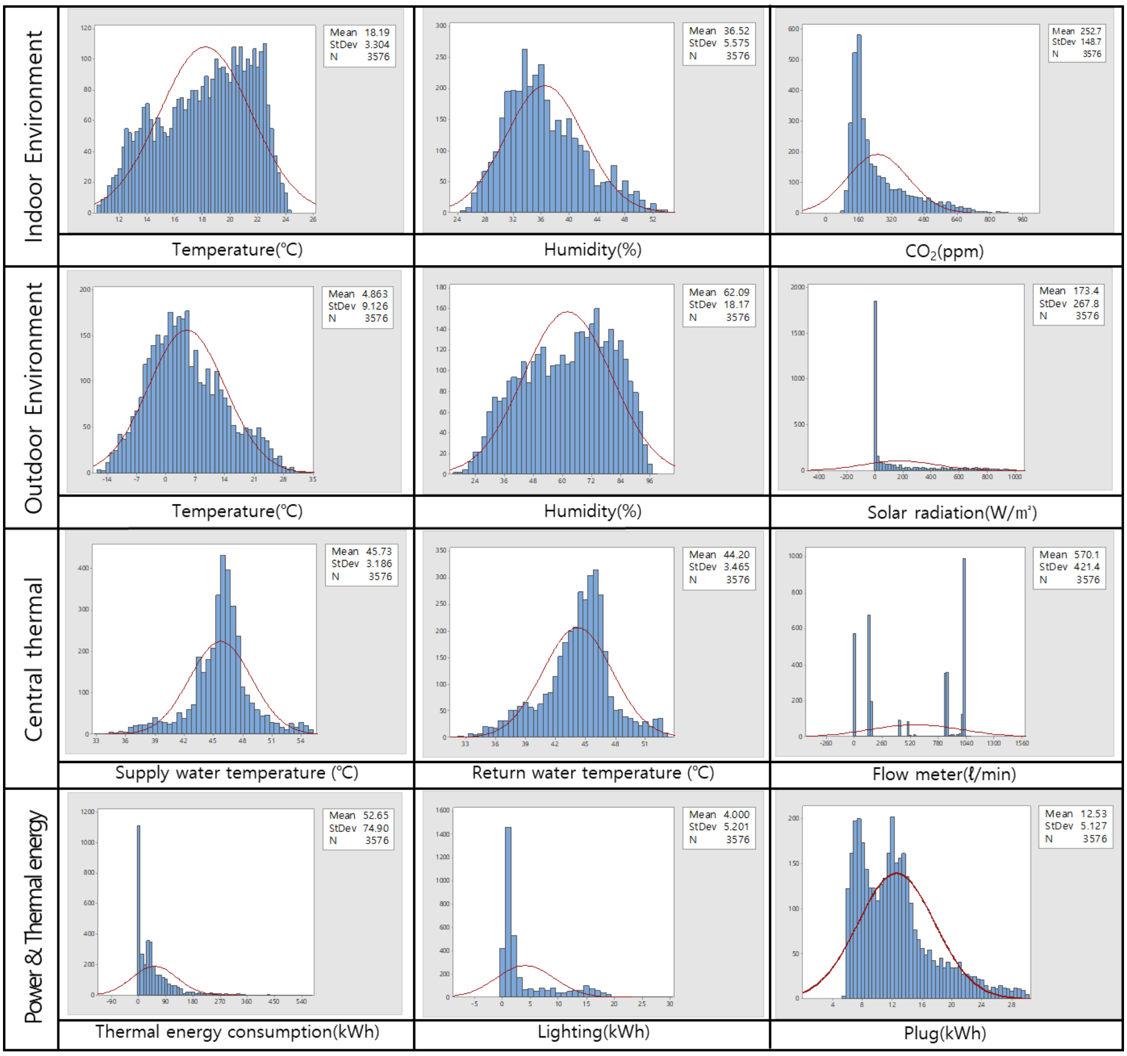


Figure 4. Distribution of measured data.

The indoor temperature was maintained at 16–17 °C when the rooms were not occupied and at nighttime, and it was maintained at 18–23 °C when the rooms were occupied and during the daytime. The indoor humidity was maintained at 36% on average, which was quite comfortable. The average indoor CO_2 concentration was approximately 252.7 ppm. From the data patterns, the CO_2 concentration varied in proportion to the number of occupants, indicating that the occupancy of a room can be estimated through the data.

The outdoor temperature was 4.8 °C on average. The hours of the day when the temperature was lower than 0 °C were fewer than those when it was higher than 0 °C, indicating that extremely cold weather was not experienced even in winter. The outdoor humidity ranged from 45% to 75%, fluctuating severely depending on the weather conditions, such as snow. The frequency of zero values for solar radiation was very high because the data were also collected during the night when the solar radiation cannot be measured.

The heat source supplied from the central machine room was transported to the subject building by the hot water supply. The supply water temperature was maintained at 45–48 °C, and the return water temperature was maintained at 43–46 °C. For the flow meter as a constant water volume supply system was adopted, and it was maintained at about 1000 ℓ/min when the rooms were occupied.

Thermal energy consumption was based on the energy used for heating and DHW in the building. When the rooms were not occupied, the value of the data was zero because no energy was consumed. For the lighting and plug electricity energy consumption data, a number of data points with low electricity energy consumption existed because standby electricity power was required during the non-occupancy period. During the occupancy period, however, constant electricity consumption was recorded.

3.2. Feature Selection by Random Forest

The purpose of feature selection is to eliminate the data, which is likely to interfere with the prediction. Random forest was used to extract major variables. The Gini importance of random forest is a coefficient for judging the importance of the influence on the output value. In the course of evaluating their importance, variables were compared with one another relatively. If the importance of certain variables was extremely high, the importance of other variables tends to be neglected in this group. Therefore, in this study, random forest was performed to avoid these issues by grouping raw data according to the characteristics of the variables. Raw data were divided into the following four groups: (i) outdoor environment data, (ii) indoor environment data, (iii) central heating supply data, and (iv) electricity energy consumption data. After grouping variables, distributions of results of Gini importance in each group should be checked. If they did not have an extremely low or high coefficient in their own group, they utilized all variables as input data of ANN. On the other hand, if extremely high or low Gini importance existed in their group, major variables were selected based on the 70th percentile in this study.

3.2.1. Outdoor Environment

Figure 5 shows the results of performing random forest for the outdoor environment data. The corresponding data were solar radiation, outdoor temperature, and outdoor humidity, which were representative variables for the outdoor environment. All of them exhibited a Gini importance of 0.3 or higher for the building thermal energy consumption, which was the output value, and it was judged that they evenly and significantly affected the output value. Therefore, for the outdoor environment data, solar radiation, outdoor temperature, and outdoor humidity were utilized as input data for the ANN model.

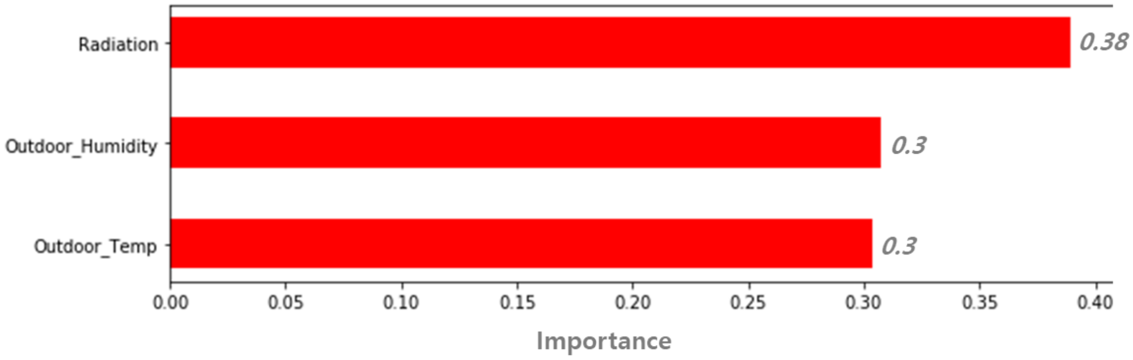


Figure 5. Gini importance of outdoor environmental data.

3.2.2. Indoor Environment

There were more variables for data related to the indoor environment compared to the other groups. The representative rooms with high occupancy density were selected from the building, and each room had data on the temperature, humidity, and CO_2 concentration. Three rooms in the chosen building had the same number of sensors to measure indoor conditions. Figure 6 shows the results of performing random forest for the indoor environment. Because there were several extremely low and high Gini importance in indoor conditions, major variables were selected based on the 70th percentile (The 70th percentile on the group of indoor environmental data is around 0.1, as shown in Figure 6). The variables judged to have significant impacts on the building thermal energy consumption were indoor temperature and CO_2 concentration of HS10 (classroom), and the indoor temperature of HS3 (office room). As shown in Table 3, both rooms had quite high occupancy densities and occupancy rates. Therefore, the importance of the variables was high in such rooms. HS8, which was not selected as a major variable, was a teachers' room. As teachers moved to classrooms for classes, the occupancy rate was reduced, and this appears to have lowered the importance of the variable. Finally, for the indoor environment data, the temperature and CO_2 concentration of the classroom and the temperature of the office were utilized as input data.

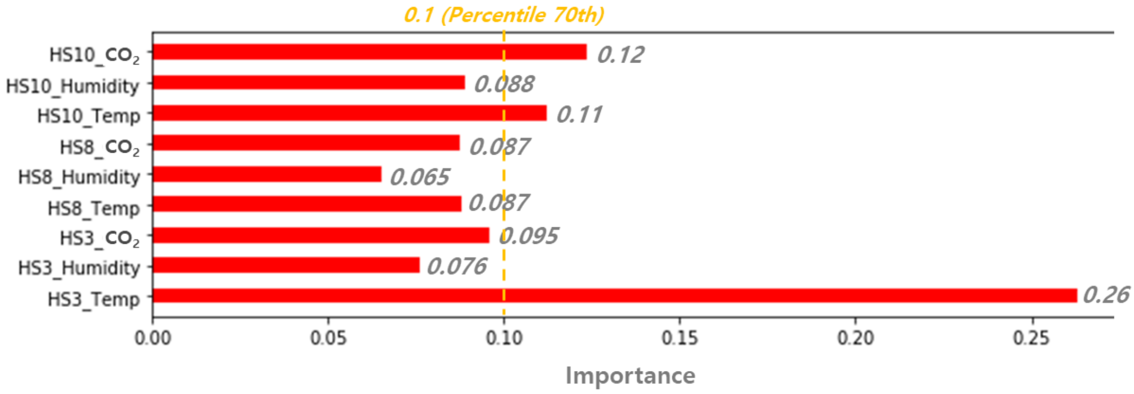


Figure 6. Gini importance of indoor environmental data.

3.2.3. Central heating Supply Data

The heating supply data included the temperature and flow rate when the central heat source was supplied to the building. Figure 7 shows the results of feature selection by random forest. As the Gini importance of all the variables exceeded 0.25, and there were no extreme coefficient values in this group, all of them were utilized as input data. In general, as the actual building starts operation, the supply flow rate increases, and changes in the return water temperature and the supply water temperature become larger. Owing to these correlations, the Gini importance results of random forest were quite high for all the variables.

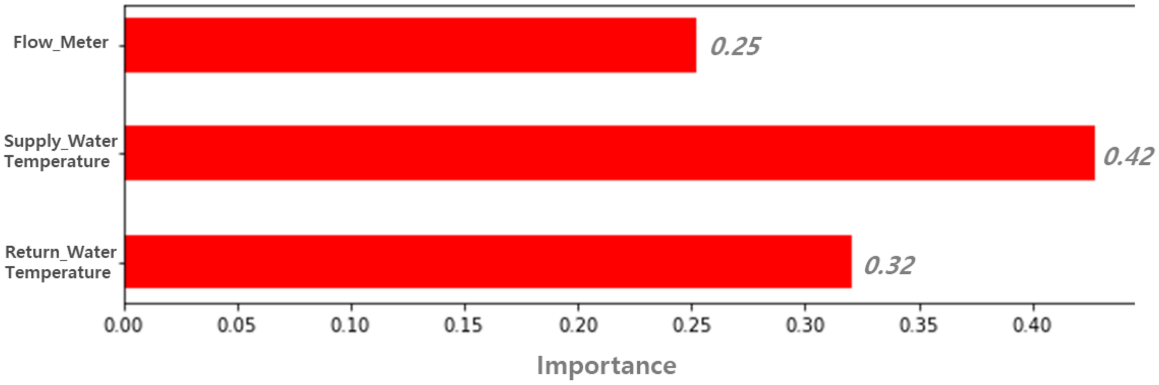


Figure 7. Gini importance of heating supply data.

3.2.4. Electricity Energy Consumption Data

The electricity energy consumption data were used to determine the occupancy status of the building. The results shown in Figure 8 were obtained because the use of equipment and lighting is closely related to the occupancy rate. It was found that both equipment and lighting significantly affected the building thermal energy consumption. Therefore, both variables were used as input data for the ANN model.

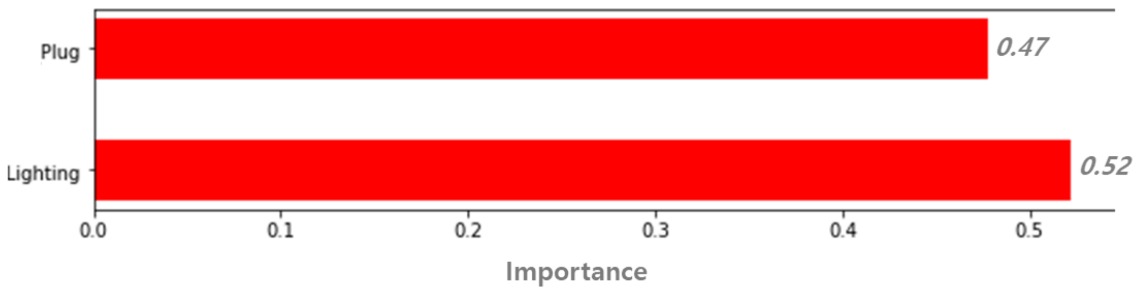


Figure 8. Gini importance of electricity energy consumption data.

As a result, 11 variables with high Gini importance out of 17 data were used as the input data of the ANN model, as shown in Figure 9.

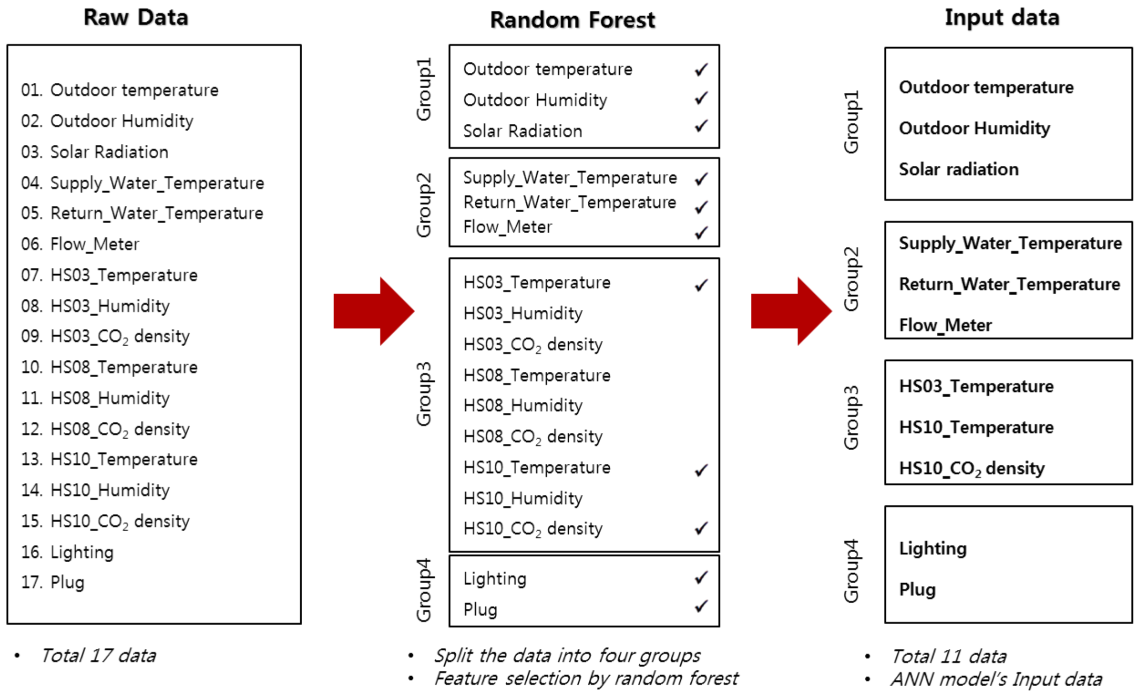


Figure 9. Results of feature selection by random forest.

3.3. Prediction of Buildings' Thermal Energy Consumption

3.3.1. Case 1

In Case 1, the ANN model was constructed using all sensor data as the input variables of the prediction model. The list of variables used to construct the ANN model contained 17 variables. The list was the same as the raw data list of Figure 9. When the prediction accuracy of the ANN model of Case 1 was evaluated, CVRMSE was approximately 40%, and MAE was approximately 11. As the prediction accuracy was not high, it was necessary to improve the model.

Figure 10 compares the empirical data collected from the sensors in the building with the predictive data of the ANN model for a week from March 16 to 22 in 2018. Although the ANN model of Case 1 predicted a similar building energy consumption pattern with that of empirical data, the prediction for each time period was not accurate. In particular, the prediction was found to be inaccurate in the early morning immediately before the occupants entered the building and at lunch time. In the morning, energy consumption rapidly increased as students began to enter the building. Conversely, the energy consumption was dramatically reduced as the occupants left the classrooms at lunch time. The prediction accuracy was low because the existing data combination could not reflect the rapidly changing situation, or it acted as noise when such rapid changes in the energy consumption occurred.

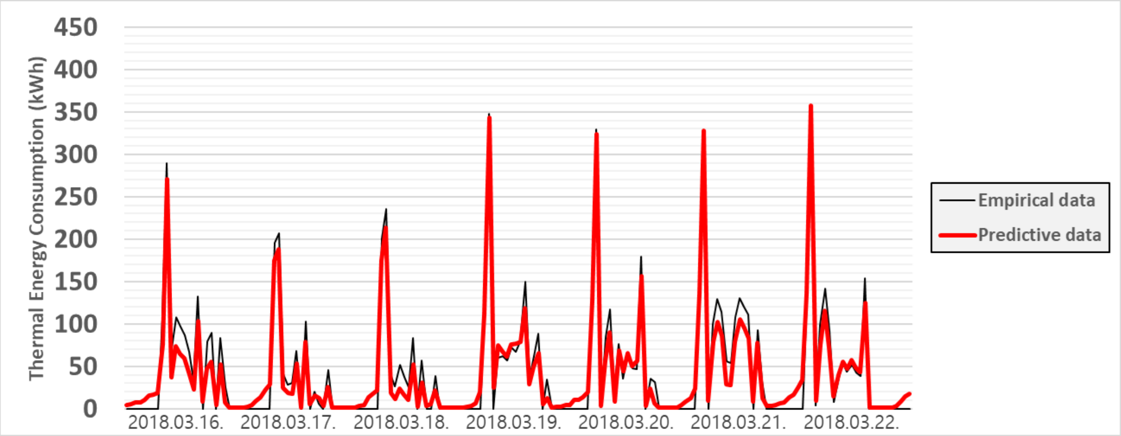


Figure 10. Case 1: Comparison between empirical and predictive data.

3.3.2. Case 2

In Case 2, major variables were selected by performing feature selection by random forest, and then an ANN model was constructed using the selected variables as input data. The variables used were the same as those in the input data list of Figure 9, and 11 variables were selected. For the prediction accuracy of Case 2, CVRMSE was approximately 35%, and MAE was approximately 10. The prediction results were improved by approximately 5% compared to Case 1. In particular, the accuracy in the time period when the occupants were in the rooms was significantly improved.

When the results of Case 2 were compared with those of Case 1, the prediction of the energy consumption exhibited more accurate results when a small number of variables extracted through feature selection were used than when all data were used. This indicates that using all the building data is not beneficial for prediction and that it is necessary to consider the combination of variables for prediction to secure excellent prediction accuracy. The prediction accuracy of Case 1 was lower even though all data were used because several data acted as noise that interfered with prediction. Therefore, it is important to increase the prediction accuracy of the model by finding the optimal variable combination.

However, as shown in Figure 11, the energy consumption prediction accuracy was still significantly low for the early morning. In particular, the model performed poorly for the time period immediately before people began to enter the rooms. This appears to be because there was no variable that can be used as a criterion for the occupancy status of people among the input data for training the ANN model.

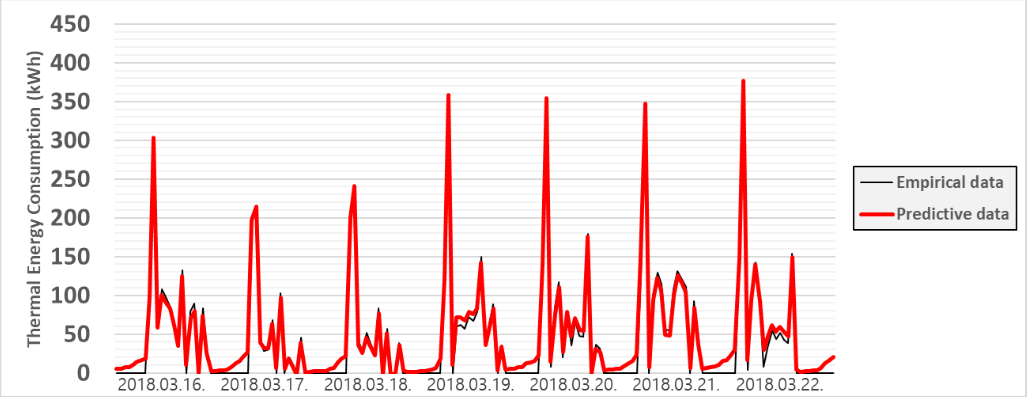


Figure 11. Case 2: Comparison between empirical and predictive data.

3.3.3. Case 3

Case 3 is an ANN model created by adding important variables to the input data used in Case 2. The chosen building was a school that had regular occupancy schedules during the weekend and the weekdays. For schools, the schedule involving school time, break time, lunch time, and home time is fixed, and the occupancy rate is also very predictable. Therefore, the energy consumption patterns of school buildings tend to be regular and clear. For this reason, the hour of the day data was used as a variable to reflect such regularity of the building schedule. As the raw data of the existing sensors could not be used as criteria for the occupancy hour, the energy prediction accuracy was significantly low in the interval where the energy consumption rapidly changed. To address this problem, the hour of the day data were utilized as input data.

In Case 3, the ANN model was constructed with 12 input data, adding the hour of the day data to the 11 variables used in Case 2. For Case 3, CVRMSE was approximately 25%, and MAE was 6.88, indicating the highest accuracy among all cases. The results of the cases confirmed that Case 3 exhibited a higher prediction accuracy than Case 1 and Case 2. It was also observed that the dispersed prediction values were tightening. In particular, as shown in Figure 12, the prediction accuracy of the pattern was significantly higher compared to Case 1 and Case 2. As the hour of the day data were used as input data, a variable for the criterion of the occupancy hour was generated, and the prediction accuracy during the occupancy period was significantly improved. Moreover, prediction accuracy was significantly improved for the morning and the lunch time when the energy consumption rapidly changed, making it possible to construct a model with high prediction accuracy.

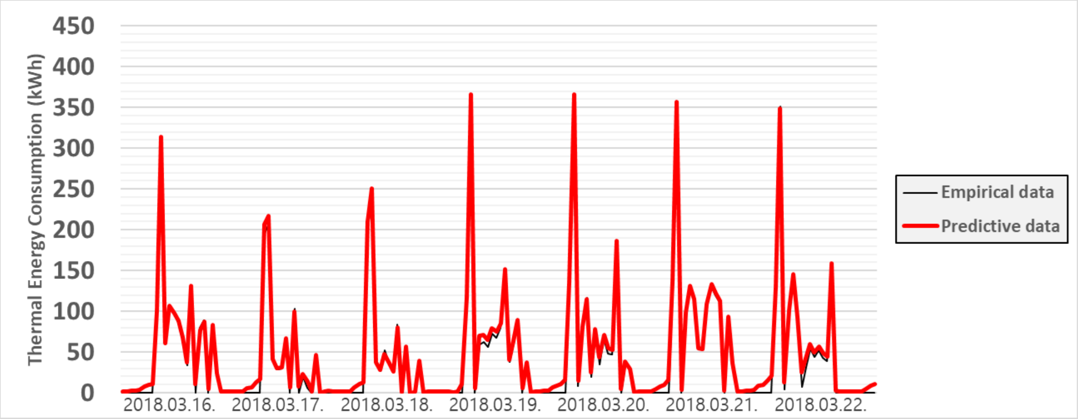


Figure 12. Case 3: Comparison between empirical and predictive data.

4. Discussion

In this study, models were constructed using combinations of various variables to increase the accuracy of the prediction model. Figure 13 compares the results of all cases using CVRMSE and MAE, which are accuracy evaluation indices. In particular, Table 8 illustrates the accuracy of each case with evaluation criterion based on the ASHRAE guideline 14. The results show that Case 2, which performed feature selection for input data, exhibited a higher prediction accuracy than Case 1 that used all data and that Case 3 that added a variable capable of having good influence on the output value of the prediction model showed higher prediction accuracy than Case 2. Therefore, an excellent prediction model for the thermal energy consumption of a building can be constructed by extracting major variables through feature selection and adding significant variables to the input data used for the model rather than by using all raw data as input data.

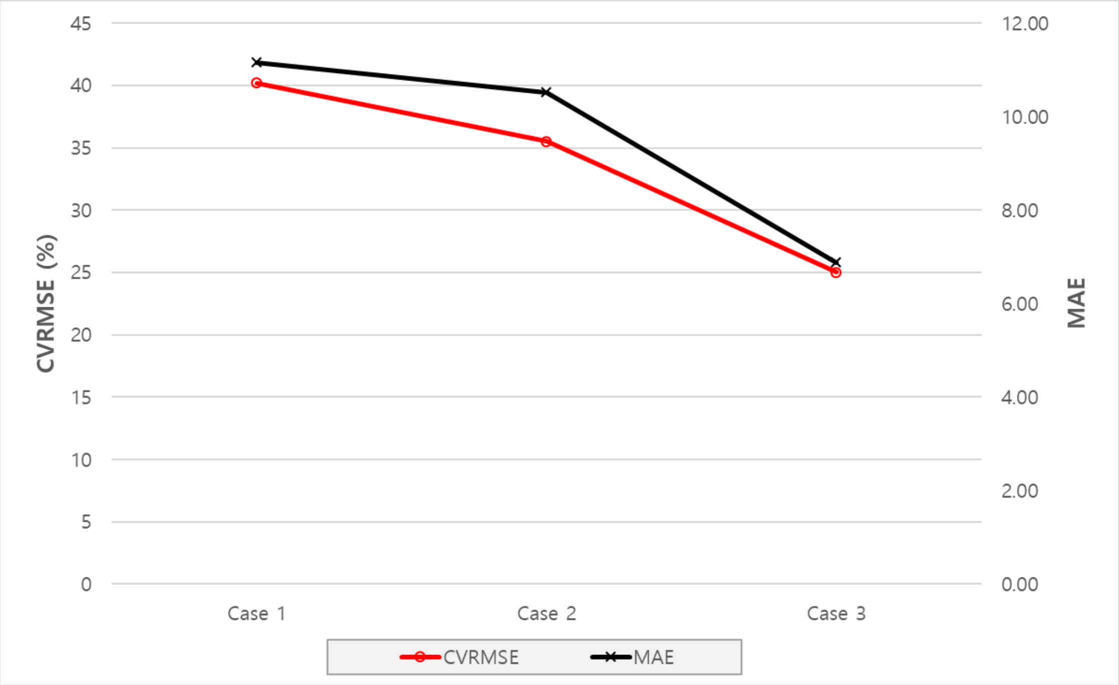


Figure 13. Comparison of prediction accuracy of all three cases.

Table 8. Comparison of prediction accuracy of all three cases with the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) Guideline 14.

Accuracy	ASHRAE Guideline 14	Case 1	Case 2	Case 3
CVRMSE	Criterion: Less than 30%	40.19%	35.52%	25.01%
MAE	-	11.16	10.52	6.88

As in Case 2 of this study, random forest was used to perform feature selection for extracting major variables. Random forest creates models with various combinations, and the importance of variables is determined in the process. Therefore, the importance of the input variables to the output value can be distinctly determined through the accuracy index. Moreover, the results of this study showed that such feature selection was actually quite helpful in predicting energy consumption and improving the prediction accuracy.

Moreover, the results of Case 3 show that the prediction accuracy can be significantly improved by adding significant variables based on data analysis rather than by using only the raw data collected from sensors in the building. If input data alone are not sufficient for predicting the output value, it is quite helpful to add variables that may bring better results when combined with the input data. In this study, a variable for hour of the day was added. As there was no variable to be used as a criterion for the occupancy status of the building among the existing sensor data, the hour of the day variable was used, which significantly improved the prediction accuracy of the model.

The accuracy of all cases is compared, as shown in Figure 14, and the R2 values of Case 1, Case 2, and Case 3 were 0.9492, 0.9779, and 0.9877, respectively. In particular, the resulting slope of Case 3 was close to 1, indicating a high prediction accuracy. Unlike the dispersed prediction results for Cases 1 and 2, the prediction results of Case 3 exhibited a high density, indicating that an excellent model was constructed. Moreover, Case 3 met the 25% CVRMSE criterion of ASHRAE, indicating that reliable results were drawn.

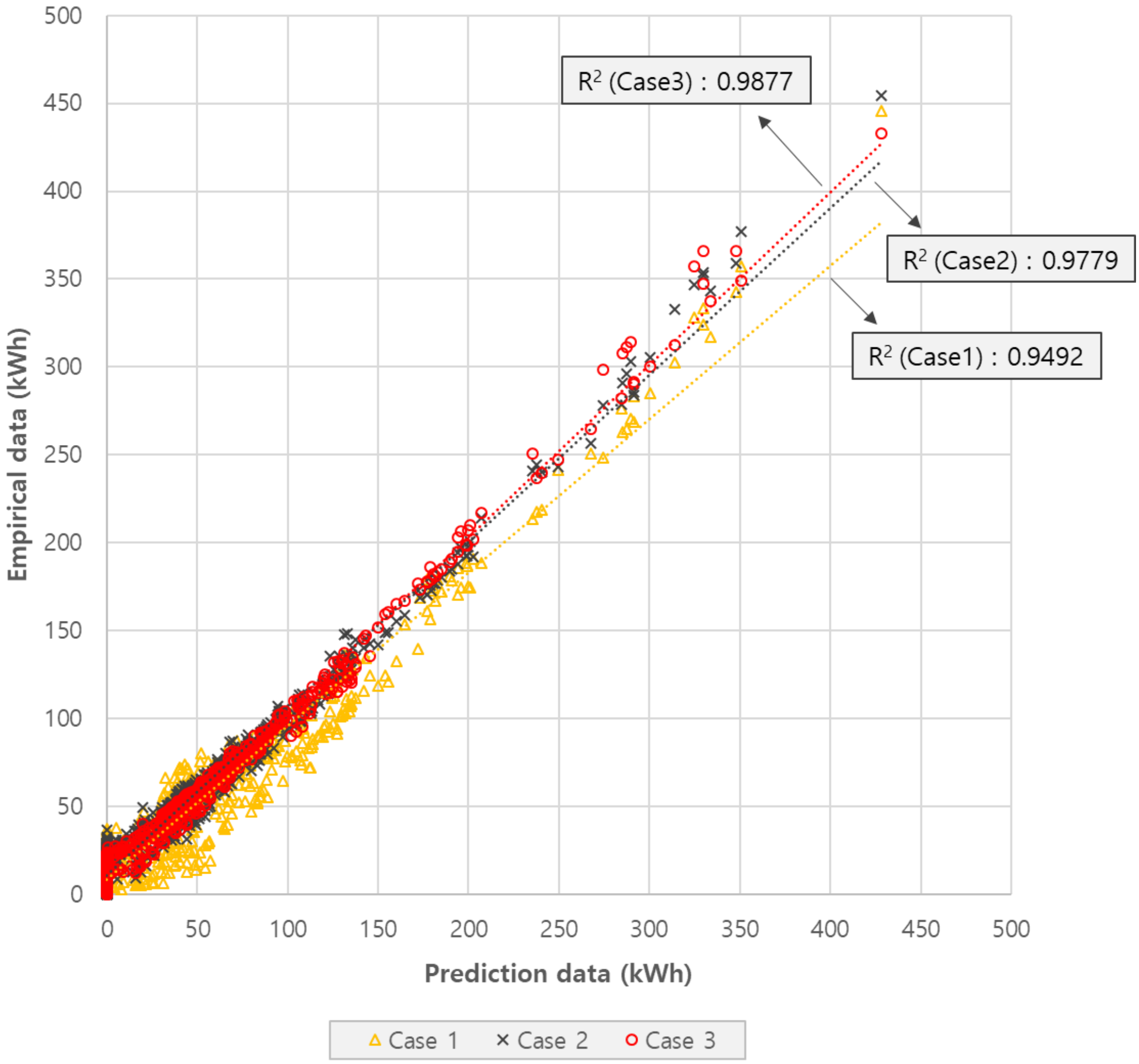


Figure 14. Comparison of the artificial neural network (ANN) model's prediction results.

5. Conclusions

In this study, a method for creating an optimal variable combination was used to construct a model for predicting the thermal energy consumption of a high school building. High-quality input data were created using a reduced data set through feature selection, and the prediction accuracy was improved by adding a significant variable to the input data combination. In this study, feature selection for extracting major variables was constructed using the Gini importance of random forest, and the prediction model for building thermal energy consumption was implemented using ANN. The accuracies of three prediction models were compared to create the optimal variable combination.

When a model for predicting the energy consumption of a building is to be implemented, the variable combination should be fully considered. Many current prediction models are constructed in the same direction as Case 1, which uses all the raw data available. In addition, when raw sensor data collected from an actual building are used, it is difficult to predict the energy consumption in many cases because they are empirical data. In this case, a fairly excellent prediction model can be obtained using the following steps: (i) extraction of major variables through appropriate feature selection and (ii) addition of significant input data that may have a good influence on the output value. When this method was used in this study, the Case 3 prediction model was found to be more accurate than the prediction model that used all raw data. If combinations producing a synergistic effect between input data are used, as in this study, it can be possible to implement a highly accurate model for predicting the thermal energy consumption of a building.

Compared with existing traditional methods, such as computer simulation and statistical method to predict energy consumption in buildings, it is not easy for those models to find a correlation between non-linear variables. And it is also difficult for machine learning using empirical data collected by a lot of sensors to predict with high accuracy because of the uncertainty of data. However, if the derivation strategy suggested by this research is applied, the ANN model can improve the quality of prediction considerably rather than before. Even though this research focused on the elimination of data, which act as a noise, an improved model with only a small number of variables could be achieved with high accuracy than other cases. In addition, if this strategy for the improved model is applied for actual buildings, it is possible to build economic monitoring systems efficiently by optimizing and installing sensors only where necessary.

In this study, however, only the data obtained during the short period of approximately four months from December to April were used, and the prediction model is limited to the corresponding period. Therefore, it is necessary to implement a prediction model for a longer period. Moreover, in Case 3, the accuracy was improved by adding only the hour of the day variable, but there are likely many variables that can improve the output prediction results. If further studies are conducted on the input variable combination that can improve the prediction accuracy, it will be possible to construct a model that predicts the thermal energy consumption of a building more effectively. In addition, other public buildings in town can be considered in the future works to make a more systemic feature selection according to the type of buildings and build more high-quality models while this study only focused on the high school building.

Author Contributions: All authors have participated in this research in the following contributions: conceptualization, J.J.; methodology, J.J., J.L., E.S. and K.P.; software, J.J., J.L. and E.S.; validation, E.S. and K.P.; formal analysis, J.J., J.L.; investigation, G.K.; data curation, K.P. and G.K.; writing—original draft preparation, J.J., J.L., G.K.; writing—review and editing, E.S., G.K. and J.H.L.; supervision, S.-B.L. and J.H.L.

Funding: This research was sponsored by the Korean Institute of Energy Technology Evaluation and Planning (KETEP) of the Republic of Korea (No. 2018201060010A).

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

References

1. National Oceanic & Atmospheric Administration (NOAA). The NOAA Annual Greenhouse Gas Index (AGGI). Available online: https://www.esrl.noaa.gov/gmd/aggi/aggi.html (accessed on 8 July 2019).
2. Kafle, S.; Parajuli, R.; Bhattarai, S.; Euh, S.H.; Kim, D.H. A review on energy systems and GHG emissions reduction plan and policy of the Republic of Korea: Past, present, and future. *Renew. Sustain. Energy Rev.* **2017**, *73*, 1123–1130. [CrossRef]
3. Brophy, V.; Lewis, J.O. *A Green Vitruvius: Principles and Practice of Sustainable Architectural Design*, 2nd ed.; Routledge: Abingdon, UK, 2011; p. 32.
4. IPCC. *Fifth Assessment Report*; IPCC: Geneva, Switzerland, 2014.
5. Wu, H.J.; Yuan, Z.W.; Zhang, L.; Bi, J. Life cycle energy consumption and CO_2 emission of an office building in China. *Int. J. Life Cycle Assess.* **2012**, *17*, 105–118. [CrossRef]
6. Torgal, F.P.; Mistretta, M.; Kaklauskas, A.; Granqvist, C.G.; Cabeza, L.F. *Nearly Zero Energy Building Refurbishment: A Multidisciplinary Approach*; Springer: Berlin, Germany, 2014.
7. Pedersen, L.; Stang, J.; Ulseth, R. Load prediction method for heat and electricity demand in buildings for the purpose of planning for mixed energy distribution systems. *Energy Build.* **2008**, *40*, 1124–1134. [CrossRef]
8. Ma, Y.; Borrelli, F.; Hencey, B.; Packard, A.; Bortoff, S. Model Predictive Control of thermal energy storage in building cooling systems. In Proceedings of the Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference, Shanghai, China, 15–18 December 2009.
9. Dincer, I.; Dost, S.; Li, X. Performance analyses of sensible heat storage systems for thermal applications. *Int. J. Energy Res.* **1997**, *21*, 1157–1171. [CrossRef]

10. Powell, K.M.; Sriprasad, A.; Cole, W.J.; Edgar, T.F. Heating, cooling, and electrical load forecasting for a large-scale district energy system. *Energy* **2014**, *74*, 877–885. [CrossRef]
11. Idowu, S.; Saguna, S.; Åhlund, C.; Schelén, O. Applied machine learning: Forecasting heat load in district heating system. *Energy Build.* **2016**, *133*, 478–488. [CrossRef]
12. Barak, S.; Sadegh, S.S. Forecasting energy consumption using ensemble ARIMA–ANFIS hybrid algorithm. *Electr. Power Energy Syst.* **2016**, *82*, 92–104. [CrossRef]
13. Bedi, J.; Toshniwal, D. Deep learning framework to forecast electricity demand. *Appl. Energy* **2019**, *238*, 1312–1326. [CrossRef]
14. Alfares, H.K.; Nazeeruddin, M. Electric load forecasting: Literature survey and classification of methods. *Int. J. Syst. Sci.* **2002**, *33*, 23–34. [CrossRef]
15. Neto, A.H.; Fiorelli, F.A.S. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. *Energy Build.* **2008**, *40*, 2169–2176. [CrossRef]
16. Tso, G.K.F.; Yau, K.K.W. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy* **2007**, *32*, 1761–1768. [CrossRef]
17. Hor, C.L.; Watson, S.J.; Majithia, S. Analyzing the impact of weather variables on monthly electricity demand. *IEEE Trans. Power Syst.* **2005**, *20*, 2078–2085. [CrossRef]
18. Daut, M.A.M.; Hassan, M.Y.; Abdullah, H.; Rahman, H.A.; Abdullah, M.P.; Hussin, F. Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review. *Renew. Sustain. Energy Rev.* **2017**, *70*, 1108–1118. [CrossRef]
19. Almonacid, F.; Rus, C.; Pérez-Higueras, P.; Hontoria, L. Calculation of the energy provided by a PV generator. Comparative study: Conventional methods vs. artificial neural networks. *Energy* **2011**, *36*, 375–384. [CrossRef]
20. Kialashaki, A.; Reisel, J.R. Development and validation of artificial neural network models of the energy demand in the industrial sector of the United States. *Energy* **2014**, *76*, 749–760. [CrossRef]
21. Ekonomou, L. Greek long-term energy consumption prediction using artificial neural networks. *Energy* **2010**, *35*, 512–517. [CrossRef]
22. Jovanović, R.Ž.; Sretenović, A.A.; Živković, B.D. Ensemble of various neural networks for prediction of heating energy consumption. *Energy Build.* **2015**, *94*, 189–199. [CrossRef]
23. 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]
24. Abedinia, O.; Amjady, N.; Zareipour, H. A New Feature Selection Technique for Load and Price Forecast of Electrical Power Systems. *IEEE Trans. Power Syst.* **2017**, *32*, 62–74. [CrossRef]
25. Jiang, P.; Liu, F.; Song, Y. A hybrid forecasting model based on date-framework strategy and improved feature selection technology for short-term load forecasting. *Energy* **2017**, *119*, 694–709. [CrossRef]
26. Liu, Y.; Wang, W.; Ghadimi, N. Electricity load forecasting by an improved forecast engine for building level consumers. *Energy* **2017**, *139*, 18–30. [CrossRef]
27. Zhao, H.-x.; Magoulès, F. Feature selection for support vector regression in the application of building energy prediction. In Proceedings of the 2011 IEEE 9th International Symposium on Applied Machine Intelligence and Informatics (SAMI), Smolenice, Slovakia, 27–29 January 2011; IEEE: Piscataway, NJ, USA.
28. Cho, S.; Lee, J.; Baek, J.; Kim, G.S.; Leigh, S.B. Investigating Primary Factors Affecting Electricity Consumption in Non-Residential Buildings Using a Data-Driven Approach. *Energies* **2019**, *12*, 4046. [CrossRef]
29. González-Vidal, A.; Jiménez, F.; Gómez-Skarmeta, A.F. A methodology for energy multivariate time series forecasting in smart buildings based on feature selection. *Energy Build.* **2019**, *196*, 71–82. [CrossRef]
30. Yi, X.; Liu, F.; Liu, J.; Jin, H. Building a network highway for big data: Architecture and challenges. *IEEE Netw.* **2014**, *28*, 5–13. [CrossRef]
31. Jang, J.; Baek, J.; Leigh, S.B. Prediction of optimum heating timing based on artificial neural network by utilizing BEMS data. *J. Build. Eng.* **2019**, *22*, 66–74. [CrossRef]
32. Kim, M.H.; Kim, D.; Heo, J.; Lee, D.W. Energy performance investigation of net plus energy town: Energy balance of the Jincheon Eco-Friendly energy town. *Renew. Energy* **2020**, *147*, 1784–1800. [CrossRef]
33. Kim, M.H.; Kim, D.; Heo, J.; Lee, D.W. Techno-economic analysis of hybrid renewable energy system with solar district heating for net zero energy community. *Energy* **2019**, *187*, 115916. [CrossRef]
34. Deb, C.; Zhang, F.; Yang, J.; Lee, S.E.; Shah, K.W. A review on time series forecasting techniques for building energy consumption. *Renew. Sustain. Energy Rev.* **2017**, *74*, 902–924. [CrossRef]

35. Nguyen, C.; Wang, Y.; Nguyen, H.N. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *J. Biomed. Sci. Eng.* **2013**, *6*, 551–560. [CrossRef]
36. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Zhang, Q.; Aires-de-Sousa, J. Random Forest Prediction of Mutagenicity from Empirical Physicochemical Descriptors. *J. Chem. Inf. Modeling* **2007**, *47*, 1–8. [CrossRef] [PubMed]
38. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958. [CrossRef] [PubMed]
39. Han, H.; Guo, X.; Yu, H. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. In Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS, Beijing, China, 24–26 November 2017.
40. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity (reprinted from bulletin of mathematical biophysics. *Bull. Math. Biol.* **1990**, *52*, 99–115. [CrossRef] [PubMed]
41. American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). ASHRAE guideline 14-2002 Measurement of Energy and Demand Saving: How to Determine What Was Really Saved by the Retrofit. In Proceedings of the Fifth International Conference for Enhanced Building Operations, Pittsburgh, Pennsylvania, 11–13 October 2005.