




Article

Extracting Influential Factors for Building Energy Consumption via Data Mining Approaches

Jihoon Jang ¹, Jinmog Han ², Min-Hwi Kim ³, Deuk-won Kim ³ and Seung-Bok Leigh ^{1,*}

¹ Department of Architectural Engineering, Yonsei University, Seodaemun-gu, Seoul 03722, Korea; eands777@yonsei.ac.kr

² Lyles School of Civil Engineering, Purdue University, 550 Stadium Mall Dr., West Lafayette, IN 47907, USA; han710@purdue.edu

³ Solar Thermal Convergence Laboratory, Korea Institute of Energy Research, 152 Gajeong-ro, Yuseong-gu, Daejeon 34129, Korea; mhkim001@kier.re.kr (M.-H.K.); dwkim15@kier.re.kr (D.-w.K.)

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

Abstract: To effectively analyze building energy, it is important to utilize the environmental data that influence building energy consumption. This study analyzed outdoor and indoor data collected from buildings to find out the conditions of rooms that had a significant effect on heating and cooling energy consumption. To examine the conditions of the rooms in each building, the energy consumption importance priority was derived using the Gini importance of the random forest algorithm on external and internal environmental data. The conditions that had a significant effect on energy consumption were analyzed to be: (i) conditions related to the building design—wall, floor, and window area ratio, the window-to-wall ratio (WWR), the window-to-floor area ratio (WFR), and the azimuth, and (ii) the internal conditions of the building—the illuminance, occupancy density, plug load, and frequency of room utilization. The room conditions derived through analysis were considered in each sample, and the final influential building energy consumption factors were derived by using them in a decision tree as being the WFR, window area ratio, floor area ratio, wall area ratio, and frequency of use. Furthermore, four room types were classified by combining the room conditions obtained from the key factor classifications derived in this study.

Keywords: heating and cooling energy consumption; influential factors; random forest; decision tree; data mining



Citation: Jang, J.; Han, J.; Kim, M.-H.; Kim, D.-w.; Leigh, S.-B. Extracting Influential Factors for Building Energy Consumption via Data Mining Approaches. *Energies* **2021**, *14*, 8505. <https://doi.org/10.3390/en14248505>

Academic Editor:
Dimitrios Katsaprakakis

Received: 9 November 2021
Accepted: 13 December 2021
Published: 16 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The importance of total energy management is evolving as the need for energy saving is being increasingly emphasized worldwide. In particular, building energy—which accounts for 20–40% of total energy consumption—is one of the items that requires active management [1]. Because the demand for building energy is continuously increasing, energy-efficient building design and building energy performance improvements are essential [2]. Recently, with the development of new technologies such as information and communication technology and big data, energy management systems (EMSs) that comprehensively manage energy have drawn attention. These systems efficiently manage energy consumption, maintaining a comfortable indoor environment by collecting and analyzing data in real time using sensors installed throughout the building [3]. Representative examples include home energy management systems and building energy management systems (BEMSs). Moreover, South Korea has announced a policy to obligate zero-energy buildings for all private and public buildings by 2030, zero-energy building certification being enforced as part of this policy [4]. To receive zero-energy building certification, buildings must satisfy various criteria and conditions, one of which is to install a BEMS [5]. Owing to such efforts, the application of EMSs to buildings is on the increase.

An EMS contributes to the formation of a comfortable indoor environment and the balanced optimization of energy consumption by analyzing building energy consumption using various data to make decisions regarding the operational conditions for the heating, ventilation, and air conditioning (HVAC) control system. Consequently, it is important to use data that are highly correlated with energy consumption to accurately analyze building energy. However, much of the sensor data recorded has included significant missing values and outliers [6], and building energy may be affected by various factors in combination. Hence, it is essential to reduce unnecessary data during the analytical process [7]. Unclear or low-quality data are undesirable in the analysis of building energy; thus, they are not useful for energy management. Moreover, when less important variables are removed, the comprehensibility, scalability, and accuracy of the analysis results can be improved [8]. Therefore, it is important to obtain the required data accurately to effectively analyze building energy consumption.

In actual buildings in operation, the behavior of occupants can appear in a variety of ways [9] and also affect building energy consumption dynamically. In particular, because unexpected situations can happen in non-residential buildings (e.g., significant increase or decrease of the number of occupants, energy consumption by night worker during the non-operating hours, etc.), it is necessary to consider real-time empirical data collected by sensors to figure out what affects building energy consumption. The data collected from actual buildings have generally nonlinear patterns because they depend on the HVAC system that operates dynamically. [10] Thus, it is important to analyze for feature importance in buildings to not only reveal the complex nonlinear relationship between variables, but also the most influential factors. [11] In order to reflect this characteristic, researchers have used various methods to analyze the data required for building energy analysis, a representative method being feature selection using machine learning. Among the building energy analysis methods, finding an appropriate solution by predicting building energy consumption has been used for some time [12]. The building energy prediction process requires various input data, and feature selection is one of the methods used to increase the prediction model accuracy when there are many input data candidates [13]. Zhao et al. [14] implemented a feature selection method to predict building energy consumption using a statistical machine learning approach. They found that among the five feature set cases, the smallest set (6) exhibited better prediction performance than the largest set (14). Zhang et al. [15] selected four models with different numbers of features using a data-driven model for feature selection procedure and compared the building energy prediction performance. Once again, the results showed that the model with a small number of variables (6) exhibited better prediction performance than the model with a large number of variables (24). Robinson et al. [16] analyzed building energy consumption and showed the top 10 most important features using gradient boosting regression models for higher quality. Liu et al. [17] 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. Sooyoun et al. [18] selected electric energy as a subset of total building energy consumption and identified the variables that contribute to electric energy use. 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. González-Vidal et al. [19] dealt with multivariate time-dependent series of data points for energy forecasting in smart buildings. They applied different types of feature selection methods to find out influential factors for regression tasks. The results of the experiments carried out show that the proposed methodology electively reduces both the complexity of the forecast model and their root mean square error and mean absolute error. Salah et al. [20] utilized wrapper, embedded feature extraction method and genetic algorithm to figure out best combinations of dataset for models. A LSTM model using only selected features captured all the characteristics of complex time series and showed high accuracy for medium to long range forecasting. Zeyu et al. [21] developed

building energy use prediction model with feature selection. Ensemble Bagging Trees (EBT) was utilized to build more compact input dataset which is formed by retaining important features and removing less important ones. X.J. Luo et al. [22] built an integrated artificial intelligence-based approach consisting of feature extraction to forecast week-ahead hourly. Mean absolute percentage error of the training and testing cases of the proposed predictive model is 2.87% and 6.12%, which has a 24.6% and 11.9% decrease compared to DNN model with a fixed architecture. Kusiak et al. [23] compared the accuracy of steam-load prediction through variable reduction utilizing data-driven approach. Using only two of the selected eight parameters yielded a higher prediction accuracy than using eight parameters. The authors found that fewer inputs tended to produce more stable prediction accuracy with decreased variance. To find the optimal prediction model variable combination, studies have been conducted that add new variables to the input variables. These studies improved the accuracy of building energy consumption predictions by adding meaningful variables, such as working days and hours of the day [24,25]. Ivanko et al. [26] tried to find appropriate influential variables to build an accurate heat use prediction model for domestic hot water (DHW). When “guest presence,” an artificial variable, was added to the existing historical DHW data, the prediction accuracy improved from 76% to 83%. The significant variables utilized in existing studies are summarized in Table 1.

Table 1. The significant variables utilized in previous researches.

Catagory	Variables	Reference
Weather conditons	Outdoor dry bulb temperature	[14,15,20–23,25,26]
	Outdoor wet bulb temperature	[15]
	Outdoor dew point temperature	[21,22]
	Outdoor humidity	[20,21,23,26]
	Solar irradiation	[15,21,26]
	Precipitation	[21]
	Wind speed	[20]
	Cloud cover	[22]
Indoor conditions	Indoor temperature	[14,15,25,26]
	Indoor setpoint temperature	[15]
	Infiltration and ventilation rate	[14,15]
Building conditions	Space Area	[16]
	Number of floors	
Time data	Time of day	[15,21,22,25]
	Month of day	[18,22]
	Type of day (Weekend, holiday, etc.)	[20–22,24]
Energy and Load	Past energy consumption	[15,18,24]
	Heating/cooling degree days	[16]
Building control	AHU control temperature	[15,26]
	AHU control ratio	[15,26]
Occupancy	Number of occupants	[14–16,21]

Consequently, selecting and using meaningful data had a positive and helpful effect on the analysis or prediction of building energy consumption, reducing the run time of the prediction model [13]. However, most previous studies utilized only sensor data to figure out influential factors, and there were few studies reflected the design characteristics of the building (building design and internal design conditions) together.

An EMS requires various outdoor and indoor environmental data, and should be considered with care in building energy analysis because it is closely related to building energy and resident comfort [27]. Indoor environmental data—that is, indoor dry-bulb temperature, relative humidity, air quality, etc.—were generally collected in real time using various sensors installed inside the building [28]. Outdoor environmental data—that is, outdoor dry-bulb temperature, relative humidity, solar radiation, etc.—were generally

collected through the Meteorological Administration or sensors outside the building [29]. Appropriate sensors must be installed to collect the data. In particular, the placement of environmental sensors can be an important issue as they affect the cost and detection function of the sensors [30]. The installation location of outdoor environmental sensors is somewhat more specific as there are fewer limitations to measurements based on the installation space. By contrast, indoor environmental sensors need to be installed by selecting representative locations that represent the condition of the building as the sensors cannot be installed throughout the entire building. Thus, it is necessary to determine the appropriate location for sensors according to the room conditions. Yoganathan et al. [31] attempted to find the optimal environmental sensing points that could provide the best overall indoor conditions of the target building. The authors provided the optimal number and locations of sensors through a data-driven approach and could reduce the number of sensors by 80% while minimizing the loss of information from them. Mousavi et al. [32] investigated the optimal number and location of sensors in the hospital. They focused on finding a space that could identify the indoor thermal conditions. The authors discovered that the optimal placement of sensors was highly sensitive to exterior factors. Suryanarayana et al. [33] attempted to determine the optimal placement of sensors for the best control and monitoring of a multi-zone building. They proposed an appropriate sensor placement for the target building by providing importance ranking to every sensor location through a data-driven methodology. Wagiman et al. [34] suggested an optimal lighting sensor placement method for indoor lighting control systems using a mathematical model. The authors focused on the process of saving electrical energy and improving visual comfort through a sensor placement method. The results showed that the proposed methodology fully satisfied the visual comfort of the room, achieving a 24.5% energy saving. Consequently, since it is difficult to find representative environmental conditions from buildings in operation, many other studies have focused on determining the optimal locations and number of data points that best represented the environmental conditions of a building.

The development of building energy analysis methods has enabled precise analysis using various types of building data. In particular, as more diverse building energy prediction methods have been used—from simulation methods to data mining—the diversity and accuracy of input data are being emphasized more. Active research has been conducted to find input data combinations based on which data are useful for energy analysis among the collected environmental data. In particular, it has become possible to find appropriate variables through a feature selection approach, enabling more effective analysis results to be derived. Moreover, deciding which data point sensors should be installed in a building to select the optimal data type is considered to be an important research topic. Previous studies have focused mostly on which types and which data points best represent the environmental condition of buildings. In addition, these feature selection methods used in the previous studies can effectively reduce large amounts of variables. However, since existing researches generally focus on the combination of variables that improve the accuracy of the model, it is difficult to identify the reason why such variables from buildings were selected. The results of such studies have difficulty in explaining the building and internal design conditions of rooms of high importance in building energy consumption. In particular, there is insufficient analysis on which characteristics of a room have an influence on heating and cooling energy, the main consumption sources of building energy. If it is possible to know why influential factors affect energy consumption of buildings according to design conditions, optimal location of sensors for energy consumption analysis can be determined efficiently. Consequently, this study aims to provide a guide to factors that should be prioritized in terms of building design and internal conditions. In order to achieve research purpose, this study investigates which characteristics of space have a significant effect on the heating and cooling energy consumption using outdoor and indoor environmental data collected from buildings. In order to figure out influential factors, both empirical sensor data and building design conditions were utilized to construct machine learning models.

The conditions of spaces that have a large effect on energy consumption are classified using quantitative values, enabling the selection of the characteristics of space that are important.

2. Methodology

2.1. Influential Factor Analysis Method

To investigate which characteristics of space in buildings have a significant effect on the heating and cooling energy consumption, this study was conducted in three steps, as shown in Figure 1. In step 1, indoor and outdoor environmental data were collected by installing sensors for five different types of buildings. For the indoor environment in particular, the temperature and humidity data were collected by installing sensors in each room in the building.

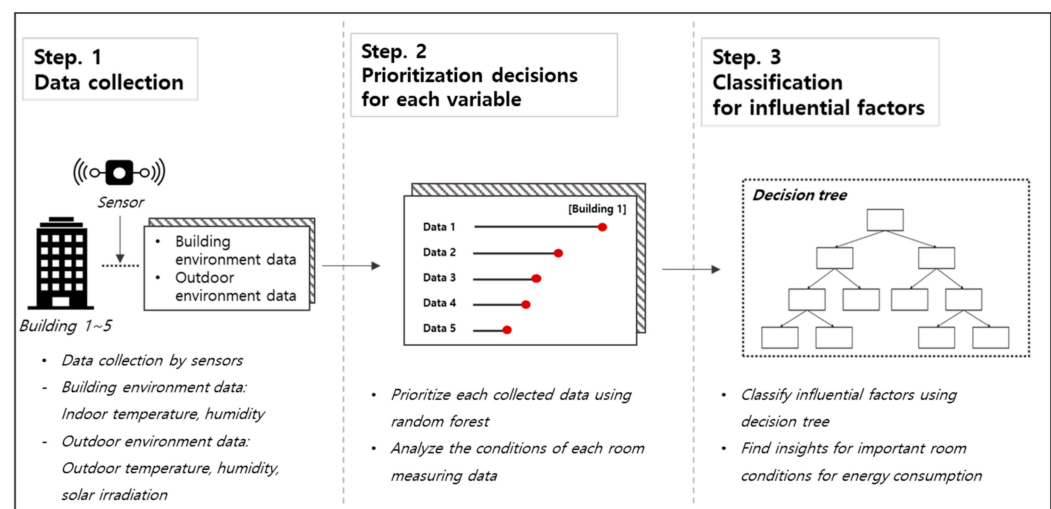


Figure 1. Flow chart of research methodology.

In step 2, the priority of importance of the data collected in step 1 was determined. To that end, the Gini importance of the random forest method was used, prioritizing and grouping the data based on the variable characteristics and building type. A list of utilized input data for each building is shown in Table 2. To analyze the differences based on the characteristics of each period, the heating and cooling seasons were analyzed separately. In this section, this study analyzed which conditions have effect on building energy consumption through sensor data with high importance.

In step 3, the conditions of the room with high importance were identified through classification in terms of building design and internal conditions by applying a decision tree using the conditions that had a significant effect on the heating and cooling energy—derived in step 2—as variables. In particular, the factors influencing energy consumption were derived through the classification of the room conditions.

2.2. Site Descriptions

In this study, data were collected from the buildings, as listed in Table 3, to determine the factors influencing heating and cooling energy consumption. The data collection periods were as follows: from November 2019 to September 2020 for a day care center, public health center, library, and high school buildings, and from November 2019 to March 2020 for a cultural center building. To perform analysis by period, the data for November to March in the data from November 2019 to September 2020 were used for the heating season, and the data from June to September were used for the cooling season. The target buildings were five publicly operated non-residential buildings characterized by regular operation based on similar operational schedules. All of the buildings were located in Jincheon, South Korea. The average temperature on site during the experiment period is about 15.9 °C (Max temperature: 40.3 °C, min temperature: −9.1 °C, standard deviation:

10.5 °C). Heating degree day (HDD) and cooling degree day (CDD) are about 2800 °C·days and 140 °C·days respectively.

Table 2. A list of input data for each case. (Temp. is indoor temperature and Hum. is indoor humidity).

Type	Culture Center	Daycare Center	Healthcare Center	Library	High School
Indoor environment	01. Office—Temp.	01. Room1—Temp.	01. Treatment room—Temp.	01. Reference room—Temp.	01. Hall—Temp.
	02. Office—Hum.	02. Room1—Hum.	02. Treatment room—Hum.	02. Reference room—Hum.	02. Hall—Hum.
	03. Hall—Temp.	03. Room2—Temp.	03. Waiting room—Temp.	03. Study room—Temp.	03. Kitchen—Temp.
	04. Hall—Hum.	04. Room2—Hum.	04. Waiting room—Hum.	04. Study room—Hum.	04. Kitchen—Hum.
	05. Auditorium—Temp.	05. Room3—Temp.	05. Multipurpose—Temp.	05. Hall—Temp.	05. Office1—Temp.
	06. Auditorium—Hum.	06. Room3—Hum.	06. Multipurpose—Hum.	06. Hall—Hum.	06. Office1—Hum.
	07. Lecture room—Temp.	07. Living room—Temp.	07. Chief’s room—Temp.	07. Office—Temp.	07. Class room1—Temp.
	08. Lecture room—Hum.	08. Living room—Hum.	08. Chief’s room—Hum.	08. Office—Hum.	08. Class room1—Hum.
			09. Office—Temp.		09. Class room2—Temp.
			10. Office—Hum.		10. Class room2—Hum.
			11. Lounge—Temp.		11. Cafeteria—Temp.
			12. Lounge—Hum.		12. Cafeteria—Hum.
			13. Hall—Temp.		13. Office2—Temp.
			14. Hall—Hum.		14. Office2—Hum.
Outdoor environment			01. Outdoor Temp.		
			02. Outdoor Hum.		
			03. Solar Irradiation		

Table 3. Building data.

Category	Types of Buildings				
	Culture Center	Daycare Center	Healthcare Center	Library	High School
Total floor area (m ²)	1456	572	542	1790	10,432
Building area (m ²)	728	572	271	1492	3872
Floor	Two-story building	One-story building	Two-story building	One-story building with one underground floor	Four-story building with one underground floor
Operating hours	08:00–22:00	09:00–20:00	08:00–22:00	08:00–22:00	08:00–17:00
Holidays	Sat., Sun.	Sat., Sun.	Sat., Sun.	Tue.	Sat., Sun.
HVAC			Fan Coil Unit (FCU)		

The conditions of rooms for each building are listed in Table 4. Among the criteria, room utilization was determined by field investigation, the other conditions being based on the design guide-lines for non-residential buildings [35,36].

Table 4. List of internal conditions for each room.

Building	Type	Illuminance (Lux)	Occupancy Density (m ² /Person)	Plug Load (W/m ²)	Frequency of Room Utilization
Culture center	1F Administration office	500	10	10	Always
	1F Hall	200	12	4	Intermittent
	2F Lecture room	300	3	4	Intermittent
	2F Hall	200	12	4	Intermittent
Daycare center	Nursery room (1-year-old)	300	4	4	Always
	Living room	300	3	2	Intermittent
	Nursery room (5-year-old)	300	4	4	Always
	Nursery room (4-year-old)	300	4	4	Always
Healthcare center	1F Treatment room (North)	500	14	7	Always
	1F Waiting room (Hall)	300	3	2	Always
	2F Lounge	300	3	2	Intermittent
	1F Multipurpose room	500	1	4	Intermittent
	2F_Hall	300	3	2	Intermittent
Library	2F Administration office	500	10	10	Intermittent
	1F Storage	100			Intermittent
	Reference room	500	2.5	0	Always
	Administration office	500	10	10	Always
	Study room	500	2.5	0	Always
High school	Hall	200	0.7	0	Intermittent
	1F Administration office	500	10	10	Always
	1F Hall	200	0.7	0	Intermittent
	1F Kitchen	300	10	300	Intermittent
	2F Classroom 2	300	3	4	Intermittent
	1F Cafeteria	200	1.2	2	Intermittent
	1F Teacher's office	500	10	10	Intermittent
2F Security office	500	14	7	Always	
	2F Classroom 1	300	3	4	Intermittent

There is a large solar thermal storage system in Jincheon that supplies heat to every building there. In winter, the solar heat stored is supplied for heating and DHW. If the supply is insufficient, a heat pump in the machine room is activated to supply heat. In summer, the solar heat is supplied only for DHW—that is, for cooling, a low-temperature fluid is produced through a heat pump and supplied to each building. The energy for heating and cooling in Jincheon is supplied in a manner similar to district heating [37]. Since it is difficult to install experimental equipment for radiant temperature and air velocity from buildings in real operation, basic indoor environment data (e.g., indoor temperature, humidity) were collected. Sensors are installed in the rooms of each building to check the basic indoor environmental data—temperature and humidity—and sensors to check the outdoor environmental data—temperature, humidity, and solar irradiation—are installed outside the building. Calorimeter sensors were installed to measure the heat for heating in the mechanical room of the building. This sensor measures supply and return temperature and flow meter data to calculate building energy consumption. Building environment data (Calorimeter data and indoor environment data) is measured every minute and outdoor environment data is measured every 30 s. Collected data was converted into an hour unit and utilized. These data are stored in the server sequentially over time. In this study, the data collected from 27 rooms in five buildings were used for analysis.

2.3. Random Forest

A large amount of data collected from an actual building have nonlinear patterns because they depend on the current state of the system that operates dynamically. If it is difficult to find out relationship between variables in building, data mining method is effective to solve this problems. Random forest is a data-driven learning algorithm for classification or regression to establish the meaning of data. It is a supervised learning method that randomly generates multiple decision trees in the learning process. In particular, because the variables used to separate each node are randomly selected, the random forest method is a complex model that has advantages when there are many types of data and datasets [13,38]. Ensemble or supervised learning refers to a method that trains multiple models instead of a single model, producing one output by using the prediction results of all of the models. The random forest method generates multiple models with slightly different characteristics by randomness, selecting the result that receives the largest number of votes as the final model through a voting scheme based on the results of each model. Through this process, one model of high accuracy is selected among the many models, the final selected model exhibiting assured generalization performance [39]. Furthermore, the learning process of the random forest algorithm is advantageous for variables that do not have a significant effect on the output or for variables that include noise—this is characterized by the ease of judging the importance of variables when there are many data [40,41]. When compared with traditional methods (e.g., correlation) that determine the association between variables, the random forest method has advantages in terms of the generalization of results because it draws results through a combination of various variables and the training of various models. When a model is configured as a random forest, the Gini impurity can be used to measure the impurity of each node, the Gini importance of which can be calculated using the Gini impurity [42], representing the impact of the input variables on the output value. The priorities of the variables can be listed using the Gini importance of the random forest, which is a quantified value that can be used to compare the relative importance of input variables [40]. The Gini importance is a value between 0 and 1, a value closer to 1 indicating the higher importance of the variable. Moreover, the input variables applied to the random forest method have their own Gini importance values, the sum of all values being 1.

Consequently, the Gini importance of the random forest algorithm is suitable for comparison between input variables and has been used as an indicator of the impact of indoor and outdoor environmental data on the building energy consumption. Constructed random forest model's hyper parameters are as follows: Package—Python scikit-learn RandomForestRegressor; the number of estimators—1000; the number of minimum sample split—2; the number of minimum sample leaf—1. The priorities in each parameter group are listed using the Gini importance, which is a calculated intermediate value. The parameters were grouped based on the data characteristics of each parameter and building type based on the operating time of the building. The input variables are as follows: (i) outdoor environmental data of each building for heating and cooling seasons; (ii) indoor environmental data of each building for heating and cooling seasons. Thus, the importance of each room was determined for each building, and the conditions of the rooms with high importance in terms of energy consumption were analyzed based on the results.

2.4. Decision Tree (J48)

A decision tree is a simple learning model that classifies or regresses data based on certain rules. It is a type of data mining technique that can discover meaningful patterns of a dataset using data mining methods, which may be difficult to determine using conventional statistical methods [43,44]. A decision tree is composed of a root node, intermediate node, and terminal node. Node branches are formed after analysis of the input data distribution. The criterion for separation from each node is determined by the impurity—that is, how much different data are mixed within the corresponding category. This is determined by the characteristics of the target parameters. If the data are discrete, it is determined by the

frequency of belonging to a corresponding category. If the target parameter is continuous, it is determined based on the average and standard deviation of the parameter [45].

This decision tree can draw information from the data and express it in a simple form. In general, the result of a decision tree is illustrated in a tree form, which has the advantage of easy analysis and understanding compared to other analytical models [44]. Decision trees are also good for making a judgment about important parameters because even if the number of types of handled parameters becomes large, less important parameters are removed and only parameters that classify data well remain when classification is performed. Moreover, it is easy to analyze and interpret importance because the criterion for classification can be determined through the pruning for each parameter [46].

The classification of a decision tree that has the advantage of discovering meaningful patterns was used in this study to identify influential factors that have a large impact on the building energy consumption among the room conditions. In this process, combinations of room conditions were used as input variables, and room importance derived by random forest was used as output variable. Because the number of sample datasets for room conditions including heaters and coolers was only 50, a simple model was used for classification instead of a complex model. Consequently, this study used the J48 decision tree of Weka, a data-driven tool developed by the University of Waikato in New Zealand, to derive influential factors [47]. Decision tree model's hyper parameters are as follows: Program—Weka 3.8; Confidence factor—0.25; the number of minimum sample split—2; the number of seeds—1.

3. Results

A random forest model was implemented using the collected data to analyze the factors influencing the heating and cooling energy consumption in buildings. In addition, using the Gini importance, the relative priorities of the outdoor and indoor environmental parameters were derived for each building.

In Section 3.1, the conditions are analyzed in terms of design that could influence the building through a priority analysis of the outdoor environmental data. In Section 3.2, the internal conditions of each room are analyzed based on the priorities of the indoor environmental data. In particular, when the indoor environment is analyzed, the sum of the Gini importance values of the parameters collected from each room is considered to be the representative room importance value of the corresponding room—that is, the sum of the Gini importance values for the temperature and humidity of the room is used as the importance of the corresponding room. In Section 3.3, the list of conditions that can influence the building analyzed in Sections 3.1 and 3.2, and the derived room condition samples are used in Section 4.

3.1. Prioritization for Outdoor Environmental Data

In this section, the priorities of the outdoor environmental data are analyzed. To that end, the Gini importance, an intermediate value of the random forest method (as explained in Section 2) was used as a quantified value for the importance of the sensor data, the relative importance of the data being compared to analyze the impact based on the heating and cooling seasons in each building.

The importance prioritization results using the Gini importance are shown in Figure 2, where the rows represent the five target buildings, and the columns are separated by heating and cooling results. It was found that in all five target buildings, solar irradiance had the greatest effect on building energy consumption regardless of the heating and cooling seasons. In buildings other than the school, outdoor temperature ranked second and outdoor relative humidity ranked third in the Gini-importance-derived priority rankings. When the totals were assessed, the Gini importance of solar irradiance showed very high average values of 0.53 and a standard deviation 0.07. This suggests that solar irradiance had a significant effect on the energy consumption patterns at the site of the buildings. This phenomenon appears to be due to the fact that because of the small size of each room

in the four buildings (culture center, daycare center, healthcare center, and library)—the high school excepted—the ratio of the perimeter zone exposed to the outdoor environment of the building was high, the façades of rooms being designed to be affected by solar irradiance because windows were installed in most of them. However, in the high school building, the Gini importance value of solar irradiance was slightly lower than that of other buildings, whereas the Gini importance of values for outdoor temperature and outdoor relative humidity were higher. In particular, the Gini importance of relative humidity was higher than that of temperature. This was because the school building had a larger area than the other buildings, and as a result, the ratio of the perimeter zone in the room that could be affected by solar irradiance was low and the number of occupants was very high (approximately 660 persons during the semester).

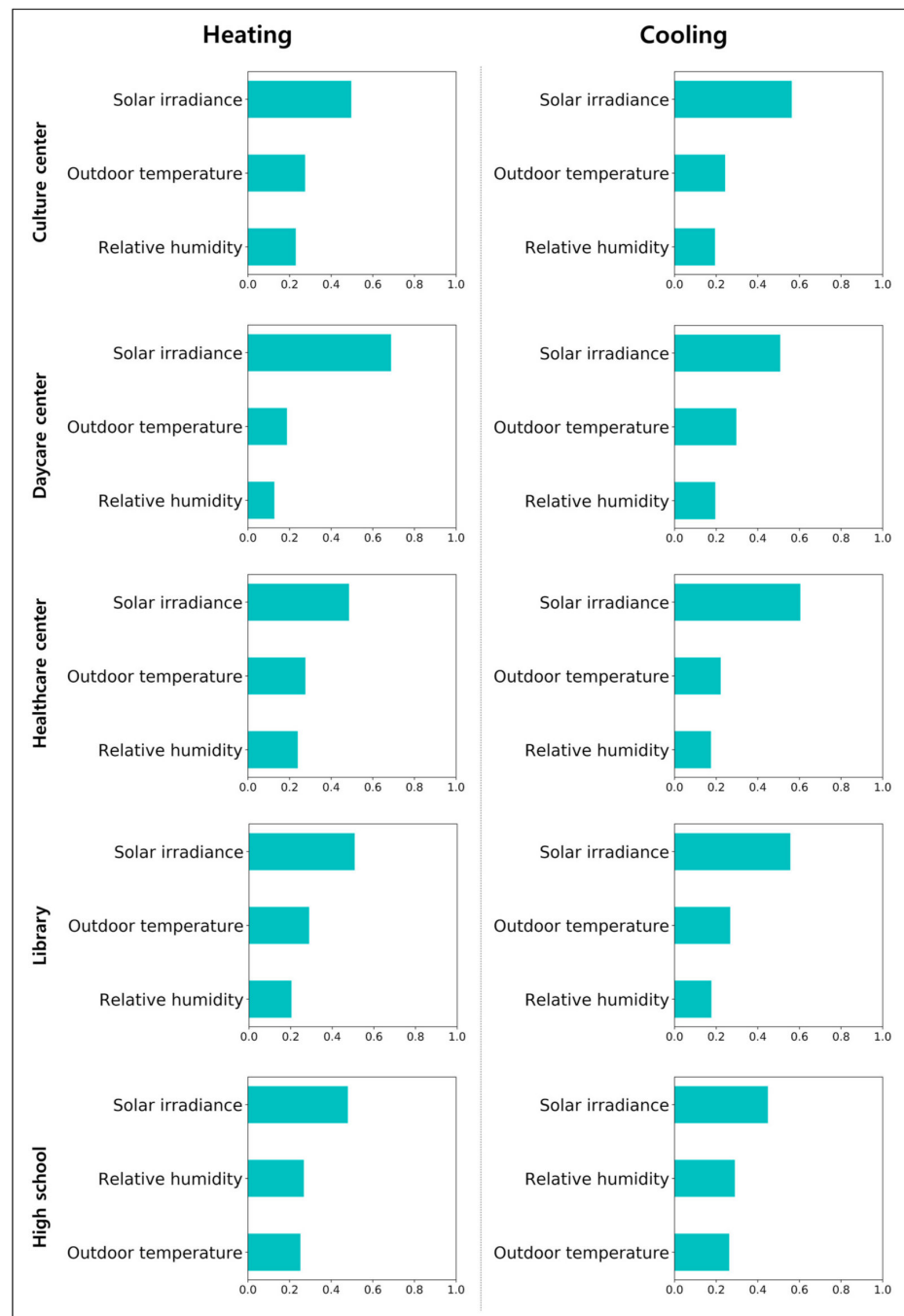


Figure 2. The results of prioritization using the Gini importance (outdoor environmental data).

3.2. Prioritization for Indoor Environmental Data

In this section, the priorities of the indoor environmental data collected from buildings are analyzed. As in Section 3.1, the relative importance values of rooms in the buildings were compared using the Gini importance. Moreover, the analysis of the indoor environment was performed separately for the heating and cooling seasons, the relative importance being compared. This study was conducted to identify the effect of design conditions of rooms on energy consumption rather than what kind of sensor data is important (Temperature, humidity, etc.). Thus, in order to evaluate the room's indoor parameters comprehensively, the sum of Gini importance values for temperature and humidity data collected from each room were used as the importance values for comparison of the indoor environmental conditions of rooms.

Because the impact of the building indoor environmental sensor data was determined by internal conditions, the analysis of the room importance focused on the analysis of internal conditions in Table 4, which were analyzed based on illuminance, occupancy density, plug load, and frequency of room utilization.

The prioritization results for the room indoor environment are shown in Figure 3. The rows represent the five target buildings, the columns being divided into heating and cooling results. For the culture center, only the heating season was analyzed as the data were collected only from November 2019 to March 2020. For the other buildings, both heating and cooling seasons were analyzed.

Firstly, the characteristic indoor environment pattern was the result of the culture center building during the heating season. In this building, the importance of the administration office was very high at 0.54, it being a room that was always occupied by administrative staff during the building's operating hours. The other rooms were intermittently used only when necessary. In particular, the occupancy density of the administration office was 10 m²/person, the room having the largest number of occupants in the building. Plug load and illuminance were also the highest among the rooms. Hence, it was the room that was most affected by occupants in the building.

The room that showed similar results was the administration office of the high school building, which served a similar purpose. This room was also affected by occupants as it was always occupied by administrative staff. The Gini importance was 0.32 during the heating season and 0.27 during the cooling season, exhibiting the highest importance value among all rooms in the building, regardless of the period. As with the office of the culture center, the office room of the high school also had the highest criteria for illuminance, occupancy density, and plug load among the rooms that were always used in the building.

In the case of the daycare center, most rooms were used to care for children, each room having similar internal conditions. In the daycare center during the cooling season, the importance of the room for one-year old infants was the highest at 0.34. This phenomenon seems to have appeared because the building environment was operated with a focus on this room to manage it more specifically than other rooms, infants being sensitive to the indoor environment.

In the healthcare center, the results differed between the heating and cooling seasons. In the heating season, the importance of the treatment room was 0.3, the 2F hall exhibiting a high importance at 0.35 in the cooling season. The treatment room had a smaller number of occupants than other rooms at 14 m²/person, but its utilization rate was very high as patient treatment was performed continuously during operating hours. Consequently, it was a room that was significantly affected by occupants. During the cooling season, the 2F hall was highly important. Because the 2F was closed in this building, the 2F reacted sensitively to internal heat gain compared to the 1F during the cooling season. The 2F hall was used intermittently, but it had a higher occupation density of 3 m²/person, the highest among the rooms on the 2F. This suggests the high importance of the hall, where the number of occupants and floating population were the highest among the rooms on the 2F. For this reason, the utilization rate was high during the cooling season, but the importance of the 2F hall was higher than that of the treatment room with a low number of

occupants. The importance of the administration office in the healthcare center was low as it was intermittently used, unlike other buildings. During the cooling season, the hall in the library building exhibited a much higher room importance (0.4) than the other rooms, as was the case in the healthcare center during the cooling season.

The library hall, located in the interior zone at the center of the building where there was no external effect, served as a passageway to each room and had a large floating population (occupation density 0.7 m²/person). Consequently, the importance of building cooling energy consumption was the highest because of the effect of internal heat gain due to the occupants in a closed space, even though the room was used intermittently. In the analysis process, the indoor environment room importance of the library in the heating season and the daycare center in the cooling season were excluded because the Gini importance values were not significantly different and it was difficult to derive characteristic patterns.

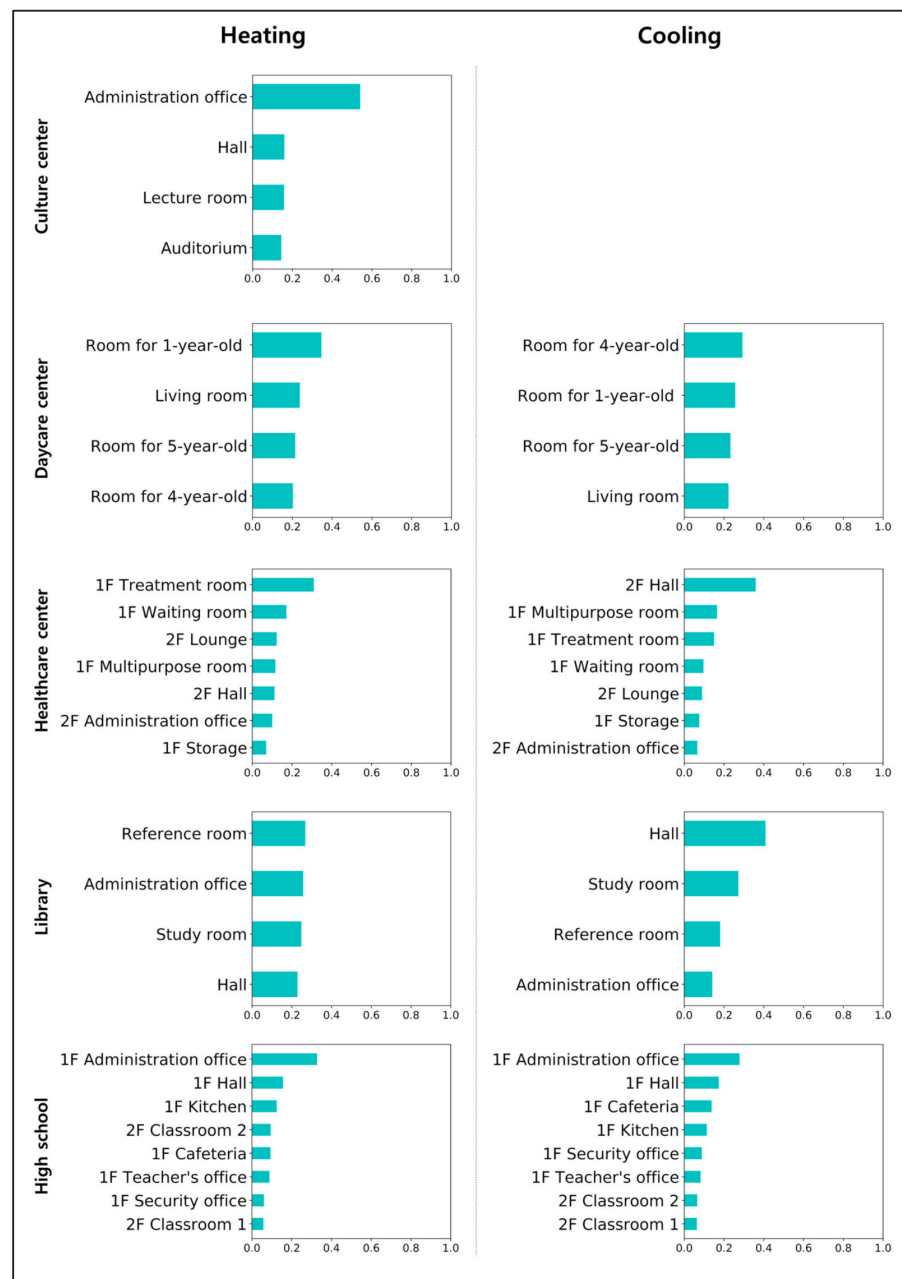


Figure 3. The results of prioritization using the Gini importance (indoor environmental data).

3.3. Summary

The analysis result of the Gini importance in Section 3.1 suggests that the outdoor environmental factor affecting energy consumption was determined by the design conditions of the building. It appears to have been greatly affected by the building size, which determines the perimeter and interior zones and exposure through the windows of the façade that can affect the solar irradiance. Furthermore, as described in Section 3.2, the indoor environmental factors influencing building energy consumption were determined by internal conditions and affected by the environmental conditions of the occupants.

Consequently, through the outdoor environmental analysis, this study determined the design factors of the building that affect the building energy consumption to be as follows: wall area ratio, floor area ratio, window area ratio, window-to-wall ratio (WWR), window-to-floor area ratio (WFR), and the azimuth. The area ratios (wall, floor, window) represent the ratio of the area of the target rooms in the building. Moreover, through the analysis of data from the building's indoor environmental sensors, the internal room conditions for occupants were selected as follows: illuminance, occupancy density, plug load, and frequency of room utilization. Through the analysis of Section 3, the room conditions that could influence the building energy consumption were selected, the basic statistical values of each room being outlined in Table 5.

Table 5. Statistical information of the building and internal design conditions for each room.

	Conditions	Min	Mean	Max	Standard Deviation
Building design conditions	Wall area ratio	0	0.18	0.45	0.12
	Floor area ratio	0.04	0.18	0.41	0.11
	Window area ratio	0	0.18	0.5	0.12
	WWR	0	0.33	0.84	0.2
	WFR	0	0.23	0.78	0.16
	Azimuth (°)	-	145	-	91
Internal conditions	Illuminance (lux)	100	352	500	125
	Occupancy density (m ² /person)	0.7	5.6	14	4.3
	Plug load (W/m ²)	0	16.6	300	59
	Frequency of room utilization		Always use: 11 rooms Intermittent use: 16 rooms		

3.4. Classification for Influential Factors

This study used various room conditions and the building energy consumption data of actual target buildings to investigate the correlations between the two datasets. Because room conditions directly affect heating and cooling energy consumption, grouping related conditions into a combination could help build an energy-efficient design and EMS. Consequently, this study used the decision tree (J48) algorithm as a data mining tool to examine the significant factors influencing building energy consumption among the room conditions. Subsequently, based on the classification results of the decision tree, a combination of conditions that exhibited high importance of energy consumption in terms of building design and internal conditions were presented.

As summarized in Table 5, the room conditions that had significant effects on heating and cooling energy consumption were used as the input variables for the decision tree, and the room importance of the building energy consumption as the output variable. To simplify the classification of the decision tree, the output variable was converted into two types: (1) high: the room importance value was larger than the average of the building and (2) low: the room importance value was smaller than the average of the building. The rooms classified as high or low meant that they had a high or low impact on energy consumption, and not that they had a high or low energy consumption. The confusion matrix of the classification decision tree and the overall reliability are summarized in Figure 4. Overall,

the accuracy of the decision tree classification was 94% (correctly classified instances: 47; incorrectly classified instances: 3).

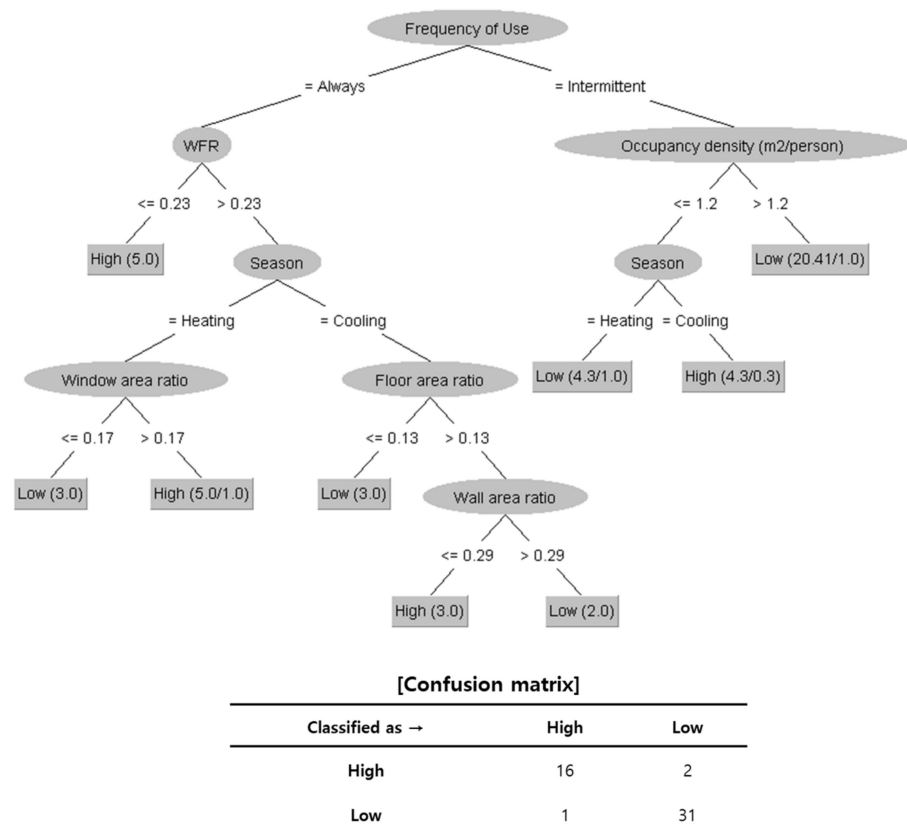


Figure 4. Decision tree based on the building and internal design conditions.

4. Discussion

In this section, based on the results presented in the preceding section, the characteristics of influential factors for building energy consumption are discussed.

Figure 4 shows the results of the decision tree based on the building and internal design conditions. As shown in Figure 4, a hierarchical approach was proposed on the basis of the importance of the conditions influencing the building energy consumption. The initial input variables of the decision tree were five among the internal design conditions and six among the building design conditions. However, the variables that were actually used in the classification were four among the building design conditions—that is, WFR, window area ratio, floor area ratio, and wall area ratio—and two internal design conditions—that is, frequency of use and season. This meant that the rooms of the target buildings could be classified using 6 variables instead of 11. There were four combinations of high importance, which were named A-1, A-2, A-3, and A-4, as listed in Table 6.

The A-1 type room is a combination of the frequency of use (always) and WFR (low). Here, the classification criterion of WFR was 0.23, which was the same as the average of all rooms (0.23). Based on this, the A-1 type room was interpreted to be a room that was always used, reacting sensitively to energy consumption regardless of cooling and heating if the WFR was smaller than the average—that is, the window area was smaller than the floor area. In this case, the possibility of an interior zone appearing is large. The room importance of building energy becomes large as the window area is relatively small, the exposure to the external environment diminishing, the tendency of the interior zone increasing, and the change rate of the internal conditions by occupants increasing.

Table 6. Classification based on the internal and building design conditions.

Type	Season	Internal Condition			Building Design Condition		
		Frequency of Use	Occupancy Density [m ² /Person]	WFR	Window Area Ratio	Floor Area Ratio	Wall Area Ratio
A-1		Always		Low			
A-2	Heating	Always		High	High		
A-3	Cooling	Always		High		Except for extremely low	Low
A-4	Cooling	Intermittent	Extremely low				

The A-2 type is a combination of season (heating), frequency of use (always), WFR (high), and window area ratio (high). Here, the classification criterion of the window area ratio was 0.17, which was an approximate value of the average of all rooms (0.18). Hence, the A-2 type room had a WFR and a window area ratio higher than the average. This meant that the exposure to the outdoor environment through the window was large as the window area relative to the room area was large. In other words, it was interpreted that the A-2 type room reacted sensitively to the outdoor environment because it was always used and had a large window area.

The A-3 type room is a combination of season (cooling), frequency of use (always), WFR (high), floor area ratio (except for extremely low), and wall area ratio (low). For the classification criterion of the floor area ratio, 0.13 was used. Based on the actual data, the cases classified as less than 0.13 had extremely low floor area ratios—such as 0.04, 0.08, and 0.1, which in Table 6 is expressed as ‘except for extremely low’ in the sense that rooms with a very small floor area ratio were excluded from consideration even in the target buildings. Based on the combination of the A-3 type room, it was interpreted that a room with these conditions was an ordinary room whose floor area ratio was not extremely low, and as the wall area ratio was small while the WFR was high, the wall area exposed to the outside was smaller than that of other rooms, but the window area ratio was high. As the window area was large, it was exposed to the outdoor environment in the perimeter zone, but the wall area exposed to the outside was smaller than in other rooms. As a result, the room had an interior zone inside the building. In other words, a room with the characteristics of the A-3 type room simultaneously exhibited the characteristics of a perimeter zone and an interior zone. This classification result was obtained as in the perimeter zone, the cooling energy became sensitive due to the heat gain by solar irradiance, whereas in the interior zone, the cooling energy became sensitive due to the effect of internal heat gain by illuminance, the occupants, and devices.

The A-4 type room is a combination of season (cooling), intermittent, and occupancy density (extremely low). The classification criterion for occupancy density, 1.2 m²/person, was much lower than the average of all rooms (5.6 m²/person). Based on this, the A-4 type room had a large effect on the cooling energy when it was used intermittently, and the number of occupants was extremely high. When the actual data were examined, the cases that belonged to this type of room were rooms with a large-area interior zone and a large number of occupants—for example, the hall of a building). This type of room had a significant effect on the cooling energy because although it was used intermittently, its internal heat gain by occupants being much higher than that of other rooms.

Most of existing researches for feature selection from a lot of building data focus on certain variables that improve the accuracy of the model. In addition, such studies just showed what type of sensor data (Temperature, humidity, power, etc.) has an important role for energy analysis model. In this cases, it is not easy to explain the reason why variables were selected and what kind of design conditions has an important effect on energy consumption of buildings. Thus, in this study, a method using data mining was conducted to find out design conditions that affects building energy consumption. The accuracy of classifying factors for energy consumption was considerably high at 94%. As a

result of classification for influential factors using decision tree, it was found that the room with the following conditions had a significant effect on energy consumption:

1. A room with a huge interior zone (For both heating and cooling season)
2. A room with a large window area (For both heating and cooling season)
3. A room with a large number of occupants (For cooling season)

Rooms that meet these conditions are likely to have a large impact on energy consumption. Therefore, environmental sensors installed in such rooms could well represent energy consumption of whole building rather than other rooms. In addition, if the design conditions resulting from the study are considered first, it will be possible to proceed more efficiently when analyzing energy consumption of buildings. The decision tree as a classification tool enabled the identification of various layers of room conditions related to building energy consumption. Although various types coexisted in the classification results of the decision tree, the meanings of each type could be interpreted consistently. The influence could be explained using only a few variables among the many input room conditions of the early decision tree because the combinations of data were meaningful. Furthermore, the factors that had a significant effect on building energy consumption could be compared with quantified values, and these values could be used as a guide for conditions that could be prioritized in terms of building design and internal conditions.

Consequently, it is crucial to clarify such classification results and classify room conditions based on their relevance and priority. Using this method, energy analysis could be performed more efficiently if the influential factors for building energy consumption are used in actual spaces, along with the results presented in this study.

However, this study had several limitations. Total number of buildings examined is a small size for generalization. Although this study focused on various 27 types of room conditions from five buildings in order to overcome these problem, there was a limit to expressing all kinds of room conditions. In particular, many rooms in the target buildings, except high schools, had shallow characteristics with many perimeter zones having a high exposure to the sun. Thus, it is difficult to conclude that various conditions were comprehensively considered. However, since the accuracy of the decision tree was excellent at 94% to classify room importance for energy consumption, it is judged to be worth applying to other types of room conditions. The second limitation is the type of indoor data used in this study. This study utilized basic indoor data from sensor, such as temperature and humidity. If more various types of sensors could be used to express indoor environment, such as CO₂, wind velocity, mean radiant temperature (MRT), etc., the result for representative room conditions might be more reliable.

5. Conclusions

This study investigates which characteristics of space have a significant effect on the heating and cooling energy consumption using outdoor and indoor environmental data collected from buildings. In order to figure out influential factors, both empirical sensor data and building design conditions were utilized to construct machine learning models. For analysis, the priorities of the data collected from buildings were derived using the Gini importance of the random forest method. In particular, room conditions could be compared using quantitative values for each room. As a result, the conditions that influence energy consumption were determined to be as follows: (i) conditions related to the building design—that is, the wall, floor, and window area ratio, WWR, WFR, and the azimuth; (ii) the internal conditions of buildings—that is, the illuminance, occupancy density, plug load, and frequency of room utilization.

Furthermore, the influential factors of rooms having a significant effect on building energy consumption were derived through a decision tree using the selected conditions, the accuracy of decision tree classification being 94%. The variables used in the classification of the decision tree were four building design conditions (WFR, window area ratio, floor area ratio, and wall area ratio) and two internal design conditions (frequency of use and season). This study derived four combinations of high room importance, and it was found

that in general, cooling had a high correlation with the combination of room conditions sensitive to internal conditions, and heating had a high correlation with the combination of room conditions that had a high exposure to the outdoor environment. As a result of classification for influential factors using decision tree, it was found that the room with the following conditions had a significant effect on energy consumption: A room with a huge interior zone (For both heating and cooling season); a room with a large window area (For both heating and cooling season); a room with a large number of occupants (For cooling season).

Finally, the major contributions of this study are as follows.

- Factors that have a significant effect on building energy consumption can be compared through quantified values;
- The methodology proposed in this study can identify rooms of high importance in building energy consumption and prevent indiscriminate installation of sensors;
- Efficiency can be improved as rooms of low importance can be excluded from consideration in the process of audit, analysis, and prediction of building energy consumption; and
- This study contributes to energy-efficient design and EMS construction as it provides a guide to conditions that should be prioritized in terms of exterior and internal building design.

In addition, this study had possible improvements to consider in future research. If more types of rooms and sensors are dealt to identify influential factors for building energy consumption, more reliable results could be obtained. Especially, if future research utilize a lot of cases, certain calculated values for branch from decision tree could be utilized for specific guidelines to make efficient monitoring systems.

Author Contributions: Conceptualization, J.J.; methodology, J.J.; software, J.J.; validation, J.J., J.H.; formal analysis, M.-H.K., D.-w.K.; investigation, J.H., M.-H.K., D.-w.K.; resources, M.-H.K., D.-w.K.; data curation, J.J., J.H.; writing—original draft preparation, J.J., J.H.; writing—review and editing, S.-B.L.; visualization, J.H.; supervision, S.-B.L.; project administration, S.-B.L. All authors have read and agreed to the published version of the manuscript.

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

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

References

1. Pérez-Lombard, L.; Ortiz, J.; Pout, C. A review on buildings energy consumption information. *Energy Build.* **2008**, *40*, 394–398. [CrossRef]
2. Yu, Z.; Haghghat, F.; Fung, B.C.M.; Yoshino, H. A decision tree method for building energy demand modeling. *Energy Build.* **2010**, *42*, 1637–1646. [CrossRef]
3. Roccotelli, M.; Rinaldi, A.; Fanti, M.P.; Iannone, F. Building Energy Management for Passive Cooling Based on Stochastic Occupants Behavior Evaluation. *Energies* **2020**, *14*, 138. [CrossRef]
4. Kim, Y.; Yu, K.H. Study on the certification policy of zero-energy buildings in Korea. *Sustainability* **2020**, *12*, 5172. [CrossRef]
5. Korea Energy Agency, Criteria of Zero Energy Building Certification. Available online: https://zeb.energy.or.kr/BC/BC03/BC03_05_002.do (accessed on 14 June 2021).
6. Xiao, F.; Fan, C. Data mining in building automation system for improving building operational performance. *Energy Build.* **2014**, *75*, 109–118. [CrossRef]
7. Zhao, H.-X.; Magoulés, 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 (SAMII), Smolenice, Slovakia, 27–29 January 2011; IEEE: Piscataway, NJ, USA, 2011.
8. Wang, J. *Data Mining: Opportunities and Challenges*; IGI Global: Hershey, PA, USA, 2003.
9. Nikdel, L.; Schay, A.E.S.; Hou, D.; Powers, S.E. Data-driven Occupancy Profiles for Apartment-style Student Housing. *Energy Build.* **2021**, *246*, 107205. [CrossRef]
10. Ruiz, L.; Cuéllar, M.; Calvo-Flores, M.; Jiménez, M. An Application of Non-Linear Autoregressive Neural Networks to Predict Energy Consumption in Public Buildings. *Energies* **2016**, *9*, 684. [CrossRef]

11. Pan, Y.; Zhang, L. Data-driven estimation of building energy consumption with multi-source heterogeneous data. *Appl. Energy* **2020**, *268*, 114965. [[CrossRef](#)]
12. Judkoff, R.; Wortman, D.; O'Doherty, B.; Burch, J. *A Methodology for Validating Building Energy Analysis Simulations*; Technical Report NREL/TP-550-42059; National Renewable Energy Laboratory: Colorado, CO, USA, 2008.
13. Jang, J.; Lee, J.; Son, E.; Park, K.; Kim, G.; Lee, J.H.; Leigh, S.B. Development of an improved model to predict building thermal energy consumption by utilizing feature selection. *Energies* **2019**, *12*, 4187. [[CrossRef](#)]
14. Zhao, H.X.; Magoulés, F. Feature selection for predicting building energy consumption based on statistical learning method. *J. Algorithms Comput. Technol.* **2012**, *6*, 59–77. [[CrossRef](#)]
15. Zhang, L.; Wen, J. A systematic feature selection procedure for short-term data-driven building energy forecasting model development. *Energy Build.* **2019**, *183*, 428–442. [[CrossRef](#)]
16. Robinson, C.; Dilkina, B.; Hubbs, J.; Zhang, W.; Guhathakurta, S.; Brown, M.A.; Pendyala, R.M. Machine learning approaches for estimating commercial building energy consumption. *Appl. Energy* **2017**, *208*, 889–904. [[CrossRef](#)]
17. Liu, Y.; Wang, W.; Ghadimi, N. Electricity load forecasting by an improved forecast engine for building level consumers. *Energy* **2017**, *139*, 18–30. [[CrossRef](#)]
18. 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](#)]
19. 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](#)]
20. Bouktif, S.; Fiaz, A.; Ouni, A.; Adel Serhani, M. Optimal Deep Learning LSTM Model for Electric Load Forecasting using Feature Selection and Genetic Algorithm: Comparison with Machine Learning Approaches. *Energies* **2018**, *11*, 1636. [[CrossRef](#)]
21. Wang, Z.; Wang, Y.; Srinivasan, R.S. A novel ensemble learning approach to support building energy use prediction. *Energy Build.* **2018**, *159*, 109–122. [[CrossRef](#)]
22. Luo, X.J.; Oyedele, L.O.; Ajayi, A.O.; Akinade, O.O.; Owolabi, H.A.; Ahmed, A. Feature extraction and genetic algorithm enhanced adaptive deep neural network for energy consumption prediction in buildings. *Renew. Sustain. Energy Rev.* **2020**, *131*, 109980. [[CrossRef](#)]
23. Kusiak, A.; Li, M.; Zhang, Z. A data-driven approach for steam load prediction in buildings. *Appl. Energy* **2010**, *87*, 925–933. [[CrossRef](#)]
24. Jurado, S.; Nebot, À.; Mugica, F.; Avellana, N. Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques. *Energy* **2015**, *86*, 276–291. [[CrossRef](#)]
25. 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](#)]
26. Ivanko, D.; Sørensen, Å.L.; Nord, N. Selecting the model and influencing variables for DHW heat use prediction in hotels in Norway. *Energy Build.* **2020**, *228*, 110441. [[CrossRef](#)]
27. Han, J.; Bae, J.; Jang, J.; Baek, J.; Leigh, S.B. The derivation of cooling set-point temperature in an HVAC system, considering mean radiant temperature. *Sustainability* **2019**, *11*, 5417. [[CrossRef](#)]
28. Ryu, S.H.; Moon, H.J. Development of an occupancy prediction model using indoor environmental data based on machine learning techniques. *Build. Environ.* **2016**, *107*, 1–9. [[CrossRef](#)]
29. Wennerström, H. Meteorological Impact and Transmission Errors in Outdoor Wireless Sensor Networks. Bachelor's Thesis, Uppsala University, Uppsala, Sweden, December 2013.
30. Sanders, D. Environmental sensors and networks of sensors. *Sens. Rev.* **2008**, *28*. [[CrossRef](#)]
31. Yoganathan, D.; Kondepudi, S.; Kalluri, B.; Manthapuri, S. Optimal sensor placement strategy for office buildings using clustering algorithms. *Energy Build.* **2018**, *158*, 1206–1225. [[CrossRef](#)]
32. Mousavi, E.; Khademi, A.; Taaffe, K. Optimal sensor placement in a hospital operating room. *IISE Trans. Healthc. Syst. Eng.* **2020**, *10*, 212–227. [[CrossRef](#)]
33. Suryanarayana, G.; Arroyo, J.; Helsen, L.; Lago, J. A data driven method for optimal sensor placement in multi-zone buildings. *Energy Build.* **2021**, *243*, 110956. [[CrossRef](#)]
34. Wagiman, K.R.; Abdullah, M.N.; Hassan, M.Y.; Radzi, N.H.M. A new optimal light sensor placement method of an indoor lighting control system for improving energy performance and visual comfort. *J. Build. Eng.* **2020**, *30*, 101295. [[CrossRef](#)]
35. American Society of Heating, Refrigerating and Air-Conditioning Engineers. Ventilation for acceptable indoor air quality. In *ASHRAE Standard 62.1*; American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE): Atlanta, GA, USA, 2019; pp. 17–22.
36. Din, V. 18599, *Energetische Bewertung von Gebäuden—Berechnung des Nutz-, End-und Primärenergiebedarfs für Heizung, Kühlung, Lüftung, Trinkwarmwasser und Beleuchtung*; Part 10: Boundary Conditions of Use, Climatic Data; Fraunhofer-Gesellschaft: Berlin, Germany, 2007; pp. 18–62.
37. Kim, M.; Kim, D.; Heo, J.; Lee, D. Energy performance investigation of net plus energy town: Energy balance of the Jincheon Eco-Friendly energy town. *Renew. Energy* **2020**, *147*, 1784–1800. [[CrossRef](#)]
38. 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](#)]

39. Xu, G.; Liu, M.; Jiang, Z.; Söffker, D.; Shen, W. Bearing fault diagnosis method based on deep convolutional neural network and random forest ensemble learning. *Sensors* **2019**, *19*, 1088. [[CrossRef](#)]
40. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
41. Menze, B.H.; Kelm, B.M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F.A. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinform.* **2009**, *16*, 213. [[CrossRef](#)] [[PubMed](#)]
42. 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.
43. Witten, I.H.; Frank, E.; Hall, M.A. *Chapter 1. Data Mining*, 3rd ed.; Elsevier: Amsterdam, The Netherlands, 2013; pp. 46–48.
44. Somvanshi, M.; Chavan, P.; Tambade, S.; Shinde, S.V. A review of machine learning techniques using decision tree and support vector machine. In Proceedings of the International Conference on Computing Communication Control and Automation, Pune, India, 12–13 August 2016.
45. Choi, J.; Moon, J. Impacts of human and spatial factors on user satisfaction in office environments. *Build. Environ.* **2017**, *114*, 23–35. [[CrossRef](#)]
46. Pal, M.; Mather, P.M. An assessment of the effectiveness of decision tree methods for land cover classification. *Remote. Sens. Environ.* **2003**, *86*, 554–565. [[CrossRef](#)]
47. Machine Learning Group, Machine Learning Project at the University of Waikato in New Zealand. Available online: <http://www.cs.waikato.ac.nz/ml/index.html> (accessed on 20 February 2016).