Fang, Haizhou; Tan, Hongwei; Dai, Ningfang; Liu, Zhaohui; Kosonen, Risto

Hourly Building Energy Consumption Prediction Using a Training Sample Selection Method Based on Key Feature Search

Published in:
Sustainability (Switzerland)

DOI:
10.3390/su15097458

Published: 01/05/2023

Document Version
Publisher's PDF, also known as Version of record

sustainability

MDPI

Article

Hourly Building Energy Consumption Prediction Using a Training Sample Selection Method Based on Key Feature Search

Haizhou Fang [1,2], Hongwei Tan [1,3,4,*], Ningfang Dai [1], Zhaohui Liu [1,5] and Risto Kosonen [2]

[1] School of Mechanical Engineering, Tongji University, Shanghai 201804, China
[2] School of Mechanical Engineering, Aalto University, 02150 Espoo, Finland
[3] Research Center of Green Building and New Energy, Tongji University, Shanghai 200092, China
[4] UNEP-Tongji Institute of Environment for Sustainable Development, Tongji University, Shanghai 200092, China
[5] Shanghai Construction Group Co., Ltd., Shanghai 200080, China
* Correspondence: hw_tan@tongji.edu.cn; Tel.: +86-18621941089

Abstract: For the management of building operations, hourly building energy consumption prediction (HBECP) is critical. Many factors, such as energy types, expected day intervals, and acquired feature types, significantly impact HBECP. However, the existing training sample selection methods, especially during transitional seasons, are unable to properly adapt to changes in operational conditions. The key feature search selection (KFSS) approach is proposed in this study. This technique ensures a quick response to changes in the parameters of the predicted day while enhancing the model's accuracy, stability, and generalization. The best training sample set is found dynamically based on the similarity between the feature on the projected day and the historical data, and feature scenario analysis is used to make the most of the acquired data features. The hourly actual data in two years are applied to a major office building in Zhuhai, China as a case study. The findings reveal that, as compared to the original methods, the KFSS method can track daily load well and considerably enhance prediction accuracy. The suggested training sample selection approach can enhance the accuracy of prediction days by 14.5% in spring and 4.9% in autumn, according to the results. The proposed feature search and feature extraction strategy are valuable for enhancing the robustness of data-driven models for HBECP.

Keywords: hourly prediction; training set selection; feature search; scenario analysis; machine learning

Citation: Fang, H.; Tan, H.; Dai, N.; Liu, Z.; Kosonen, R. Hourly Building Energy Consumption Prediction Using a Training Sample Selection Method Based on Key Feature Search. *Sustainability* **2023**, *15*, 7458. https://doi.org/10.3390/su15097458

Academic Editor: Lin Lu

Received: 22 February 2023
Revised: 3 April 2023
Accepted: 28 April 2023
Published: 1 May 2023

1. Introduction

In the wake of the world's rising carbon emissions, lowering building carbon emissions is becoming increasingly crucial [1]. The advancement of building energy management and autonomous operation control is a must tool for decreasing operational carbon emissions. Hourly Building energy consumption prediction(HBECP) is one of the most critical responsibilities in energy management and control [2]. White-box, black-box, and grey-box models make up the HBECP model [3,4]. The establishment of a white-box model is far more difficult, and the grey-box model has theoretical constraints [5]. Therefore, many previous researchers have selected the black-box model technique because of the huge quantity of building-related data obtained. Time-series methods [6], multiple regression methods, support vector machine [7,8], random forest [9], integrated methods [10,11], and deep learning methods [12,13] are all covered by the algorithms. In addition, to enhance prediction performance, a number of merger algorithms [14,15] and deep reinforcement learning techniques [16] have been created. With the use of Reinforcement Learning (RL) agents and Long Short Term Memory (LSTM) models, Zhou et al. [17] increased the predictive models' accuracy by up to 23.5%. Almost all the algorithms were mentioned by Zhang et al. [18]. In addition, he emphasized that research articles on algorithms are much more prevalent than those on data.

There are studies on data in addition to the numerous studies on algorithms. Particularly in recent years, scholars have focused more on the study of the data itself. There are now three categories of study data sources: data created by physical models, data generated by machine learning, and actual energy consumption data from buildings. Amasyali et al. [19], for instance, used EnergyPlus to produce a sizable amount of simulated data for machine learning to research the effect of occupant behavior on office building energy consumption. Oh, et al. [20] models that include simulation findings as input into data-driven models enhance learning performance in a data-scarce environment. Kim et al. [21] gathered the target data set through physical simulation in various weather scenarios to enhance the model's capacity for generalization. Tian et al. [22] combined the hybrid data created by Generative Adversarial Networks (GAN) with the original data to create data that was comparable to the original data in order to enhance the prediction model. Several researchers have used transfer learning approaches to get better practical value [23]. Fan et al. [24,25] evaluated the use of transfer learning technology in energy consumption prediction in the absence of building data, and the findings demonstrate that transfer learning has significant application value. However, this situation only applies to new buildings that lack historical operation data, and it does not apply to older buildings that have a massive portion of historical operation data.

The bulk of researchers does studies based on historical data using data mining techniques. Pattern recognition technologies were often used to classify operating data into numerous categories, and data with the same pattern is treated as a class of training samples. Using the k-means clustering method, for example, operational data from the literature [26,27] was used to classify energy consumption patterns into three categories. Acquah et al. [28] produced strong predictive accuracy by classifying energy consumption patterns into three groups using K-means clustering with dynamic time wrapping. Chen et al. [29] used fuzzy C-means (FCM) clustering to identify the usage patterns of energy consumption. Jallal et al. [30] used FCM as a component to establish energy consumption models. Piscitelli et al. [31] established a non-intrusive way to effectively distinguishing different building functions using classification algorithms. These researchers' clustering approaches were often used in the creation of HBECP as a data-centric strategy. Clustering approaches, on the other hand, will reduce training samples for each class in the case of insufficient samples, potentially resulting in poor model adaptability. Another group of researchers [32–34] employed time series features to partition the energy consumption data into periodic and random components, which they then used as training sample sets for modeling. As well, the accuracy of the model was improved by separating periodic and random signals. Empirical mode decomposition with adaptive noise was utilized by Zhou et al. [35] to partition the data and enhance the model's performance. Peng et al. [36] decomposed the data using an empirical wavelet transform and then independently modeled the decomposed data. The findings indicated that this strategy can increase the model's accuracy. Obviously, the selection of training samples using such a data decomposition method was simply dependent on mathematical principles, and while it does have some effect on enhancing the prediction accuracy, the total effect was quite small.

There are differences in the selection of data types, as well as differences in the sample size of training samples. The selection of features is equally important in the development of the model. According to the literature [37–39], past research primarily used historical data on energy consumption and outdoor weather. When Qiao et al. [40] successfully enhanced the energy consumption model by conducting feature selection research using time, weather, and historical energy consumption, they also made the observation that occupant behavior data will need to be taken into account in the future. Occupants had received increased attention in recent years, as they directly reflect the Usage of the building [41–43]. Liu et al. [44] pointed out that due to occupant behavior, there is a prediction deviation of peak and basic energy consumption. Shao et al. [45] pointed out that the sudden increase in passenger flow during holidays and other reasons would cause problems that the model could not handle in time for hotels, thus occupant

behavior data should be considered in future predictions. Ahmad et al. [46] proposed to use of a Gaussian kernel regression model with random feature expansion to improve accuracy, and they concluded that it was necessary to introduce occupant behavior to improve the model quality. Kadir et al. [19] used different combinations of occupant behaviors to simulate the impact of building energy consumption and pointed out that the complexity of the real world should be considered to obtain actual occupant behavior data. Das et al. [47] established an improved LSTM model to improve short-term prediction accuracy, considering device-utilization patterns and occupant interactions with these devices. It can be seen that the acquisition and selection of features also determine the quality of the model, especially occupancy behavior and outdoor weather data.

Researchers have conducted extensive work to improve prediction model generalization and interpretability. The preceding literature indicates that research on algorithms covers conventional algorithms to deep learning or combination application. The main focus of data-focused research was clustering, decomposition, or combination utilization. Also, a lot of literature [40,48] discusses the value of occupancy patterns and climatic variables as predictive input features, although these feature data are rarely accounted for when identifying training samples. All the same, because they have a large influence on building energy consumption, these two feature parameters should receive adequate attention when choosing training samples. The accuracy of the HBECP model is wildly disproportionate to the number of training samples when compared to other domains (e.g., image recognition and speech recognition). Instead of only using these two parameters as features to train the model, occupancy behavior and climatic parameters must be thought about during the data selection stage before creating a machine learning model. Additionally, when running prediction algorithms, the precise time point of the prediction object is frequently considered. As a result, in order to provide a more suitable training sample that meets the need, the requirements for specific scenarios must be considered in the initial data selection stage. Scientists have constantly argued that the greatest difficulties in the current modeling process are how to choose HBECP training samples and how to improve model interpretability. However, prior studies have not discussed how to choose training samples while taking occupancy behavior and climatic conditions into attention. The technical methods of cluster and decomposition emerged since previous studies largely focused on how to construct a model based on the collected data. The time and space situational factors that need to be predicted in the future are not thought about. This work analyses two major feature parameters and prediction periods in the training sample selection step in order to address the two issues of stability and interpretability. As a result, more trustworthy training samples are obtained, and the model's stability is improved. Furthermore, the model's prediction outcomes are explained using two distinctive features and the prediction period to increase the model's interpretative capacity.

In order to fill this gap, In this paper, a key feature search selection(KFSS) method is proposed. The AC utilization rate is recalculated by getting the start-stop state of all fan coil units in the building, which is reasonably simple to obtain and may be used to depict the building's occupancy behavior. As well, one of the elements impacting building energy consumption is the outdoor temperature. As a result, both are chosen as search features. The weather data can be predicted by meteorological predictions, and the AC utilization rate data on the predicted day can be predicted or artificial settings. Indoor usage pattern information and outdoor environmental information are represented by the two features. The timeframe takes the cooling season, transition season, and heating season into care. The earliest work concentrated on the cooling and heating seasons, while the training sample selection in this work is intended for the above three periods, particularly the transition season. This is because there are various operating modes for buildings during transitional seasons, the law of energy consumption is complex, and prediction is harder.

The following is a summary of this paper's significant innovations:

(1) Ingeniously incorporate the essential distinctive parameters for future energy consumption situations into the model training sample shortlisting. This makes it possible

to quickly acquire data and create precise predictive models. Similarly, the model's interpretability is improved.

(2) The KFSS technique fixes the clustering method's issue with insufficient training sets while also allowing for changes on the predicted day (e.g., HBECP problem under the transition of the AC switch-on mode in the transitional season). This effort includes the model's predicting abilities throughout the year, particularly during the transition season, rather than simply during the cooling and heating seasons.

(3) Mixing feature scene analysis and actual feature acquisition can increase not only the feasibility of actual engineering modeling, but also provide a leading direction for the next step in improving model capabilities. The proof existence can offer a fresh concept for mining building data that will become available in enormous amounts in the future and that can be swiftly implemented on energy platforms, offering a reliable assurance for HBECP.

This paper is organized as follows. Section 2 presents the research outline, the dataset of two-step selection, and the feature set scenario. Section 3 presents the case study of the building, including dataset selection in detail and feature selection. Section 4 analyzes and discusses the performances of the prediction models. Section 5 discusses the application and conclusions are drawn in Section 6.

2. Methodology

2.1. Framework

In this research, we propose a strategy for selecting training samples that incorporates occupancy behavior features for HBECP. As shown in Figure 1, the majority of the paper's innovations are clustered in a dotted box. The energy consumption modeling procedure is divided into five parts. The following is a detailed description of each step in the workflow:

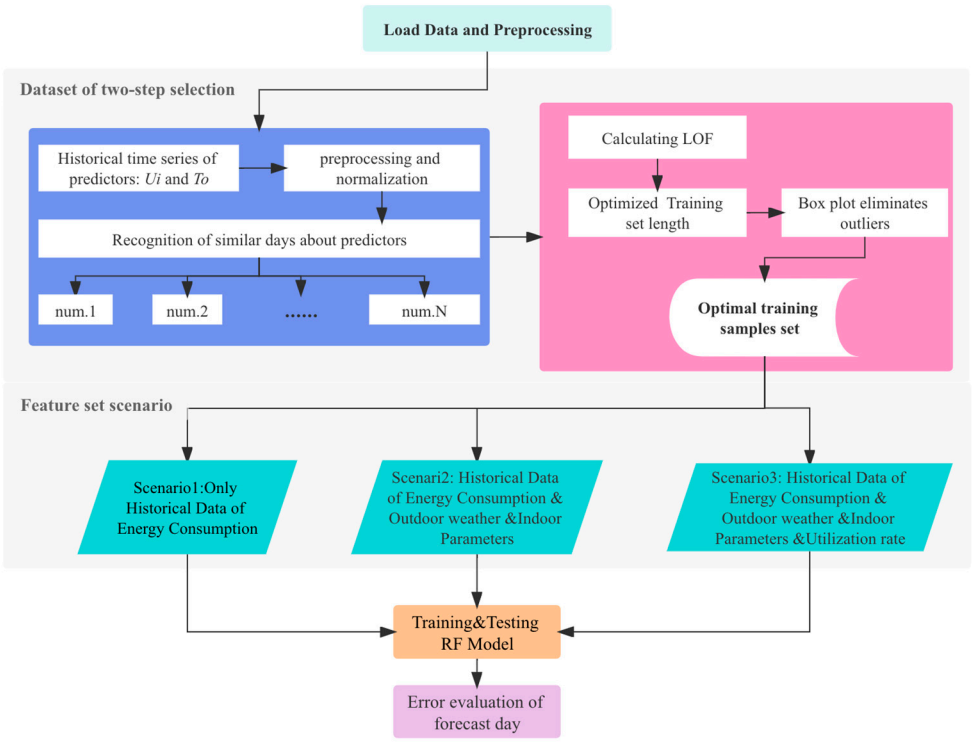


Figure 1. The research framework.

Step 1. Data loading and preprocessing:

Before being fed into the model, the raw data is pre-processed. First, raw data from several data sources are combined (e.g., historical energy consumption data, meteorological data, and utilization rate). Second, outliers in the data are cleansed and classified based on

day type. Finally, the data is standardized to the interval [0, 1] to prevent bigger or smaller data ranges from dominating model parameters.

Step 2. Dataset of two-step selection:

There are discrepancies between prediction goal time periods during dataset selection for model training (e.g., cooling season, transition season, and heating season). First, the data should be sorted by neighboring points based on historical data of the key features in the predicted day, with the distance calculation using Euclidean distance. The training dataset is then further filtered using the local outlier factor (LOF) technique, which removes training samples with greater outliers and uses the remaining samples as the final model input.

Step 3. Feature set scenario:

First, the periodicity and correlation calculations are used to extract the features. Then, the features are grouped into three situations based on the difficulty of getting them: (1) historical energy consumption data only, (2) historical energy consumption data and outdoor weather parameters, (3) historical energy consumption data, outdoor weather parameters, and fan coil utilization rate.

Step 4. RF model training and testing:

Here, partition the training and test sets and apply 10-fold cross-validation. Grid search is then performed to find the best hyperparameters. Finally, the results of model training are reviewed and saved.

Step 5. A day ahead prediction:

The CV-root mean square error (CV-RMSE), coefficient of determination (R^2), and mean absolute percentage error (MAPE) et al. are calculated by comparing the predicted day's results to the validation set.

In the real operation, the matching scenario is chosen first, then the model input features, based on Step 3. Following that, the data set is chosen in accordance with Steps 1 and 2 in order to create a prediction model.

2.2. Dataset of Two-Step Selection Using Features Search

2.2.1. Similarity Measurement

The daily and hourly energy consumptions fluctuate very smoothly over a typical cooling season, and the traditional selection training sample method can also produce reasonably excellent prediction results. The traditional method of data selection is to use data from one month earlier to the prediction day as the training sample. However, during the transitional season, daily energy consumption fluctuates non-stationarily, and hourly energy consumption fluctuates relatively widely every day. If the usual training sample selection strategy is utilized, the model's predictive performance will be severely harmed. As a result, a similarity metric is established to assess the similarity between historical feature data and daily predictions.

The KFSS method is not a carbon copy of the K nearest neighbors (KNN) algorithm, but rather borrows its concept, in which feature similarity is represented by neighboring points and quantified. The AC utilization rate and the outdoor temperature are compared to see how similar they are. The closest neighbors of the AC utilization rate in the predicted day, as well as the outdoor temperature in previous data, are searched and ranked. The similarity index is calculated using the daily average value and variance, normalized to the maximum and minimum values. The distance metric used in this study is Euclidean distance [49], which has the expression Equation (1). In the distance computation, two features with different weights are used, and their weights are determined by the RF method's feature importance module. Energy consumption in office buildings varies hugely between working days and non-working days. Only hourly energy consumption during workdays is predicted in this paper.

$$\mathrm{dist}(F_{pre}, F_{his}) = \sqrt{\sum_{i=1}^{n} (U_{pre} - U_i)^2 + \sum_{i=1}^{n} (T_{pre} - T_i)^2} \quad (1)$$

where F_{pre} is the key factor of the predict day; F_{his} is the key factor of the historical data; U_{pre} is the daily mean value of the predict daily AC utilization rate; U_i is the daily mean value of the historical AC utilization rate; T_{pre} is the daily mean value of the predict daily outdoor dry bulb temperature; T_i is the daily mean value of historical outdoor dry bulb temperature.

2.2.2. Feature Outlier-Based Selection for Dataset

Following the similarity assessment, the number of samples to be used as the training set must be determined. The LOF technique [49] is used with the box plot method to exclude aberrant hour samples, ensuring that the selected data is closest to the predicted daily features and no abnormal historical data. LOF calculates the reachability distance between the object p and the object o, as well as the object p's local reachability. Finally, the formula for computing the local outlier factor of the object p is Equation (2) [49]. When computing the locally accessible distance, the feature set must be normalized by min max and weighted due to the use of 2-dimensional features (air conditioning utilization rate and outdoor temperature). The feature weight uses the features significance module in the RF technique to assign a weight to each feature. The outlier factor of point p is the average of the local reachable density of the point in the k-th distance field of the point p divided by the local reachable density of the point p. The formula is:

$$\mathrm{LOF}_k(p) = \frac{\frac{\sum_{o \in N_k(p)} \mathrm{lrd}_k(o)}{|N_k(p)|}}{\mathrm{lrd}_k(p)} \tag{2}$$

where $\mathrm{lrd}_k(p)$ is the local reachable density of point p, which refers to the reciprocal of the average value of the k-th reachable distance from point p to the k-th reachable distance from point p. The formula is:

$$\mathrm{lrd}_k(p) = \frac{1}{\frac{\sum_{o \in N_k(p)} r - d_k(p,o)}{|N_k(p)|}} \tag{3}$$

The $r - d_k(p,o)$ is the k-th reachable distance from point o to point p, which refers to the maximum value of the k-th distance from point p and the distance from point p to point o. The formula is:

$$r - d_k(p,o) = max\{d_k(p), d(p,o)\} \tag{4}$$

where $d(p,o)$ is the distance between point p and point o, and $d_k(p)$ is the k-th distance of point p, which is the distance between the k-th farthest point from point p and point p except for point p; $N_k(p)$ is the k-th distance neighborhood of point p, which refers to the set of points within the k-th distance of p plus points at the k-th distance.

Therefore, the outlier factor of point p determines the relative density of point p and the surrounding points. If $\mathrm{LOF}_k(p) > 1$, it could be a problem. As illustrated in Figure 2, k was adjusted to 5.

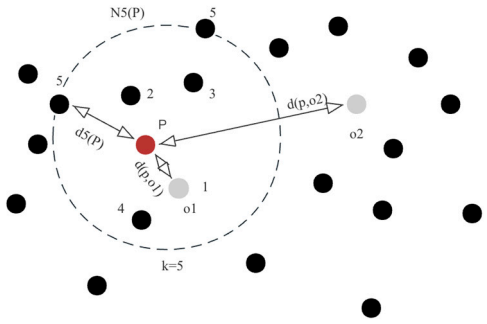


Figure 2. The schematic diagram of LOF algorithm (The distance to point P is indicated by the arrows in the diagram. The numbers represent the five closest points to point P).

2.3. Adapting Feature Set

2.3.1. Feature Types

Influencing variables are particularly significant throughout the building operation phase, which has a direct impact on mode reliability, and feature extraction is critical to the modeling's success. Sun et al. [4] categorized the different types of features in building HBECP models into seven categories: meteorological parameters, indoor environment information, building usage-related data, time stamp, building characteristic data, social information and historical data. Structure feature data (window-to-wall ratio, heat transfer coefficient, and so on) were determined after the building was completed, therefore they have no bearing on the scale of single-building modeling. Other sorts of features are included in this paper, but 3 and 6 are used in regional building or building group modeling.

The types of features extracted and summarized in this work are listed in Table 1, and they are classified into two categories: sequence features and associated features.

Table 1. Summary of feature types of building energy consumption factors.

Feature Source	Feature Types	Feature Name	Access
Sequence features	Time features	Time-related factors (day type, hour type)	Only mining the characteristics of historical energy consumption data
	Period features	At the same time the day before, etc. (discrete historical data)	
Associated features	External interference features	Outdoor weather parameters (temperature, humidity, etc.)	It needs to be obtained through information technology such as sensors
	Internal interference features	Indoor temperature and humidity. utilization rate	

(1) Sequence features are chosen based on historical operating energy consumption time series data. Data mining is used to investigate data features, which are then extracted as feature variables for prediction. A time series (e.g., time, date, holiday) and historical data are both examples of distinctive factors. The periodic intensity is identified using Fourier transform technology, and the discretized historical data is recovered as a feature using the person correlation coefficient.

(2) Associated features need to be obtained through sensors, which include the external interference features (e.g., outdoor meteorological parameters) and internal interference features (e.g., indoor dry bulb temperature, relative humidity, and utilization rate).

This paper's features are similar to those seen in most papers. However, in comparison to other research literatures [4], this paper is distinctive in that it combines time domain and frequency domain analysis to extract key historical time series features and defines occupancy behavior using the utilization rate of fan coil units.

2.3.2. Feature Set Selection

1. Extract Sequence Features

The sequence features are extracted based on previous energy consumption data, and several sorts of features affecting energy consumption are introduced above. The time information is obtained using one-hot encoding, while the period feature is extracted using the Fourier transform and person coefficient methods based on the periodicity and correlation of historical data. The fast Fourier transform (FFT) is frequently used in data processing to save time and cost. The correlation coefficient (r) is a common and helpful approach in energy consumption time series analysis. It is primarily used to extract historical data from time series that has a strong link with current hourly energy

consumption, as well as to investigate the relationship between energy consumption and associated variables. The technique of calculation is as follows [50]:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \quad (5)$$

where n is the number of samples, X_i is variable at the i time, Y_i is the other variable at the i time, $\overline{X}$ is the mean of the variable, $\overline{Y}$ is the mean of the other variable.

2. Different Feature Set Based on Availability

Some occupancy-related parameters are difficult to get in real-world engineering applications. However, this study investigates the use of AC utilization rate data, which is currently gathered in many buildings, to represent occupancy behavior information. As indicated in Table 2, this research provides three application scenarios based on the complexity of getting a feature set. In real-world engineering project modeling, features are chosen based on the actual application feature set to provide the greatest modeling effect and to identify the feature acquisition approach for model improvement.

Table 2. The settings of the three scenarios.

	Scenario 1	**Scenario 2**	**Scenario 3**
Feature set	Sequence features	Sequence features + External interference features + indoor temperature and humidity	Sequence feature + external disturbance factor + indoor temperature and humidity + utilization rate

The prediction and evaluation error is made based on the preceding conditions so that the feature set, which is chosen based on the features of their own data, may be quickly recognized. This study presents a gradient scenario analysis based on the availability of feature variables, which can increase performance the efficiency of actual modeling tasks. Furthermore, all numerical features are normalized by their maximum and minimum values, and categorical features are coded in one-hot mode.

2.4. RF Model

2.4.1. RF

The RF algorithm is a more enhanced form of bagging and one of the most advanced ensemble methods. The random tree algorithm in random forest [51] is shown in Figure 3. The RF method is a type of ensemble learning algorithm that has a lot of flexibility and superiority [51]. The self-service sampling approach is used to randomly select a sample from a training data set of N samples and place it in the sampling set, as shown in Figure 4. The sample is then reinserted into the original data set, producing a sampling set of m samples. Different sub-datasets are created in the same way. The sampling set contains approximately 63.2 percent of the samples from the initial training set [52]. A decision tree is a fundamental learner that is deployed for training. The introduction of feature selection is the key difference between RF and bagging approaches. When each decision tree chooses a segmentation point, RF will choose a feature subset at random first, then the typical bagging method's segmentation point. Not only does it increase the variety of base learners through sample perturbation (random sampling and selection of sub-datasets), but it also does so through feature perturbation (random selection of sub-feature sets). Finally, the difference in the base learner improves the generalization capacity of the integrated algorithm. The random regressor in the SciKit-Learn application programming interface is used to develop the RF technique in this paper.

Input： dataset D = {(x1, y1) , (x2, y2) , …., (xm, ym)}; Feature subset size K.
Steps：
(1). N ⟵ Construct a node for a give dataset D.
(2). If All of samples belong to one category, then return N.
(3). $\mathcal{F}$⟵ A set of features that can continue to be classified. (4). If $\mathcal{F}$ empty, then return N.
(5). $\tilde{\mathcal{F}}$ ⟵ Random selection of K features from the $\mathcal{F}$;
(6). $N.f$⟵ Features in feature $\tilde{\mathcal{F}}$ with the best segmentation points.
(7). $N.p$⟵ Best segmentation point in the $N.f$;
(8). D_l⟵ Sample subset with $N.f$ value less than $N.p$ in the D.
(9). D_r⟵ Sample subset with $N.f$ value not less than $N.p$ in the D.
(10). N_l⟵ Continue calling this program (D_l, K) arguments.
(11). N_r⟵ Continue calling this program (D_l, K) arguments.
(12). Return N.
Output: A random decision tree.

Figure 3. The random tree algorithm in RF.

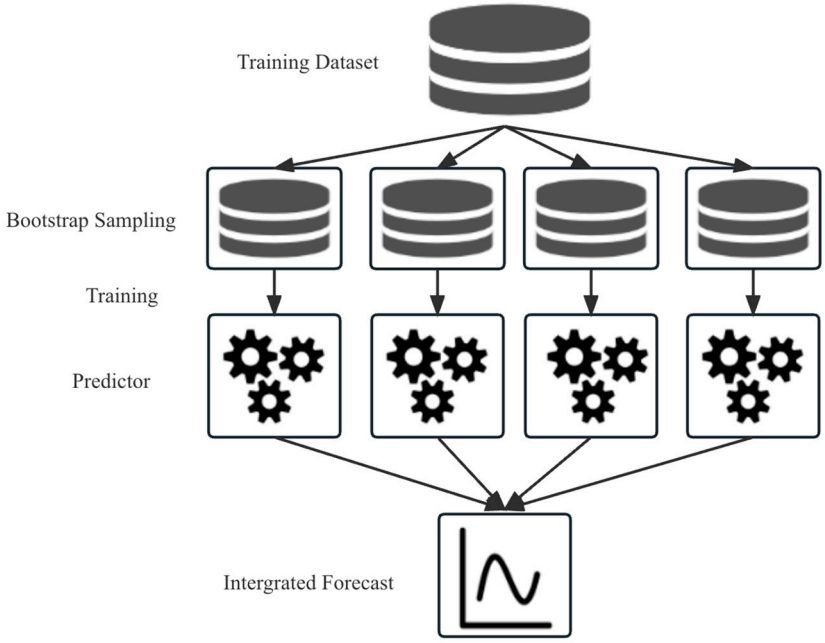


Figure 4. The prediction process of RF.

2.4.2. Parameters Optimization of RF

The machine learning model's hyperparameters must be set. There are also inaccuracies in the various hyperparameter models. The normal practice is to choose based on experience in order to find the best hyperparameters, however, optimality cannot be guaranteed. The grid search provides each hyperparameter's value range and step size and then determines the optimum hyperparameters (e.g., the maximum depth, and the minimum number of leaf nodes). This method can significantly identify the global ideal value if a broader range and smaller step size are employed, but it consumes computational resources and time. This study employs grid search to discover the essential hyperparameters of the RF algorithm, which is part of the SK-learn algorithm package. The number of trees in this paper is set to 100, the maximum depth to 6, and the minimum number of leaf nodes to 2. As well, all other hyperparameters are set to their default values.

2.5. Performance Evaluation Indices

To assess the model's performance, a performance evaluation index is required. The machine learning regression model evaluation standard is mostly employed in the modeling of power consumption time series.

The standard of model generalization error is as follows [53]:

(a) CV-Root mean squared error (CV-$RMSE$):

$$CV - RMSE = \frac{\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(E^{(i)} - \hat{E}^{(i)}\right)^2}}{\sum_{i}^{m} E^{(i)}/m} \quad (6)$$

(b) R square (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{m}\left(E^{(i)} - \hat{E}^{(i)}\right)^2}{\sum_{i=1}^{m} E_i^2} \quad (7)$$

(c) Mean absolute percentage error ($MAPE$):

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{\left(E^{(i)} - \hat{E}^{(i)}\right)}{E^{(i)}}\right| \times 100\% \quad (8)$$

where m is the number of samples, $E^{(i)}$ is the true power consumption density of the sample at the i time, W/m^2. $\hat{E}^{(i)}$ is the predicted power consumption density of the sample at the i time, W/m^2.

3. Case Study

3.1. Building and Data Description

The case study makes use of data retrieved from an office building in Zhuhai, China. The climate in Zhuhai is subtropical maritime. The building has a total area of 23,471 m^2 and a height of 73 m. There are seventeen levels above ground and one basement level. It is a zero-energy building with platinum leadership in energy and environmental design and three-star green building certification. The building's overall shape is that of two leaves, and the facade is a large-area glass curtain wall with photovoltaic panels for shading. The lobby, exhibition hall, meeting room, open office, enclosed office, information room, supporting restrooms, and equipment room are the main aspects of the case study building.

The office building's energy system consists of the air conditioning system, equipment system, and lighting system. The lighting system uses LED bulbs with low lighting power densities as one of them. The general layout of an office building's equipment system consists of office equipment (typical desktop computers, notebook computers, and printers), power equipment (two elevators, ventilators, and water pumps), electronic equipment (an information data center), and auxiliary facilities and equipment. Two 908.5 kW water-cooled screw chillers serve as the cold source for the centralized AC system. On the cooling side, there are two cross-flow cooling towers, and the cooling and refrigeration pumps are all frequency conversion water pumps. The AC's terminal employs a fan coil unit and a fresh air system. Besides this, the fresh air unit has a comprehensive heat recovery system for evaporative condensation.

In Section 2, the feature set was separated into two categories: historical energy consumption data and meteorological data. To get integrated datasets, an on-site collection of actual operating data is undertaken. The building's energy management platform provided the data used in this study. The data was collected over a one-hour period from 1 January 2018 to 31 December 2019. All of the features listed in Table 1 are covered by the types of data obtained. The fan coil unit's utilization rate is calculated by dividing the number of turns hourly by the total number, with the remainder of the data coming from sensors.

3.2. Dataset of Two-Step Selection

The working days in the transition season (spring) from April to June, the cooling season from July to August, and the transition season (autumn) from September to November are anticipated and the outcomes are examined using the method described in this

study. In addition, samples from different typical seasons are selected as sampled periods, including 12 April, 14 May, 4 July, 15 July, and 20–21 November. These samples were chosen for investigation in this work because energy demand has changed significantly these days compared to the previous day and falls within the context of changing energy demand trends.

First and primary, as illustrated in Figure 5, statistical data that meet the error standards. When comparing the KFSS approach to the conventional method, the predicted number of days whose MAPE is managed within 15% rises by 14.5% in spring and by 4.9% in fall. Both the traditional and new methods can maintain the projected day's MAPE under 15% in the summer. Second, observing statistics under the assumption of fulfilling the error is insufficient; it is important to investigate the differences in hourly prediction outcomes each day. Next, we'll look at the mistake in predicting typical seasonal days. Figure 6 shows a 24-h boxplot of the training sample selection outcomes in three typical seasons.

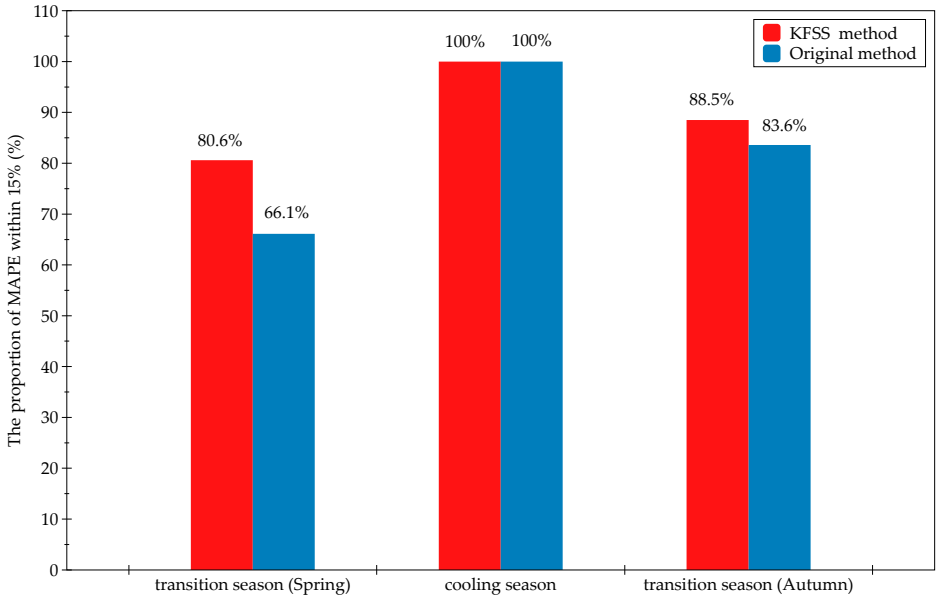


Figure 5. The daily percentage of accurate predictions.

The early phase of the transitional spring season, as illustrated in Figure 6a, occurs on 12 April. The two selection strategies in the prediction day generated obviously distinct properties in the training sample sets. The blue training sample set has a lot of strewn data, and there are a lot of sample values that do not match the predicted date. Furthermore, the red training sample set created by the KFSS approach demonstrates excellent clustering. 14 May comes between the middle and late stages of the transition season, as indicated in Figure 6b. There has rarely been a big amount of scattered data in the blue training sample data. The training sample values are typically below the projected day, however, the red training samples are in good agreement with the prediction day, compared to the early part of the transition season. If the scattered training sample set or sample values are considerably lower than those on the prediction day, the machine learning model would fail for the normal spring transition season.

The difference between the two training sample selection approaches for the HBECP model of the day before the steady cooling season is quite tiny, as shown in Figure 6c,d. The fundamental reason for this is that throughout the consistent cooling season, the building's air conditioning system is always on, and the daily operational status is relatively similar. As a result, the traditional method's training samples, as well as the similarity of the KFSS method's training samples, are all consistent with the prediction day. For the autumn transition season, as illustrated in Figure 6e,f, the two selection techniques exhibit significant variations. The traditional method's training samples are more distributed, and there are

more outliers. In contrast to the spring transition season, the training sample values on 20 and 21 November are generally higher than the predicted days. Although the majority of the KFSS method's training samples are larger than the prediction data, the data is more concentrated. Finally, the two days of the predicted day are part of the late autumn transition season, when the AC system is broken. However, because the AC system is operational most days earlier than 20 November, there is a considerable amount of blue data that is higher than the prediction day.

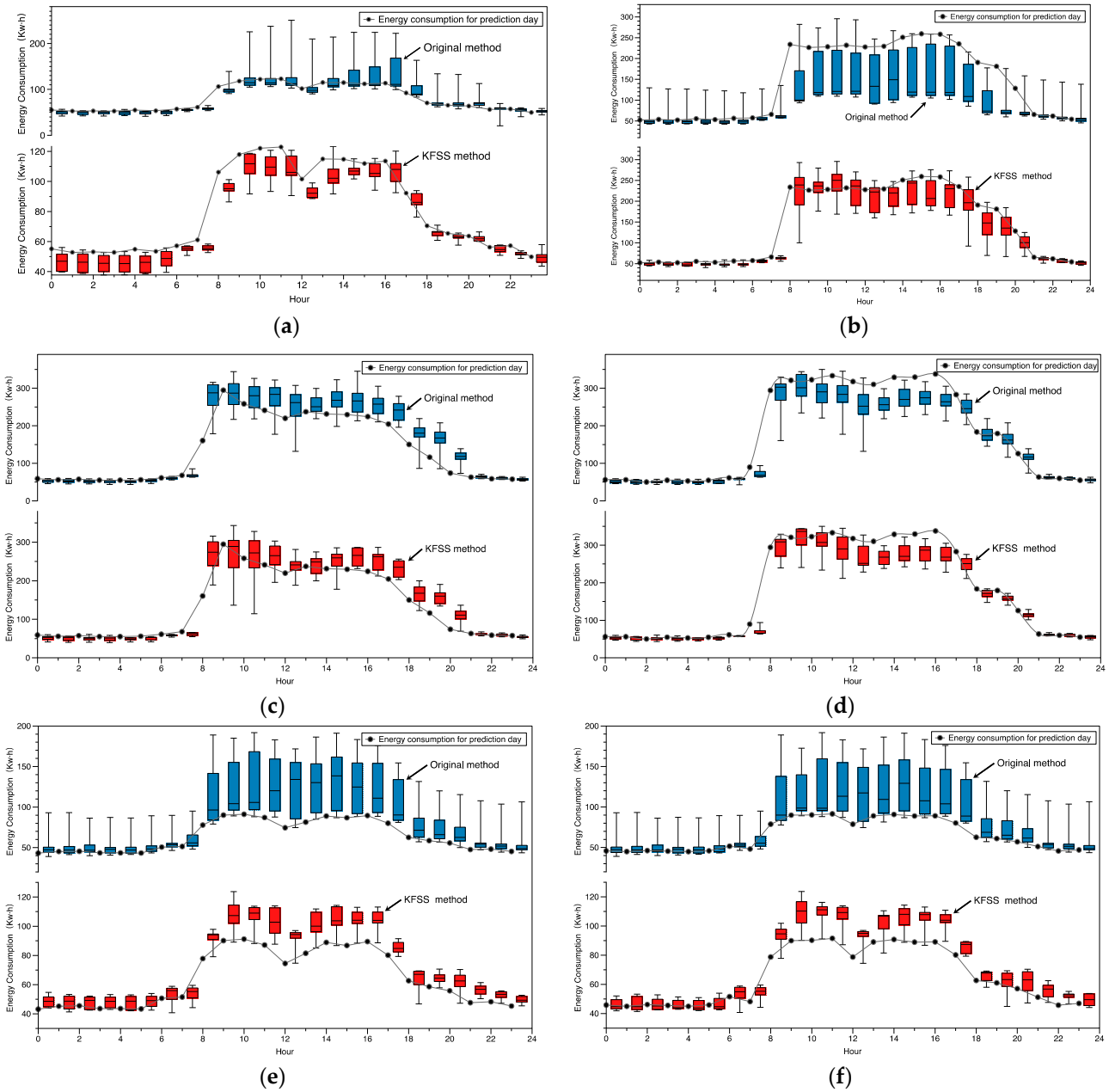


Figure 6. Comparison of the training data sets of the two methods in three typical seasons: (**a**) 12 April; (**b**) 14 May; (**c**) 4 July; (**d**) 15 July; (**e**) 20 November; (**f**) 21 November.

3.3. Feature Set Scenario

3.3.1. Sequence Features Selection

The FFT, which turns the time series of energy consumption into frequency signals for presentation, is shown in Figure 7. The signal amplitude with a one-day period is the highest, and the amplitude with a seven-day period is also greater. Other frequencies with

higher amplitudes are multiples of the one-day cycle and can be termed one-day cycles as well. As a result, based on the periodic pattern of the historical time series of energy consumption as illustrated in Figure 8, the person coefficient is used to determine the autocorrelation of energy consumption. The parameter of $E_{d\text{-}1,\,h\text{-}1}$ represents the energy consumption in the previous hour (h-1) of the previous day (d-1). A strong correlation is one with a correlation coefficient larger than 0.7. The anticipated energy consumption is strongly connected with energy consumption with a delay of fewer than 2 h and energy consumption over a seven-day cycle, with the correlation performance decreasing as the daily delay increases. The learning algorithm can preserve certain sequence feature redundancy in order to assure its stability, and the historical energy consumptions during 2 h and 2 days delay are chosen as the sequence feature. The following is a list of the seven-dimensional training feature samples included in this paper.

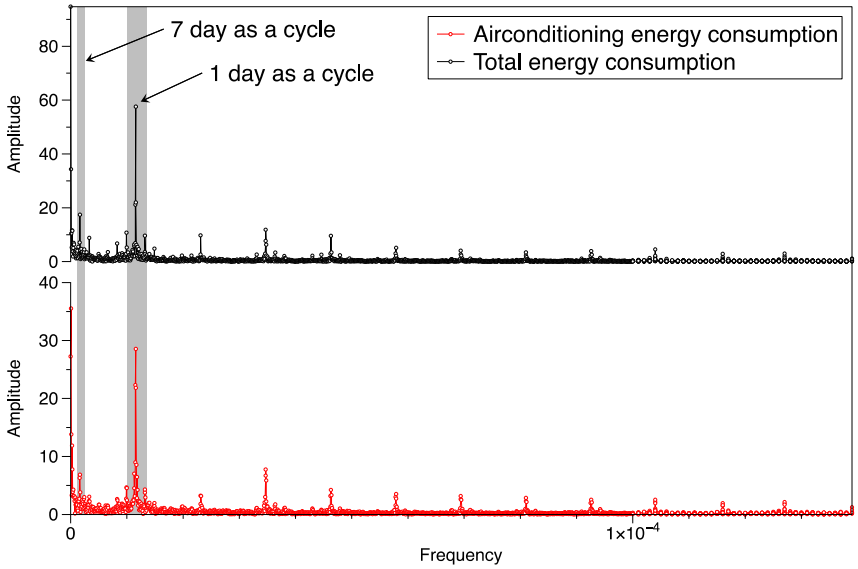


Figure 7. The result of the Fast Fourier Transform.

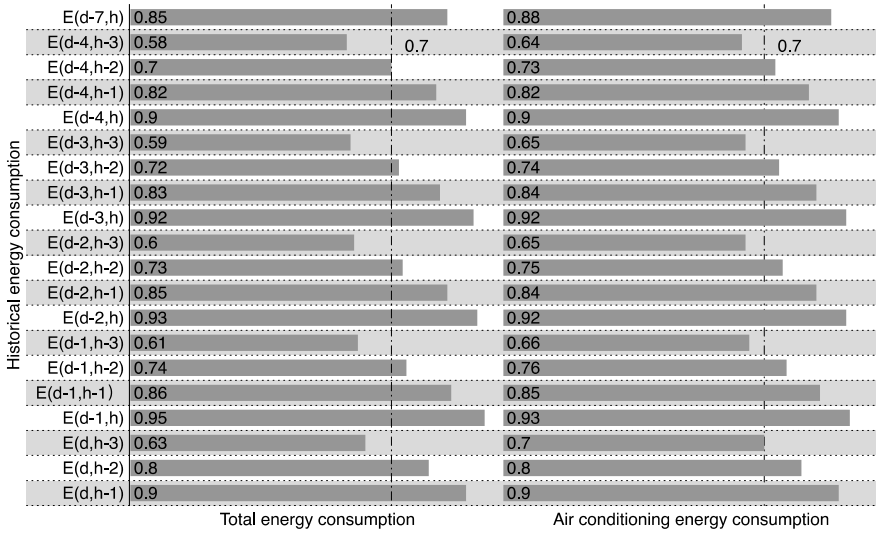


Figure 8. The result of correlation calculation between energy consumption and sequence features.

- EDI at the same hour of the day before, two days earlier, and seven days earlier: $E_{d\text{-}1,\,h}$, $E_{d\text{-}2,\,h}$, $E_{d\text{-}7,\,h}$.
- EDI at the earlier hours of one day before: $E_{d\text{-}1,\,h\text{-}1}$ and $E_{d\text{-}1,\,h\text{-}2}$.
- EDI at the earlier hours of two days earlier: $E_{d\text{-}2,\,h\text{-}1}$ and $E_{d\text{-}2,\,h\text{-}2}$.

3.3.2. Scenario of Feature Set

For specialized building energy consumption modeling, the feature set must be considered, but not as many as feasible. Because not all features are available in real-life structures. As a result, it is divided into three situations for discussion based on the capabilities available in Section 2.3.2. Each scenario is broken down into three distinct sets of factors. Table 3 shows the feature sets.

Table 3. Evaluation of prediction results of typical feature groups under different scenarios.

Scenario	Number	Feature Set
Scenario 1	Feature set 1	E (d-1, h), E (d-2, h)
	Feature set 2	E (d-1, h), E (d-2, h), E (d-7, h), E (d-1, h-1), E (d-2, h-1)
	Feature set 3	E (d-1, h), E (d-2, h), E (d-7, h), E (d-1, h-1), E (d-2, h-1), E (d-1, h-2), E (d-2, h-2)
Scenario 2	Feature set 4	outdoor temperature, outdoor humidity, indoor temperature, indoor humidity, hour type
	Feature set 5	E (d-1, h), E (d-2, h), E (d-7, h), E (d-1, h-1), E (d-2, h-1), E (d-1, h-2), E (d-2, h-2), outdoor temperature
	Feature set 6	E (d-1, h), E (d-2, h), E (d-7, h), E (d-1, h-1), E (d-2, h-1), E (d-1, h-2), E (d-2, h-2), Indoor and outdoor temperature difference
Scenario 3	Feature set 7	utilization rate
	Feature set 8	utilization rate, outdoor temperature, outdoor humidity, indoor temperature, indoor humidity, hour type
	Feature set 9	utilization rate, outdoor temperature, E (d-1, h), E (d-2, h), E (d-7, h), E (d-1, h-1), E (d-2, h-1), E (d-1, h-2), E (d-2, h-2)

4. Results and Discussion

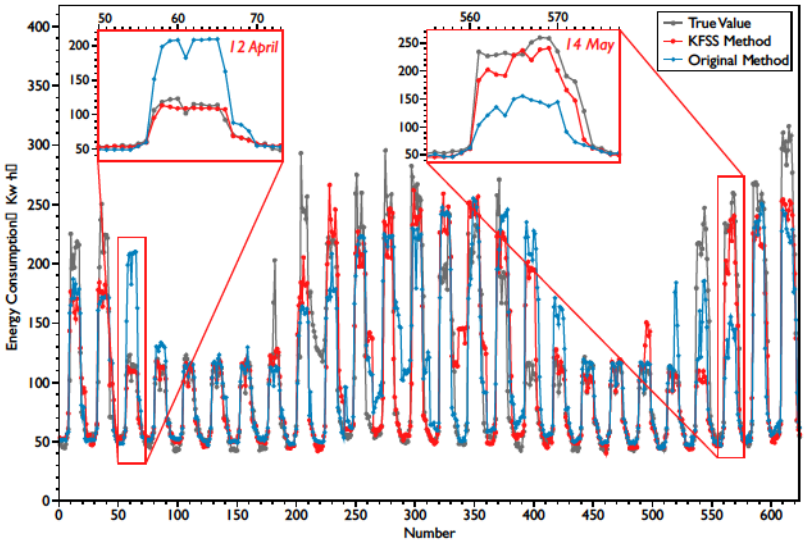
This undertaking provides additional adaptability analysis in terms of two sections, as follows, based on the KFSS approach to choosing training samples: effectiveness of training sample selection and results of feature scenario analysis.

4.1. Effectiveness Analysis of Training Set Selection

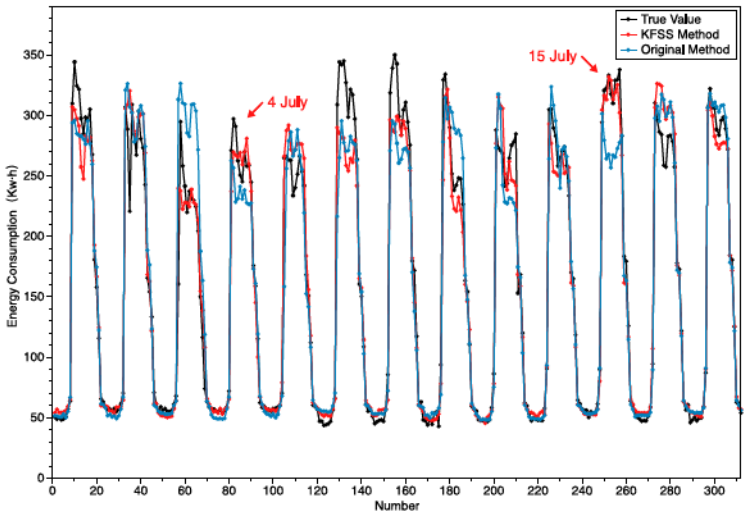
Table 4 highlights the assessment indicators, whereas Figure 9 depicts the learning algorithm's prediction performance after training. When comparing the expected outcomes of the predicted day with actual energy consumption over three typical seasons, the KFSS strategies performed admirably. It can rapidly and precisely estimate the energy consumption of the predicted day, especially during the transition season. The difference between the prediction results generated by the traditional approach and the KFSS method is minimal in a typical cooling season, but it does exist. The prediction result is consistent with the performance chosen in Section 3.2's training sample. In modeling, the selection of training samples is critical.

Table 4. Evaluation of the prediction results of the two methods.

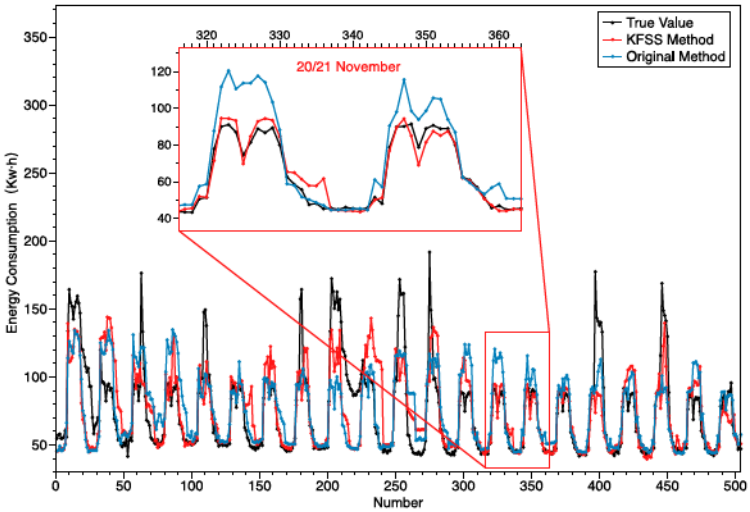
Predict Day	MAPE (%)	R^2	CV-RMSE (%)	Method
12 April	5.10	0.95	7.76	KFSS method
	10.83	0.94	14.69	Original method
14 May	12.04	0.92	16.40	KFSS method
	22.66	0.46	44.37	Original method
4 July	4.73	0.98	8.39	KFSS method
	13.94	0.89	20.46	Original method
15 July	4.58	0.99	6.25	KFSS method
	6.53	0.98	8.82	Original method
20 November	7.80	0.90	9.24	KFSS method
	15.76	0.20	26.08	Original method
21 November	3.46	0.97	5.13	KFSS method
	11.59	0.67	16.80	Original method



(**a**)



(**b**)



(**c**)

Figure 9. Comparison of prediction results of two training set selection methods: (**a**) Typical transition season (Spring); (**b**) Typical cooling season (Summer) (**c**) Typical transition season (Autumn).

The R^2 achieved by the KFSS method are all over 0.95, and the MAPE are all within 10% in terms of prediction error evaluation. During the transitional season, the traditional method's MAPE was higher than 10%. On 14 May, the MAPE was 22%, and the KFSS method had been increased to 12%. During the cooling season, the prediction findings on 15 July are nearly the same, and the MAPE is raised from 6% to 4%. The error ranges for both approaches are within acceptable limits. Overall, the KFSS method's evaluation is considerably superior to the traditional technique.

Figure 9a,c shows that the building's operation gets complex during the transition season, which makes predicting energy consumption difficult. The figure shows that, in contrast to the original method, the prediction results of the KFSS method are able to keep up with changes in the pattern of energy consumption. This is due to the fact that the classic technique's selection of training data only considers data from the previous month, whereas the KFSS method creates a model using training data that was chosen based on the traits of the predicted day. It is worth noting that the columnar portion of Figure 9a demonstrates that the real energy consumption is aberrant at night, and neither approach can produce typical prediction results. This is not a prediction error produced by the approach. This simply illustrates the role of prediction results in the early detection and management of abnormal energy consumption. Figure 9b shows that during a regular cooling season, the energy consumption pattern is largely consistent and the energy consumption of neighboring days fluctuates only a little. As a result, in terms of the predictions of energy consumption for this season, there is little difference between the original technique and the KFSS method. This is because the original method may be chosen in the stable operation mode and most data that are comparable to the forecast day can be chosen as the training set. Such a strategy, however, is rigid and fraught with risk.

In conclusion, the suggested KFSS approach efficiently selects training samples depending on the features of the target prediction day, whereas the traditional selection method is rigid. The KFSS approach outperformed the original method in the evaluation of the prediction results, especially during the transition season, while the original method surpassed the permissible range in the evaluation of the prediction results. The transition season is the most difficult to predict and accurately model in terms of energy consumption, and the KFSS strategy in this paper does a good job of solving this problem.

4.2. Results of Feature Scenario Analysis

Table 5 shows the feature sets and evaluation findings, whereas Figure 10 shows the prediction results. Only historical energy consumption data is available in scenario 1, thus sequence feature extraction can be used to make predictions. The R^2 is greater than 0.9, and the average daily MAPE performance is excellent. From feature set 1 to feature set 2, MAPE is lowered to 8.09% and R^2 is increased from 0.93 to 0.95, demonstrating that keeping some redundant features can improve model accuracy. In Scenario 2, the data includes exterior meteorological features and indoor temperature and humidity, as well as historical energy use data. As a result, as indicated in the table of feature sets 5 and 6, the model's accuracy can be enhanced further. However, the MAPE in feature set 5 is 19.60%, and the R^2 is 0.78, indicating that the prediction evaluation is quite poor. This is due to the fact that only the associated features are used, and the retrieved sequence features have a significant beneficial impact on the model.

In Scenario 3, the data for modeling was sufficient, and the building's use behavior data—AC utilization rate—was received. It can reflect the number of people that are in the building, and the utilization rate data can better capture the energy consumption trend. Even if the MAPE is 12 percent, Feature Set 7 reveals that if only the utilization rate parameter is considered, the MAPE is not the best. When employing correlation features such as utilization rate, feature set 8 shows that the MAPE is 10.71% and the R^2 is 0.95, indicating that the model has a strong assessment performance. Feature set 8 demonstrates that the addition of sequence features can increase performance regardless of the number

of correlation features employed. Because the model is so accurate, it is clear that previous data is required.

Table 5. Evaluation of prediction results of typical feature groups under different scenarios.

Scenario	Number	Features	R^2	MAPE (%)
Scenario 1	Feature set 1	E (d-1, h), E (d-2, h)	0.93	9.99
	Feature set 2	E (d-1, h), E (d-2, h), E (d-7, h), E (d-1, h-1), E (d-2, h-1)	0.95	8.09
	Feature set 3	E (d-1, h), E (d-2, h), E (d-7, h), E (d-1, h-1), E (d-2, h-1), E (d-1, h-2), E (d-2, h-2)	0.94	8.33
Scenario 2	Feature set 4	outdoor temperature, outdoor humidity, indoor temperature, indoor humidity, hour type	0.78	19.60
	Feature set 5	E (d-1, h), E (d-2, h), E (d-7, h), E (d-1, h-1), E (d-2, h-1), E (d-1, h-2), E (d-2, h-2), outdoor temperature	0.96	7.43
	Feature set 6	E (d-1, h), E (d-2, h), E (d-7, h), E (d-1, h-1), E (d-2, h-1), E (d-1, h-2), E (d-2, h-2), Indoor and outdoor temperature difference	0.96	6.96
Scenario 3	Feature set 7	utilization rate	0.94	12.26
	Feature set 8	utilization rate, outdoor temperature, outdoor humidity, indoor temperature, indoor humidity, hour type	0.95	10.71
	Feature set 9	utilization rate, outdoor temperature, E (d-1, h), E (d-2, h), E (d-7, h), E (d-1, h-1), E (d-2, h-1), E (d-1, h-2), E (d-2, h-2)	0.96	6.99

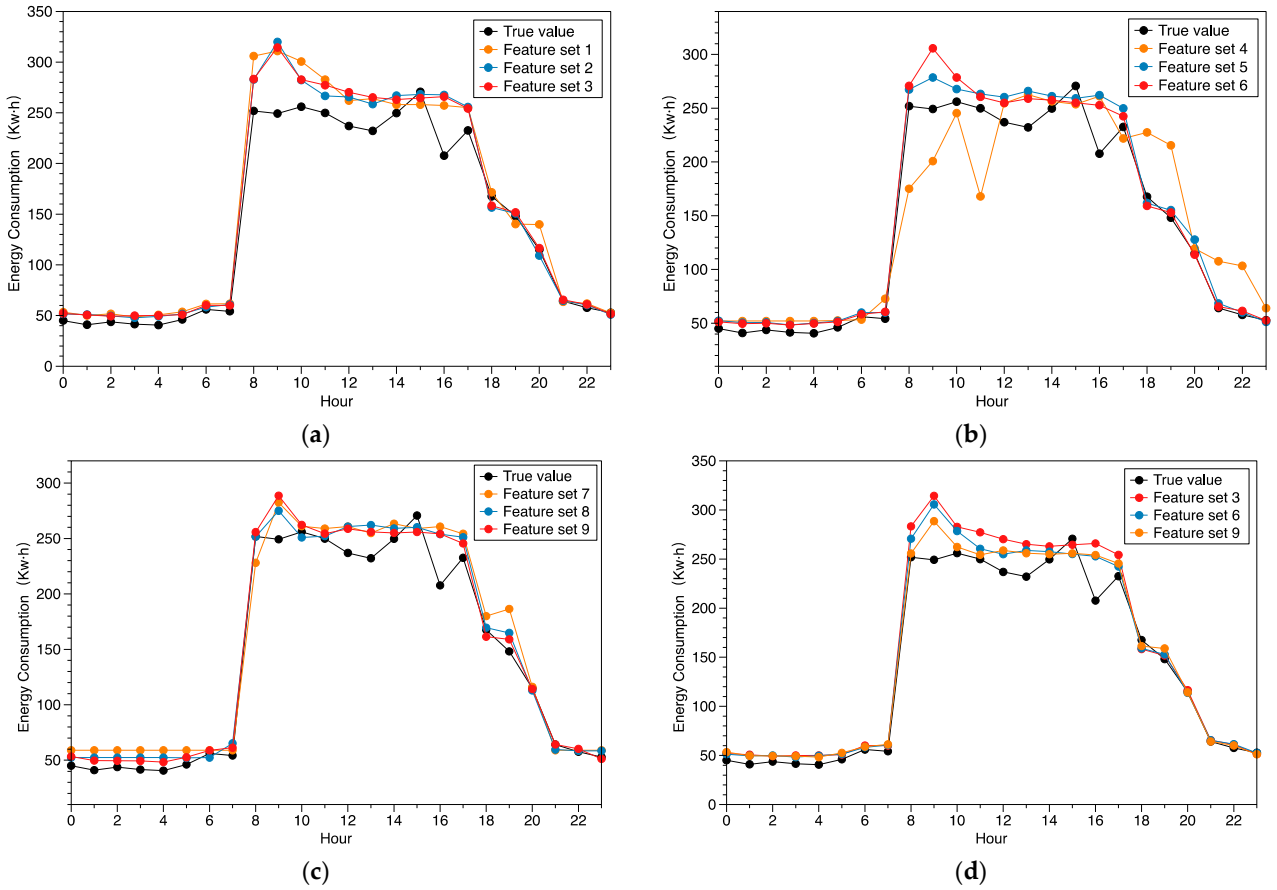


Figure 10. Typical feature group in different scenarios: (**a**) Scenario 1; (**b**) Scenario 2; (**c**) Scenario 3; (**d**) Comparison of three scenarios.

Figure 10 a–c shows the results of a 24-h prediction, focusing solely on the difference in energy consumption during working hours. In scenario 1, feature sets 2 and 3 outperform feature set 1 in terms of predictions. In Scenario 2, the prediction outcome of feature set 4 is undesirable, and the prediction error of feature set 5 is smaller than that of feature set 6 at 9 o'clock. Only the prediction of feature set 7 generates big errors at 8 and 19 o'clock in scenario 3. The prediction results of scenario 3 are better than scenarios 1 and 2 when the utilization rate is employed as the feature variable. Figure 10d compares the best feature set for each of the three scenarios and finds that feature set 9 is the best, particularly in the performance range of 8 to 9 when the building first started working.

By incorporating occupancy behavior data into consideration, this study combines the study of training sample selection and feature scenarios. This supports the significance of occupancy behavior statistics noted in the literature [54] even more. The KFSS method, developed using occupancy behavior and outdoor meteorological information as important elements, improves the model's adaptability when compared to methods described in the literature [28,29]. The study of feature scenarios can give modelers guidance for improving models in actual modeling projects in comparison to the feature selection and feature extraction techniques described in the literature [40,55]. Also, one of the most significant recent study directions is the model's interpretability. The description of the physical significance of the data itself is insufficient because much of the research in the literature [56,57] focuses on explaining the algorithm. This article can enhance the physical understanding of the data's significance and increase confidence in the model.

5. Application

5.1. Modeling Adaptability

HBECP has a huge amount of nonlinear characteristics. Prediction objects include various building types, energy consumption types, and time periods to predict. As a result, there is no universal paradigm that can solve all issues. The HBECP must be updated as needed to meet current needs. Three components of the model's adaptability can be summarized:

The training sample set's flexibility is based on KFSS. The KFSS approach can correctly capture the trend and improves prediction accuracy for all forms of energy consumption, whether in the cooling season or the transition season.

The flexibility of scenario analysis when it comes to feature selection. In the process of building energy consumption modeling, features such as building utilization rate can increase time-to-time accuracy. The situational investigation feature group assists the modeler in maximizing the model's correctness based on the real project's data characteristics.

This method is unsuccessful in predicting outside the realm of historical experience, where precise results are impossible to obtain. It will also fail on transitional seasonal days with large hourly changes within the predicted day.

5.2. Application Implications

This model is designed for buildings in the operation stage that have multi-dimensional large datasets, and it is a critical task to be accurately day-ahead hourly HBECP. In a nutshell, it serves as a foundation for effective building operation management and energy diagnosis, as well as a scientific reference for demand response. The specific application reference of the method to improve the prediction accuracy proposed are summarized as follows:

The modeling methodologies demand improvements in prediction accuracy, resilience, and generalization ability due to the complex nonstationary and nonlinear problems between inputs and outputs of HBECP models. In this study, the KFSS approach is proposed as a solution to this problem. A vast amount of multi-dimensional data will be created in the future as a result of the continual accumulation of operating data. The early training sample selection, which was based on clustering and classification notions, will have significant limits. At this time, the KFSS method will have superior performance

and may be used to properly predict various operational time periods, particularly the transition season with complex operating conditions, while the KFSS methods have strong prediction accuracy and stability.

The features scenario analysis of this paper can be used as a reference to prejudge the accuracy and robustness of modeling without the requirement for blind selection, considering the acquisition of actual buildings and the level of modelers. The paper's originality comes from its energy consumption modeling standpoint. Rather than a unique learning method, it is based on the available data of the building as the center to differentiate the contextual dialogue. There will be no universal algorithm for all buildings in the HBECP in the future. It must be a precise model of a certain structure. A reference basis for feature selection is needed at this time.

6. Conclusions

Building energy operation management relies heavily on short-term HBECP. Predictability and interpretability have been the major research focuses. The process of selecting training samples in data-driven modeling is tightly connected to the model's stability and interpretability. This work incorporates both the important characteristic aspects and the time period of predicted demand, and develops a training sample selection strategy based on characteristic scenarios to improve the model's stability and interpretability. The technique provided in this paper presents a novel perspective and methodology for selecting training samples for energy modeling.

On the basis of feature search, a KFSS technique was proposed. The AC utilization rate is used to replace occupant behavior in buildings, properly reflecting the energy consumption. As well, then looking for historical data that is highly consistent with demand using a similarity assessment based on energy demand as the data-driven model's training sample. The Scenarios with different feature sets are established based on data availability. Time series, periodic characteristics, external association features, and internal association features are split into the features, and application scenarios are developed according to the slope of acquisition ability. Unlike existing modeling techniques such as feature selection and extraction, the suggested feature scene gradient considerably improves the model's usability.

The suggested method greatly increases the hourly prediction accuracy of the full year, particularly the prediction ability of the transition season. The KFSS method can be used on a variety of building types with itemized energy use that is changed by demand fluctuations. It has good practical utility for building energy management, demand response, and energy savings since it can give precise energy consumption scheduling systems for building energy loads. This method, however, has limitations, including elements other than the time period, load rate, and outdoor temperatures, such as sensor anomaly, the occurrence of conditions outside of historical scenarios, and other considerations. This strategy cannot account for the presence of these factors, which is why the predictions of the two methods are inaccurate.

In conclusion, the case study demonstrates that the suggested method is accurate and robust and that it is appropriate for online prediction on intelligent platforms. Furthermore, because the association between energy consumption and the need for reliable lighting sockets is relatively weak, the KFSS method should be applied with caution. In the future, the suggested strategy focuses on energy consumption categories, such as AC energy consumption, total energy consumption, etc., that are closely related to the occupancy behavior of buildings. To improve the accuracy of sub-space energy zone prediction in the future, greater attention will be devoted to the needs of different functional parts of the building. Finally, the KFSS methods are refined to different places to realize the energy prediction of the building from the whole to the part, considering the imbalance between the overall structure and the functional parts.

Author Contributions: Conceptualization, H.T. and H.F.; methodology, H.F.; software, N.D. and H.F.; validation, H.F., N.D. and Z.L.; formal analysis, H.F.; investigation, H.F.; resources, H.T.; data curation, H.F.; writing—review and editing, H.F. and R.K.; visualization, H.F.; supervision, H.F. and R.K.; project administration, H.F.; funding acquisition, H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Key R&D Program of China (Grant no. 2017YFC0704200).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This study was supported by the China Scholarship Council (CSC).

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

Abbreviation

Nomenclature	
AC	Air-conditioning
HBECP	Hourly building energy consumption prediction
CV-RMSE	CV-Root mean squared error
DMT	Data mining technology
DNN	Deep neural network
FFT	Fast fourier transform
FCM	Fuzzy C-means
KNN	K nearest neighbors
LSTM	Long short-term memory
LOF	Local outlier factor
MAPE	Mean absolute percentage error
RF	Random forest
R^2	Coefficient of determination
RMSE	Root mean squared error
Symbols	
$E^{(i)}$	True power consumption density of the sample at the *i* time
$\hat{E}^{(i)}$	Predicted power consumption density of the sample at the *i* time
F_{pre}	Key factor of the predict day
F_{his}	Key factor of the historical data
r	Correlation coefficient
T_i	Daily mean value of historical outdoor dry bulb temperature
T_{pre}	Daily mean value of the predict daily outdoor dry bulb temperature
U_{pre}	Daily mean value of the predict daily AC utilization rate
U_i	Daily mean value of the historical AC utilization rate
X_i	Variable at the *i* time
$\bar{X}$	Mean of the variable
Y_i	Other variable at the *i* time
$\bar{Y}$	Mean of the other variable
$LOF_k(p)$	The average of the local reachable density of the point in the k-th distance field of the point p divided by the local reachable density of the point p
$lrd_k(p)$	The local reachable density of point p
$r\text{-}d_k(p,o)$	The k-th reachable distance from point o to point p
d(p,o)	The distance between point p and point o
$d_k(p)$	The k-th distance of point p
$N_k(p)$	The k-th distance neighborhood of point p

References

1. Wang, Z.; Huang, W.; Chen, Z. The peak of CO_2 emissions in China: A new approach using survival models. *Energy Econ.* **2019**, *81*, 1099–1108. [CrossRef]
2. Himeur, Y.; Ghanem, K.; Alsalemi, A.; Bensaali, F.; Amira, A. Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *Appl. Energy* **2021**, *287*, 116601. [CrossRef]
3. Bourdeau, M.; Zhai, X.Q.; Nefzaoui, E.; Guo, X.F.; Chatellier, P. Modeling and forecasting building energy consumption: A review of data-driven techniques. *Sustain. Cities Soc.* **2019**, *48*, 101533. [CrossRef]
4. Sun, Y.; Haghighat, F.; Fung, B.C.M. A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy Build.* **2020**, *221*, 110022. [CrossRef]
5. Li, Y.; O'Neill, Z.; Zhang, L.; Chen, J.; Im, P.; DeGraw, J. Grey-box modeling and application for building energy simulations—A critical review. *Renew. Sustain. Energy Rev.* **2021**, *146*, 111174. [CrossRef]
6. Goudarzi, S.; Anisi, M.H.; Kama, N.; Doctor, F.; Soleymani, S.A.; Sangaiah, A.K. Predictive modelling of building energy consumption based on a hybrid nature-inspired optimization algorithm. *Energy Build.* **2019**, *196*, 83–93. [CrossRef]
7. Chen, Y.; Tan, H. Short-term prediction of electric demand in building sector via hybrid support vector regression. *Appl. Energy* **2017**, *204*, 1363–1374. [CrossRef]
8. Dai, Y.; Zhao, P. A hybrid load forecasting model based on support vector machine with intelligent methods for feature selection and parameter optimization. *Appl. Energy* **2020**, *279*, 115332. [CrossRef]
9. Wang, Z.; Wang, Y.; Zeng, R.; Srinivasan, R.S.; Ahrentzen, S. Random Forest based hourly building energy prediction. *Energy Build.* **2018**, *171*, 11–25. [CrossRef]
10. Dong, Z.; Liu, J.; Liu, B.; Li, K.; Li, X. Hourly energy consumption prediction of an office building based on ensemble learning and energy consumption pattern classification. *Energy Build.* **2021**, *241*, 110929. [CrossRef]
11. Wang, R.; Lu, S.; Feng, W. A novel improved model for building energy consumption prediction based on model integration. *Appl. Energy* **2020**, *262*, 114561. [CrossRef]
12. Fan, C.; Xiao, F.; Zhao, Y. A short-term building cooling load prediction method using deep learning algorithms. *Appl. Energy* **2017**, *195*, 222–233. [CrossRef]
13. Wang, J.Q.; Du, Y.; Wang, J. LSTM based long-term energy consumption prediction with periodicity. *Energy* **2020**, *197*, 117197. [CrossRef]
14. Moon, J.; Jung, S.; Rew, J.; Rho, S.; Hwang, E. Combination of short-term load forecasting models based on a stacking ensemble approach. *Energy Build.* **2020**, *216*, 109921. [CrossRef]
15. Zhang, L.; Wen, J. Active learning strategy for high fidelity short-term data-driven building energy forecasting. *Energy Build.* **2021**, *244*, 111026. [CrossRef]
16. Liu, T.; Tan, Z.; Xu, C.; Chen, H.; Li, Z. Study on deep reinforcement learning techniques for building energy consumption forecasting. *Energy Build.* **2020**, *208*, 109675. [CrossRef]
17. Zhou, X.; Lin, W.; Kumar, R.; Cui, P.; Ma, Z. A data-driven strategy using long short term memory models and reinforcement learning to predict building electricity consumption. *Appl. Energy* **2022**, *306*, 118078. [CrossRef]
18. Zhang, L.; Wen, J.; Li, Y.; Chen, J.; Ye, Y.; Fu, Y.; Livingood, W. A review of machine learning in building load prediction. *Appl. Energy* **2021**, *285*, 116452. [CrossRef]
19. Amasyali, K.; El-Gohary, N. Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings. *Renew. Sustain. Energy Rev.* **2021**, *142*, 110714. [CrossRef]
20. Oh, K.; Kim, E.-J.; Park, C.-Y. A Physical Model-Based Data-Driven Approach to Overcome Data Scarcity and Predict Building Energy Consumption. *Sustainability* **2022**, *14*, 9464. [CrossRef]
21. Kim, D.; Lee, Y.; Chin, K.; Mago, P.J.; Cho, H.; Zhang, J. Implementation of a Long Short-Term Memory Transfer Learning (LSTM-TL)-Based Data-Driven Model for Building Energy Demand Forecasting. *Sustainability* **2023**, *15*, 2340. [CrossRef]
22. Tian, C.; Li, C.; Zhang, G.; Lv, Y. Data driven parallel prediction of building energy consumption using generative adversarial nets. *Energy Build.* **2019**, *186*, 230–243. [CrossRef]
23. Qian, F.; Gao, W.; Yang, Y.; Yu, D. Potential analysis of the transfer learning model in short and medium-term forecasting of building HVAC energy consumption. *Energy* **2020**, *193*, 116724. [CrossRef]
24. Fan, C.; Sun, Y.; Xiao, F.; Ma, J.; Lee, D.; Wang, J.; Tseng, Y.C. Statistical investigations of transfer learning-based methodology for short-term building energy predictions. *Appl. Energy* **2020**, *262*, 114499. [CrossRef]
25. Fan, C.; Lei, Y.; Sun, Y.; Piscitelli, M.S.; Chiosa, R.; Capozzoli, A. Data-centric or algorithm-centric: Exploiting the performance of transfer learning for improving building energy predictions in data-scarce context. *Energy* **2022**, *240*, 2775. [CrossRef]
26. Bedi, J.; Toshniwal, D. Deep learning framework to forecast electricity demand. *Appl. Energy* **2019**, *238*, 1312–1326. [CrossRef]
27. Somu, N.; Gauthama Raman, M.R.; Ramamritham, K. A deep learning framework for building energy consumption forecast. *Renew. Sustain. Energy Rev.* **2021**, *137*, 110591. [CrossRef]
28. Acquah, M.A.; Jin, Y.; Oh, B.-C.; Son, Y.-G.; Kim, S.-Y. Spatiotemporal Sequence-to-Sequence Clustering for Electric Load Forecasting. *IEEE Access* **2023**, *11*, 5850–5863. [CrossRef]

29. Chen, Y.; Tan, H.; Berardi, U. Day-ahead prediction of hourly electric demand in non-stationary operated commercial buildings: A clustering-based hybrid approach. *Energy Build.* **2017**, *148*, 228–237. [CrossRef]
30. Jallal, M.A.; González-Vidal, A.; Skarmeta, A.F.; Chabaa, S.; Zeroual, A. A hybrid neuro-fuzzy inference system-based algorithm for time series forecasting applied to energy consumption prediction. *Appl. Energy* **2020**, *268*, 114977. [CrossRef]
31. Piscitelli, M.S.; Brandi, S.; Capozzoli, A. Recognition and classification of typical load profiles in buildings with non-intrusive learning approach. *Appl. Energy* **2019**, *255*, 113727. [CrossRef]
32. He, F.; Zhou, J.; Feng, Z.-k.; Liu, G.; Yang, Y. A hybrid short-term load forecasting model based on variational mode decomposition and long short-term memory networks considering relevant factors with Bayesian optimization algorithm. *Appl. Energy* **2019**, *237*, 103–116. [CrossRef]
33. Zhang, G.; Tian, C.; Li, C.; Zhang, J.J.; Zuo, W. Accurate forecasting of building energy consumption via a novel ensembled deep learning method considering the cyclic feature. *Energy* **2020**, *201*, 117531. [CrossRef]
34. Zhang, L.; Alahmad, M.; Wen, J. Comparison of time-frequency-analysis techniques applied in building energy data noise cancellation for building load forecasting: A real-building case study. *Energy Build.* **2021**, *231*, 110592. [CrossRef]
35. Zhou, Y.; Wang, L.; Qian, J. Application of Combined Models Based on Empirical Mode Decomposition, Deep Learning, and Autoregressive Integrated Moving Average Model for Short-Term Heating Load Predictions. *Sustainability* **2022**, *14*, 7349. [CrossRef]
36. Peng, L.; Wang, L.; Xia, D.; Gao, Q. Effective energy consumption forecasting using empirical wavelet transform and long short-term memory. *Energy* **2022**, *238*, 121756. [CrossRef]
37. Fan, C.; Wang, J.; Gang, W.; Li, S. Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. *Appl. Energy* **2019**, *236*, 700–710. [CrossRef]
38. 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]
39. Somu, N.; Gauthama Raman, M.R.; Ramamritham, K. A hybrid model for building energy consumption forecasting using long short term memory networks. *Appl. Energy* **2020**, *261*, 114131. [CrossRef]
40. Qiao, Q.; Yunusa-Kaltungo, A.; Edwards, R.E. Feature selection strategy for machine learning methods in building energy consumption prediction. *Energy Rep.* **2022**, *8*, 13621–13654. [CrossRef]
41. Bianchi, C.; Zhang, L.; Goldwasser, D.; Parker, A.; Horsey, H. Modeling occupancy-driven building loads for large and diversified building stocks through the use of parametric schedules. *Appl. Energy* **2020**, *276*, 115470. [CrossRef]
42. Peng, Y.; Rysanek, A.; Nagy, Z.; Schlüter, A. Using machine learning techniques for occupancy-prediction-based cooling control in office buildings. *Appl. Energy* **2018**, *211*, 1343–1358. [CrossRef]
43. Wei, Y.; Xia, L.; Pan, S.; Wu, J.; Zhang, X.; Han, M.; Zhang, W.; Xie, J.; Li, Q. Prediction of occupancy level and energy consumption in office building using blind system identification and neural networks. *Appl. Energy* **2019**, *240*, 276–294. [CrossRef]
44. Liu, J.; Zhang, Q.; Dong, Z.; Li, X.; Li, G.; Xie, Y.; Li, K. Quantitative evaluation of the building energy performance based on short-term energy predictions. *Energy* **2021**, *223*, 120065. [CrossRef]
45. Shao, M.; Wang, X.; Bu, Z.; Chen, X.; Wang, Y. Prediction of energy consumption in hotel buildings via support vector machines. *Sustain. Cities Soc.* **2020**, *57*, 102128. [CrossRef]
46. Ahmad, T.; Zhang, H. Novel deep supervised ML models with feature selection approach for large-scale utilities and buildings short and medium-term load requirement forecasts. *Energy* **2020**, *209*, 118477. [CrossRef]
47. Das, A.; Annaqeeb, M.K.; Azar, E.; Novakovic, V.; Kjærgaard, M.B. Occupant-centric miscellaneous electric loads prediction in buildings using state-of-the-art deep learning methods. *Appl. Energy* **2020**, *269*, 115135. [CrossRef]
48. Wang, X.; Yuan, J.; You, K.; Ma, X.; Li, Z. Using Real Building Energy Use Data to Explain the Energy Performance Gap of Energy-Efficient Residential Buildings: A Case Study from the Hot Summer and Cold Winter Zone in China. *Sustainability* **2023**, *15*, 1575. [CrossRef]
49. Markus, M.; Breunig, H.-P.K.; Raymond, T.N.; Jörg, S. LOF: Identifying Density-Based Local Outliers. *ACM J.* **2000**, *29*, 93–104.
50. Ahlgren, P.; Jarneving, B.; Rousseau, R. Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *J. Am. Soc. Inf. Sci. Technol.* **2003**, *54*, 550–560. [CrossRef]
51. Zhou, Z. *Ensemble Methods: Foundations and Algorithms*; CRC Press: Los Angeles, CA, USA, 2012.
52. Zhou, Z. *Machine Learnng*; Tsinghua University Press: Beijing, China, 2016.
53. Khalil, M.; McGough, A.S.; Pourmirza, Z.; Pazhoohesh, M.; Walker, S. Machine Learning, Deep Learning and Statistical Analysis for forecasting building energy consumption—A systematic review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105287. [CrossRef]
54. Qiao, Q.; Yunusa-Kaltungo, A. A hybrid agent-based machine learning method for human-centred energy consumption prediction. *Energy Build.* **2023**, *283*, 112797. [CrossRef]
55. 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]

56. Moon, J.; Rho, S.; Baik, S.W. Toward explainable electrical load forecasting of buildings: A comparative study of tree-based ensemble methods with Shapley values. *Sustain. Energy Technol. Assess.* **2022**, *54*, 102888. [CrossRef]
57. Chen, Z.; Xiao, F.; Guo, F.; Yan, J. Interpretable machine learning for building energy management: A state-of-the-art review. *Adv. Appl. Energy* **2023**, *9*, 100123. [CrossRef]