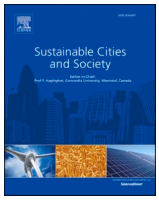
Sustainable Cities and Society 78 (2022) 103635

Contents lists available at ScienceDirect

Sustainable Cities and Society

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



An interpretable method for identifying mislabeled commercial building based on temporal feature extraction and ensemble classifier

Tong Xiao [a], Peng Xu [a,*], Renrong Ding [a], Zhe Chen [b]

[a] Department of Mechanical and Energy Engineering, Tongji University, Shanghai 201804, China
[b] School of Energy and Power Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

ARTICLE INFO

Keywords:
Primary space usage (PSU)
Mislabel
Temporal features extraction
Ensemble classification
Model interpretability

ABSTRACT

Proper building categorization is important in building energy efficiency analysis. Primary space usage (PSU) is a typical and widely used commercial building categorization method. The PSU labels are ascertained once the buildings are put into use but not always modified on time when the building usages change, which may lead to false results in analysis. In this paper, we propose a method to identify mislabeled commercial buildings based on analysis of the energy time series collected by electric meters. The method is constructed as follows: (1) data cleaning and transformation; (2) three types of temporal feature extraction; (3) several single classifier training, and the ensemble classifier building; (4) mislabel building identification and correction. The method provides a supervise way to identify mislabeled building. We applied the method to a public dataset from the Department of General Services from Washington, D.C. and found that 22.4% of the buildings were mislabeled. We also designed 1000 evaluation cases to prove the effectiveness of the method. Based on the results of the cases and the good interpretation of the method, we discuss the mislabeled buildings in reality and the temporal differences among different PSU-type buildings. We also discuss the renewal or improvement of PSU categorization.

1. Introduction

According to the annual report of the International Energy Agency (IEA), the overall energy intensity of the buildings sector is increasing during the Covid-19 crisis and the building sector has become the largest social energy consumer (Energy Efficiency 2020). Meanwhile, the building sector plays a key role in global CO2 emissions and significantly influences the way to global net zero (New Energy Outlook 2020). Building energy efficiency, especially that of commercial buildings, has become the primary target in energy saving (World Energy Investment 2020); thus, building performance analysis is vital. Building energy benchmarking, a type of building performance analysis for commercial buildings, is an important method to learn commercial building energy efficiency and usually serves in energy auditing and analysis of energy-saving scenarios (Zhan et al., 2020). Conventional building energy benchmarking aims to establish how much better or worse, a given building performs than its peer group; thus, the current energy benchmarking metrics are developed based on building categorization. A proper building categorization enables a reasonable energy performance comparison among similar buildings and thus assists in the realization of the real state of the energy efficiency of a building.

1.1. Building categorization

Primary Space Usage (PSU) is a typical commercial building categorization method and has been widely used in energy benchmarking surveys, including the Commercial Buildings Energy Consumption Survey (CBECS) in the United States Commercial Buildings Energy Consumption Survey (CBECS) 2021. Primary Space Usage refers to the classification of buildings according to their principal activity, and thus buildings are classified into different groups, such as education, office, public assembly, and lodging. Different PSUs of the area will have different contributions to the total energy consumption owing to the various occupancy schedules and equipment with variable energy intensity. For example, occupants of the office may work from 8 a.m. to 5 p.m. and use certain office equipment, whereas occupants of the retail store may work from 10 a.m. to 10 p.m. and use high energy intensity lighting; thus, the hourly energy of the office and the retail store will be different (Quintana et al., 2021).

The PSU method is believed to be useful in grouping buildings for benchmarking and is easy to obtain in practice. However, with the increasing diversity of use and loads in buildings, many buildings are of mixed-use type (Park et al., 2019); for example, some buildings may

* Corresponding author.
E-mail address: xupeng@tongji.edu.cn (P. Xu).

https://doi.org/10.1016/j.scs.2021.103635
Received 14 July 2021; Received in revised form 27 November 2021; Accepted 21 December 2021
Available online 25 December 2021

have commercial use on the lower floors and office use on the upper floors. Moreover, with the development of existing building transactions, the use types of buildings may change over time. Such scenarios provide new challenges for the building categorization such that the current PSU types recorded in buildings may no longer be no correct or accurate. The labels of the buildings need to be corrected over time, and some analyses are needed to determine whether the building categorization method should be renewed.

1.2. Previous studies

Researchers have realized the importance of building categorization and have completed research on it at the current state of building usage. Meanwhile, the wide availability of electric meter data enables researchers to learn the actual usage of a building by analyzing the energy consumption time series. Many building categorization studies have been conducted using the data-driven method by analyzing the electric meter data, which also called load profiling. Most of the studies were developed based on clustering (Wang et al., 2019), which is an unsupervised machine learning method. Benitez et al. (Benítez et al., 2014) used dynamic clustering to classify energy time series. Park, et al. (Park et al., 2019) identified the fundamental load shape profiles of buildings using several clustering methods in a large and diverse dataset and concluded that it is better at generalization than other studies. Zhan, et al. (Zhan et al., 2020) recategorized the buildings in the school based on operation quantification. The operation is quantified using several steps of k-means clustering.

The current study focuses on development of new categorizations, from inspiring the current building categorization method in use. However, fewer studies have been conducted to analyze how much the PSU label fits the description of current buildings. Mislabeled identification and the correction of the building label for the current buildings are seldom focused on. Researchers used to correct the building PSU labels by hand based on their engineering experience. Currently, machine learning is widely used in identification or detection scenario in engineering fields, such as detection of the solder paste defect (Sezer & Altan, 2021a,b). Mislabeled building may also be identified automatically using machine learning methods. Quintana, et al. (Quintana et al., 2021) applied clustering on the energy-time series to identify mislabeled buildings in two datasets. However, owing to the characteristics of the unsupervised method, the clustering results differ significantly from the current labels and are difficult to interpret. Carla, et al. (Brodley and Friedl, 1999) suggest that mislabeled data can be identified with supervised method, such as ensemble classifier. This idea has been recently tested in other fields. Feng, et al. (Feng et al., 2020) used ensemble classifier to identify the mislabeled sample in real dataset. Luengo, et al. (Luengo et al., 2021) used multiple instance classifier for mislabeled identification. However, ensemble classifier has not been used in mislabeled building identification.

Moreover, most of the current methods are based on time-series clustering and the data is normalized before clustering. Without normalization, buildings with the same energy use pattern but different energy intensities may not be assembled together (Lavin and Klabjan, 2015). However, building energy intensity is also an important feature in energy efficiency analysis, and thus should not be neglected in building categorization. Thus, in addition to the analysis of the original time series, some researchers have considered extracting temporal features from the time series for energy analysis. Temporal features are the aggregation of the behaviors exhibited in the time series data (Miller and Meggers, 2017) and have been used in several time series forecasting studies to help improve forecasting accuracy (Karasu et al., 2020). The statistical characteristics of the time series over a period such as mean, max, min are used in some studies to enrich the input variables (Grolinger et al., 2016). Miller et al. (Miller and Meggers, 2017) extracted many temporal features and divided the extracted features into three categories: statistics-based, regression model-based, and pattern-based. The statistics-based features can be calculated using statistical operations, whereas the regression model-based features are extracted from the difference between the load prediction model and the real data. Pattern-based features were extracted to describe the daily use pattern of the building. Najafi et al. (Najafi et al., 2021) applied several state-of-art feature selection methods to the temporal features and determined the most influential features for energy prediction. Proper temporal features can assist in the reduction of the computation time compared with the analysis of the original time series, while simultaneously retaining the compositions that are useful for energy analysis.

Meanwhile, as data-driven models are widely used in current studies, model interpretability is becoming important. Building professionals cannot fully trust data-driven models without enough sufficient interpretability (Fan et al., 2021). This is because researchers cannot learn sufficient knowledge on the workability of the data-driven model only by analyzing the accuracy metric. Model interpretability can assist researchers to understand what the model has learned from the data and whether there are some mistakes in the model application (Doshi-Velez and Kim, 2017). Methods for machine learning interpretability can be classified as intrinsic or post hoc (Molnar, 2019). Intrinsic interpretability refers to the use of models that are considered interpretable owing to their simple structure, such as short decision trees (Lipton, 2016). Post hoc interpretability refers to the application of interpretation methods after model training. Fan et al. (Fan et al., 2019b) develop a method to interpret building energy prediction models and the method is based on an interpretation model called the local interpretable model-agnostic explanation (LIME) (Ribeiro et al., 2016a; Ribeiro et al., 2016b). Although some studies included model interpretability analysis, many types of research did not, and thus the results are somehow not convincing.

1.3. Aim and objectives

Based on a review of the previous studies, we conclude the current gap as followed.

- Use clustering to find out new categorization.
- Always lose temporal features, such as energy intensity.
- Seldom focus on automatic mislabeled building identification.
- In need of method with a good interpretation.

Thus, we develop a method to identify mislabeled commercial buildings and the method we developed will have the following characteristics.

- The method is developed based on temporal feature extraction and attempted to maintain temporal features, such as energy intensity.
- The method is developed based on supervised learning and aims to identify mislabeled building based on the current building categorization method.
- The method is developed with a good interpretation to face practical engineering problems.

With the above study, we can identify the mislabeled building from a set of buildings more convincingly in an interpretable manner than in previous studies. The temporal features retained sufficient information of the time series in the analysis. The framework based on supervised learning makes the results closer to the existing classification situation. Thus, the method is more straightforward for misclassification recognition tasks. We apply our method in a real case and perform an evaluation. We believe that our findings will significantly assist in energy analysis in practice.

With the method we developed, we also discussed several problems to help obtain a better knowledge of the real world and attempted to make some suggestions for future studies.

- How is the current state of the mislabel of buildings?
- Should our categorization be renewed? How can we improve current building categorization?

The remainder of this paper is organized as follows. We provide a detailed description of our developed method in Section 2 and apply it to a real dataset as a case study in Section 3. We evaluate both on the features and the identification accuracy to make our model more convincing in Section 4. We analyze the results of the case study in Section 5 and attempt to answer the two questions we mentioned above. We also discuss the potential future applications and limitations of this study in Section 6 and conclude the paper in Section 7.

2. Methodology

The framework of this study is presented in Fig. 1. The framework can be divided into two parts, mislabel identification and model interpretation. This section focuses on the methods used for mislabel identification. First, we discuss how the data is processed for the following tasks; subsequently, we focus on what and how the features are extracted from the data. In addition, we describe how the ensemble classifier is built to identify the mislabeling.

2.1. Data pre-processing

The measured data obtained from the electric meter may have some problems such as outliers and are always of a different structure from the machine learning model input (Fan et al., 2015). Data pre-processing aims to obtain clean data suitable for the following procedure, as shown in Fig. 1. The major tasks in data pre-processing include data cleaning and data transformation.

2.1.1. Data cleaning

The data cleaning procedure in this study aims to remove the following types of outliers: (1) the extreme big/small values; (2) the values with extremely large front/back first-order differences in the time series; and (3) the constant values in the time series. Extreme big/small values can be visualized in box plots and detected by calculating the inter-quartile range (IQR) (Aggarwal, 2015). The values with extremely large front/back first-order difference denote values that significantly differ from the value at the time before or behind, and are always caused by the sudden abnormal state of the sensors. Such values can be determined by first calculating the front/back first-order differences and then calculating the IQR of the differences. We used the IQR for difference outlier determination. IQR is an anomaly detection method developed based on the data distribution. Mutations caused by changes in system operation are normal in the data distribution, and thus will not be considered anomalies. The constant values can be determined using the constant value detector function in tsod, a Python package (API Reference 2021).

2.1.2. Data transformation

The data transformation procedure in this study refers to the transformation of data into three types and thus aims for temporal feature extraction. The three transformation types in this study are basic data transformation, area-normalized data transformation, and static-normalized data transformation.

a. Basic data transformation

Basic data transformation aims to transfer raw data into a suitable structure for temporal feature extraction. Here, hourly electric meter data are required for the following procedure. No normalization is performed on the basic data, and thus, the basic data can represent the real hourly energy intensity of the entire building.

b. Area-normalized data transformation

Area-normalized data transformation refers to the normalization the basic data with the building area. Thus, the area-normalized data can indicate the hourly intensity per area of the building. The area-normalized data is logical because the areas among buildings are always different; thus, the basic data are not sufficient for comparison. Accurate building areas are required for this transformation.

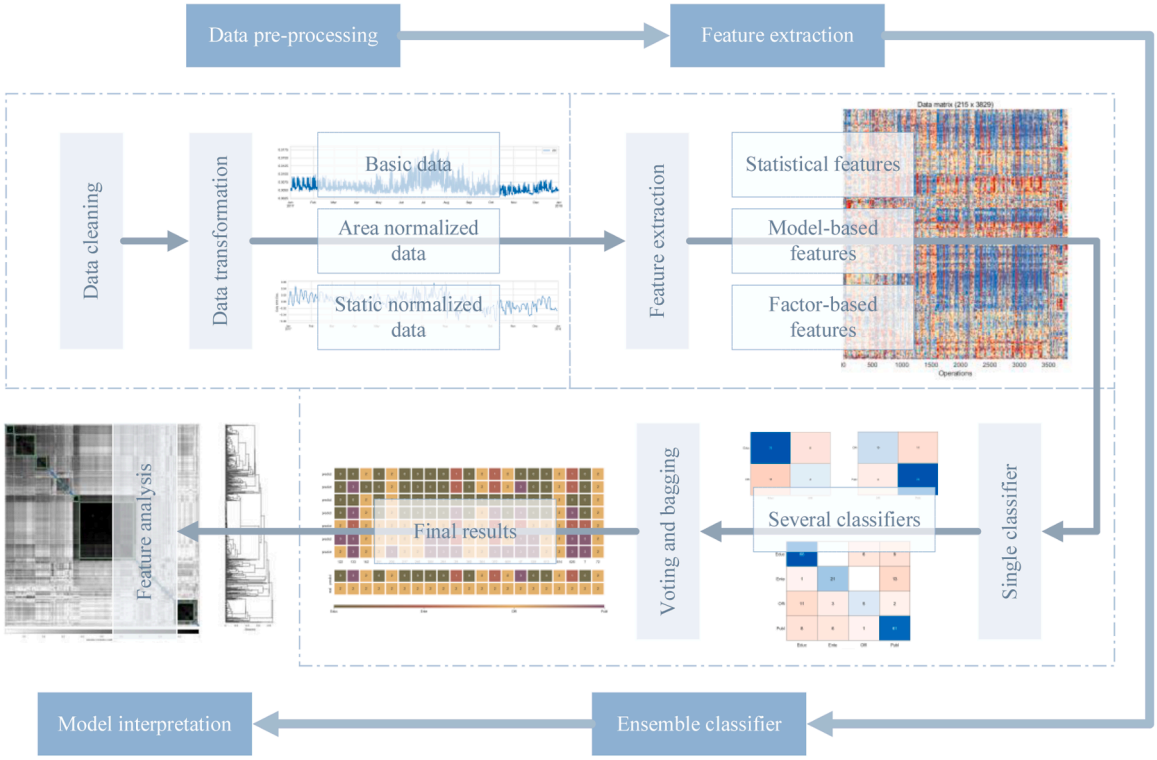


Fig. 1. Overview of the framework.

c. Static-normalized data transformation

Static-normalized data transformation attempts to remove the weather-dependence load. The large component of the weather-dependent load in a building is the heating/cooling load caused by conduction through the building envelope and air infiltration/ventilation (Kelly Kissock and Eger, 2008). The data after removing the weather-dependent load represents the energy cost by the occupancy and operations. The weather-dependent load was calculated using the multivariable change-point model in this study. The weather-dependent load consists of three components, the base load, the cooling load, and the heating load. Base load is the load that is not related to the outdoor temperature, but is required for the regular operation of the building. The multivariable change-point model is a piecewise linear model and can be calculated using Eq. (1) and Eq. (2).

$$E_c = \beta_1 + \beta_2(T - \beta_3) \quad (1)$$

$$E_h = \beta_1 - \beta_2(\beta_3 - T) \quad (2)$$

Where E_c/E_h represents the heating/cooling load, and T represents the outdoor temperature. β_1 is the temperature-independent base load of the building. β_3 is the balance temperature at which the building does not require heating/cooling. β_2 is the rate of heating/cooling energy increase owing to the outdoor air conditions.

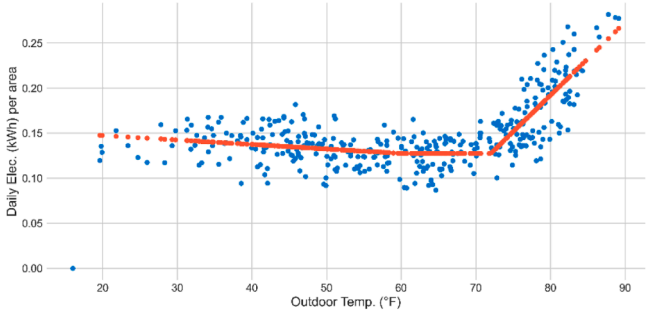
Fig. 2 shows the application of the multivariable change-point model and static-normalized data. As shown in Fig. 2(a), we first resample the raw data to daily data and then train the multivariable change-point model. We predict the base load and the cooling/heating weather-dependent load with the model we trained and then we subtracted the predicted load from the real usage for the static-normalized data. Fig. 2 (b) shows the predicted load and the real usage, and Fig. 2(c) shows the plots of the static-normalized data.

2.2. Feature extraction

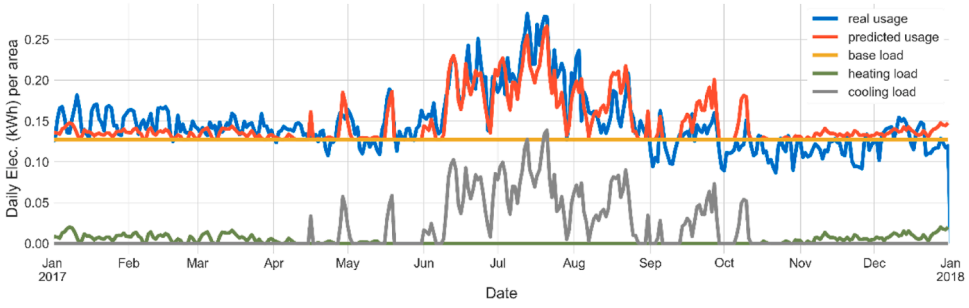
Using the transformed data, we extracted three types of features for the analysis. The three categories of the features are statistical features, model-based features, and factor-based features. The features are grouped based on how they are extracted. Features extracted via similar methods may represent a similar characteristic of the time series, and it will be addressed in the following procedure.

a. Statistical features

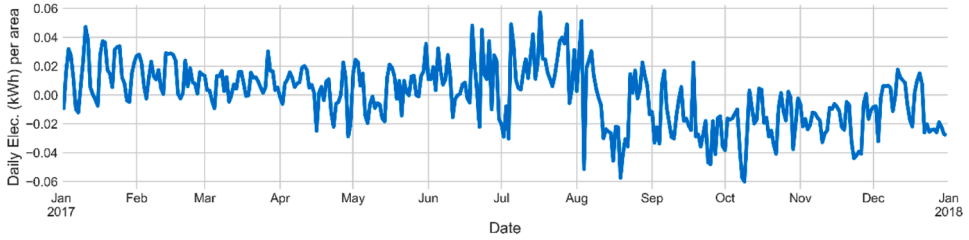
Statistical features are extracted from the time series via basic mathematical operations. In this study, we calculated the maximum, minimum, mean, and different quarter values of the time series over



(a) Multi-variable change-point model



(b) The real usage and the predicted weather dependence load



(c) The static-normalized data

Fig. 2. Multi-variable change-point model and static-normalized data.

several time ranges (season, month, week, and day). These features simply represent the energy distribution of one building. We focused on some days and day types because they are meant to analyze the load of the HVAC system. For example, the energy cost on the hottest day of the year is important to understand the cooling energy required for a building.

We also calculate the energy ratio over time to learn more about the difference in energy costs at different moments. Such features represent the energy shape of a building. The energy shape of a building is important because a building always uses energy at specific moments and based on schedule, and different use types of buildings may have significant differences in energy shape (Mathieu et al., 2011). For example, the energy ratio of the day and night for an office building can provide the energy difference between occupancy time and inoccupancy time, and the ratio will differ from that of retail stores.

b. Model-based features

We use some developed models and methods to help us extract some features from the time series and to call such features model-based features. The models and methods used for extraction are the STL method, breakout detection, and DayFilter model.

Seasonal and Trend decomposition using Loess (STL) is a classical time series decomposition method that can decompose a seasonal time series into three components: trend, seasonal, and the remainder (CLEVELAND, 1990). The building energy time series is always in strong seasonality, and their decomposition can provide quick and useful insights into the operation of the buildings (Pickering et al., 2018). In this paper, we decompose the time series in hourly time granularity and regard the day as a seasonal cycle.

The breakout detection model was proposed by Twitter, which can detect the mean shift and ramp-up breakouts in the time series (Vallis et al., 2014). The breakout detection model has been used in several building energy time series analysis (Pickering et al., 2017). Mean shift breakouts are the sudden changes from one state to another, whereas ramp-up breakouts change smoothly. Breakout detection of the building energy time series is important because there are some physical operations responsible for the breakouts. For example, the control logic of the HVAC system may change between the cooling and heating seasons, causing a breakout in the time series.

DayFilter is a day-typing model based on the symbolic aggregate approximation (SAX) method, and it can recognize the typical daily patterns of a building (Miller et al., 2015). SAX can reduce a time series of arbitrary length to a string of arbitrary length, thus help reduce the complexity of the analysis of the time series (Lin et al., 2003). With different window and alphabet sizes, the energy time series is transformed into SAX strings and thus can be easily grouped into several daily patterns. We use a set of window and alphabet sizes for transformation and then extract the pattern count and the most frequent daily patterns for each transformation.

c. Factor-based features

The factor-based features in this study describe the relationship between the time series and its influencing factors. The outdoor temperature is well known as an important factor that influences the building energy and is the primary focus of this study. Correlation analysis helps us understand the relationship between building energy consumption and influencing factors, such as outdoor temperature (Yu et al., 2013). Thus, we applied the Spearman correlation calculation on the outdoor temperature and the building energy consumption in this study and then extracted some statistical features such as mean, maximum, minimum, and standard deviation from the calculation results.

2.3. Ensemble classifier

As mentioned in the introduction, the correction of the building label task was similar to that of a mislabel identification task. The samples that be classified into different labels from their current labels may have a higher probability of mislabeling than those that are well predicted. Thus, the prediction accuracy of the classifiers is not that important in this study and the falsely predicted samples are what we mainly focused on. In this study, we built an ensemble classifier base on some single classifiers for mislabeled classification. The idea of the ensemble method is to build a predictive model by integrating several models and achieving better accuracy than a single model (Rokach, 2010). The samples that are falsely predicted by the ensemble classifier are more convincing to be the mislabeled samples than those falsely predicted by the single classifier. Furthermore, because the falsely predicted samples are reasonable only when the classifier can accurately predict the building label. Some evaluation on whether the classifier can recognize mislabeled building is needed and will be presented in Section 4.

a. Single classifier

The single classifier used in this study was developed using the support vector machine (SVM). SVM addresses the classification problem given by n data records $\{(X_i, y_i), i = 1, 2, \ldots n\}$, where X_i is the i th element of the n-dimension vector, that is, $X_i = \{x_1, x_2, \ldots x_n\} \in R^n$ and $y_i \in R$ is the aim label corresponding to X_i (Chen et al., 2017). SVM is an algorithm that is widely used in building energy analysis, such as forecasting of electricity load, electricity consumption (Amber et al., 2018). Because of its ability to handle nonlinear problems despite the high dimensionality of the data (Sun et al., 2020), SVM has been proven to achieve a good prediction accuracy in many cases (Foucquier et al., 2013).

b. Ensemble method

The ensemble method used in this study is bagging, and the ensemble workflow is shown in Fig. 3. In the bagging method, each classifier is trained on a subset of samples taken from the training set (Rokach, 2010). In our study, we divided the dataset into several groups based on their current labels and then randomly choose several groups as the training set to train each classifier. As the building energy dataset does not always include many samples for each label, such a method for dataset division and choice will enhance the prediction capability of every single classifier by ensuring sufficient input samples of each label.

Thus, the classifiers are combined with weights to build up the ensemble classifier. The weighting method used in this study is adapted from the majority voting, also known as the basic ensemble method. The majority voting means that for a given sample, all classifiers vote their class, and the ensemble classifier will predict according to the class that obtains the highest number of votes (the most frequent vote) among all the classifiers (Rokach, 2010). In this study, different classifiers will have a different label count owing to the dataset division method, and thus the vote weights for classifiers should be different. We calculate the vote weight of label A for classifier i as in Eq. (3), and then the total vote for label A using Eq. (4). The final predicted label for the ensemble classifier was calculated using Eq. (5).

$$v_i^A = \begin{cases} m, \ if \ L = A \\ 0 \end{cases} \tag{3}$$

$$V^A = \frac{\sum_i^n v_i^A}{n} \tag{4}$$

$$L_{final} = \text{argmax}\left(V^A, V^B, \ldots\right) \tag{5}$$

Where v_i^A indicates the vote of the predictor i; L means the predicted label and A means the real label of the building in the data subset; m indicate the label count in the data subset. n indicates the total count of the classifiers. And L_{final} is the final predicted label for the ensemble classifier.

3. Case study

We applied the method on a public dataset as a case study to determine mislabeled buildings in the dataset. Our computing device is a 3.8 GHz eight-core Intel Core i7 processor with 32 GB of RAM 3200 MHz DDR4. And we are using macOS 10.15.7. The feature extraction process

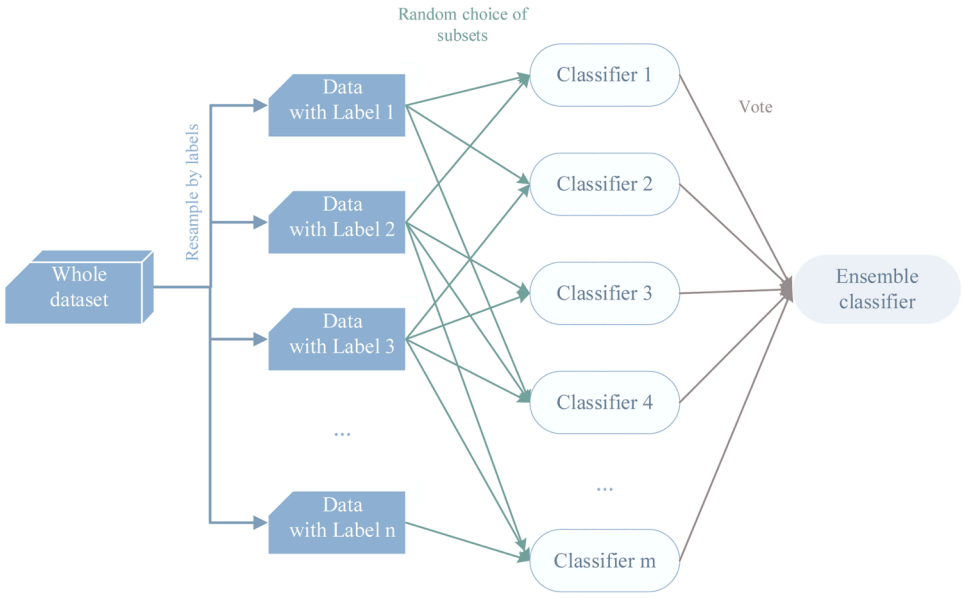


Fig. 3. The ensemble workflow in this paper.

is programmed in python, while the classifier is programmed in MATLAB. The model application and the results of the case study are described as follows:

3.1. Dataset

The dataset used in this study is from the Department of General Services of Washington, D.C. (Commercial Buildings Energy Consumption Survey (CBECS) 2021). This dataset is an electrical meter dataset, and the data are in a 15-minute interval. We choose one-yearlong data and relabeled the building types in wider groups according to the PSU categorization (Zhan et al., 2020). Building types without sufficient samples are removed; otherwise, they will function as the imbalance group in the classification and are not good for prediction. An overview of the dataset is provided in Table 1.

3.2. Data pre-processing and feature extraction

We deleted some buildings with energy time series in poor data quality, and then 228 buildings of 4 PSU types remained after processing. Fig. 4 shows the PSU types and the number of buildings in each type in the processed dataset. After data cleaning and data transformation, we obtained 3 smooth time series for each building. Subsequently, we extracted 4189 features from the time series for each building using the methods mentioned in Section 2.2.

3.3. Ensemble classifier and results

We trained 10 classifiers for each type of PSU-type combination among the 4 PSU types in processed dataset. The 4189 features were filtered before each training, and the features that had similar values over all samples were reduced to one. Fig. 5 shows the confusion matrix

Table 1
Dataset details.

Properties	Dataset (origin)	Dataset (processed)
Number of buildings	322	215
Date range	2016–02–02 – 2018–03–02	2017–01–01 – 2018–01–01
Number of days	523	365
Number of building types	22	4

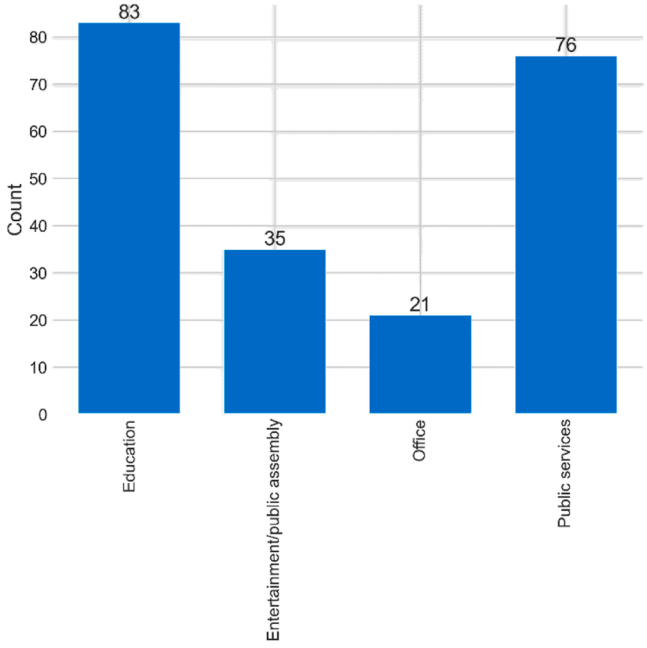


Fig. 4. Primary-Space-Usage (PSU) label distribution in the processed dataset.

of one of the ten classifiers, and the classifier shown in Fig. 5 uses the data with the education and public service PSU label.

The ensemble classifier is built using the bagging method as mentioned in Section 2.3. The training process of 10 basic classifiers and the ensemble classifier together with the visualizations of training output took 1.6288 s in total. Fig. 6 shows the prediction results (marked as predictions) of the 20 samples of each PSU label. The prediction results of the ensemble classifier (the second line from the bottom) and those of the basic classifiers (marked with the classifier name) are provided. Each column represents one building, and the row marked as real indicates the PSU label of the building in the dataset. For clarity, we will call the PSU label of the building in the dataset the real label. By comparing the prediction results of the ensemble classifier and the real label, we can determine whether the building is mislabeled.

The total rate of buildings that were predicted to be mislabeled by

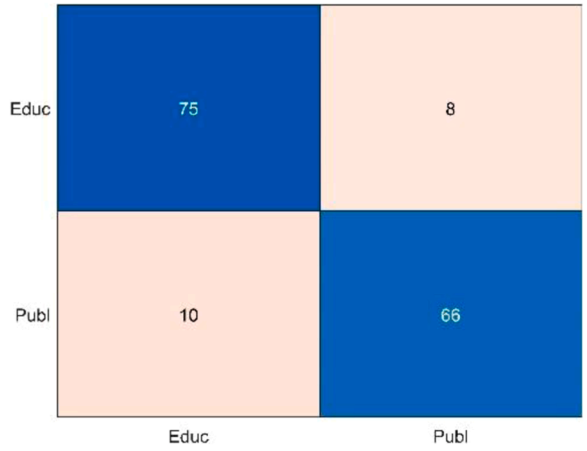


Fig. 5. The confusion matrix of the classifier on PSU label Education and Public services.

the model is 22.4%. Fig. 7 shows prediction results for each real label. There are 14% of the education buildings suggested as mislabel buildings, whereas there were 40% in entertainment/public assembly, 71% in office and 20% in public service. The suggested PSU labels are shown in Figure 7 (a) and the predictive mislabel rates are shown in Figure 7 (b).

The PSU type prediction results for each real PSU type (in the figure, Educ indicates Education, Ente indicates Entertainment/public assembly, Offi indicates Office, Publ indicates Public service. The number in each color section in (a) represents the number of buildings predicted as such PSU. The number in the blue color section in (b) represents the percentage of buildings predicted as the same PSU label as the real label. The number in gray color section in (b) represents the percentage of buildings predicted as the different PSU label from the real label.)

4. Model evaluation

As mentioned in Section 2.3, an evaluation of whether the classifier can recognize mislabeled buildings is needed. Thus, we first analyze whether the features we extracted can describe the PSU types and be useful for classification by random feature test. Secondly, we design a case in this study to evaluate whether the model could recognize mislabeled buildings.

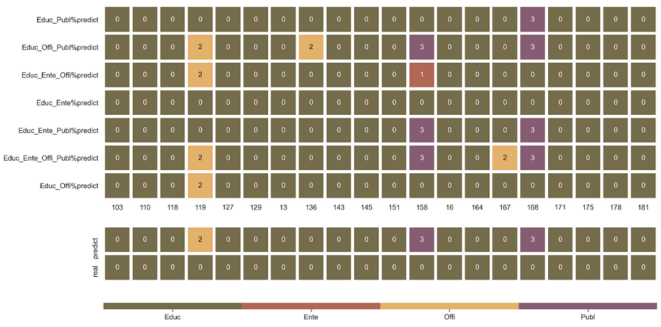
4.1. Random feature test

In a random feature test, we randomly choose values for features before inputting them to train the classifier. Such a test can simply determine whether the extracted features are useful for classification. If the classification accuracy with the random features is nearly the same as that with the real extracted features, the extracted features can be regarded as meaningless for classification. We perform random feature tests on every single classifier to determine whether the extracted features can describe the PSU labels.

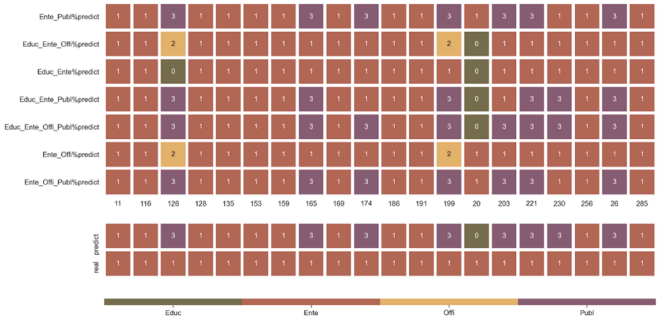
Fig. 8 shows the results for the classifier that includes all the four PSU labels. In the figure, the bars and lines in blue indicate the results calculated with the extracted features while the extracted features, whereas those in red indicate the results calculated with random features. The x-axis of the figure shows the prediction accuracy of the classifier with a single feature, and most of the extracted features have higher accuracy than the random features. Thus, most of the features we identified were important for the classification task.

4.2. Evaluation case study

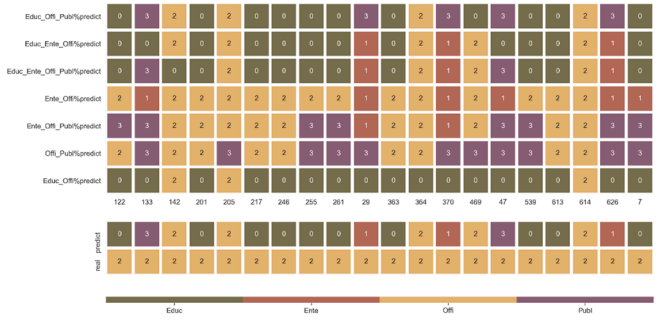
The evaluation case is designed to prove that the buildings suggested as mislabel buildings by the model are those behaving differently from



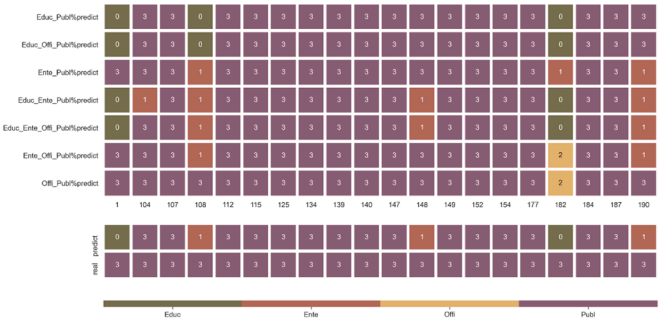
(a) 20 samples of the building in Education label



(b) 20 samples of the building in Entertainment/public assembly label



(c) 20 samples of the building in Office label



(d) 20 samples of the building in Public service label

Fig. 6. The prediction results of several samples for each PSU label (in the figure, Educ indicates Education, Ente indicates Entertainment/public assembly, Offi indicates Office, Publ indicates Public service).

their PSU types. Thus, we randomly choose several buildings and provided false labels before inputting the model to test whether they could be identified. Such tests were performed 1000 times in this study, and Table 2 provides a brief description and an example of these tests.

Fig. 9 shows the prediction results of the evaluation case example No.965, and a brief description of this case is presented in Table 2. In the figure, each column indicates a building, and the prediction PSU label,

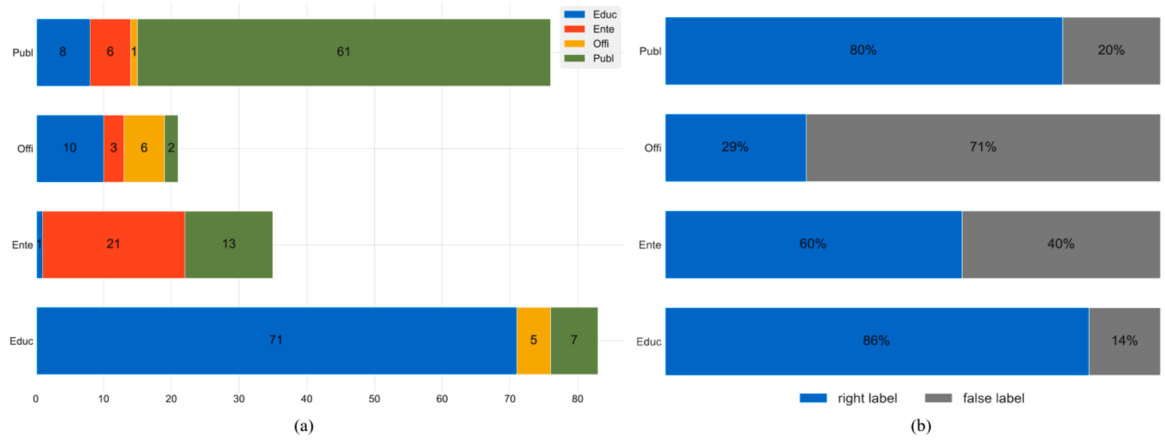


Fig. 7. . The PSU type prediction results for each real PSU type (in the figure, Educ indicates Education, Ente indicates Entertainment/public assembly, Offi indicates Office, Publ indicates Public service. The number in each color section in (a) represents the number of buildings predicted as such PSU. The number in the blue color section in (b) represents the percentage of buildings predicted as the same PSU label as the real label. The number in gray color section in (b) represents the percentage of buildings predicted as the different PSU label from the real label.).

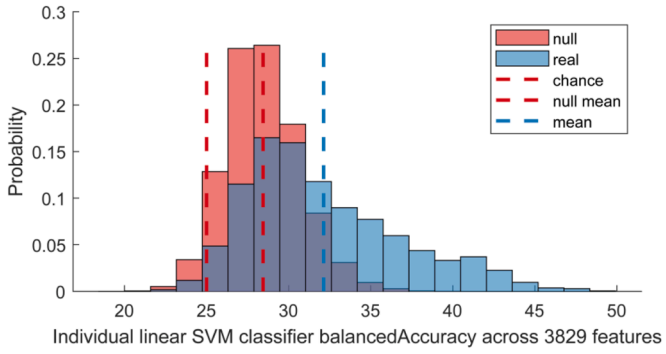


Fig. 8. . Random feature test results for a single classifier.

Table 2
Evaluation case details.

Properties	Evaluation case	Example
Evaluation count	1000	No. 965
Manual mislabel count	1 - 46	45
Mislabel rate in a single label	0% - 46.8%	22.86% in Educ
	(random choice for label)	38.46% in Ente
		16.67% in Offi
		26.32% in Publ

the real label, and the manual mislabel are listed in the bottom three rows. Because the label is predicted with classification, the label indicates that the building is then similar in behavior to the rest of the buildings tagged with that label. Buildings with the same label are more likely to be in the same use type. If the prediction label differs from the manual mislabel, this means that the mislabel is determined by the model. We define recognition as the proportion of the mislabeled buildings recognized by the model. Accuracy is defined as the proportion of mislabeled buildings that are predicted right by the model. Because the developed model in this study focuses on the mislabeling in the dataset, the situation of whether the prediction label is the same as the manual mislabel will be addressed in the following analysis.

To discuss the prediction results more clearly, we define two evaluation metrics to describe the results. These two evaluation metrics are adapted from, the recall and precision metrics, for binary classification. Recall is the proportion of real positive cases that are correctly predicted as positive, and precision denotes the proportion of predicted positive cases that are correctly real positives (Powers, 2010). In this study, we adapt Recall as the proportion of the mislabeled building that are correctly recognized, and Precision as the proportion of the recognized mislabeled buildings that were actually mislabeled. Thus, Recall indicates whether the model can recognize all the mislabeled buildings, and Precision indicates how many of the buildings that suggested mislabeling by the model were mislabeled. It should be noticed that Recall has the same meaning as the recognition earlier mentioned, and it is the main metric of focus.

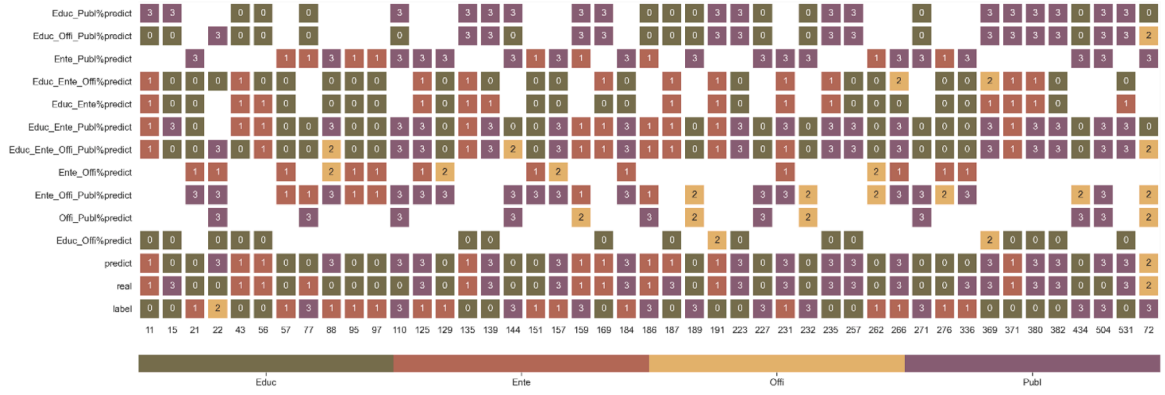


Fig. 9. . The prediction results of case No.965 (in the figure, Educ indicates Education, Ente indicates Entertainment/public assembly, Offi indicates Office, Publ indicates Public service. Predict indicates the prediction label, real indicates the PSU label in the dataset, and the label indicates the manual mislabel).

Table 3 and Fig. 10 lists the results of the evaluation tests and all the values are mean in the group. Because the buildings were randomly chosen for mislabeling, we obtained 3852 groups of metric results for every 4 PSU types from 1000 tests. As the mislabel rate increased from 0 to 46.8%, Recognition maintains a stable and high value of approximately 95% such that the model could recognize the mislabel buildings accurately among the mislabel rates in the evaluation tests. The accuracy decreases when the mislabel rate increases; however, it remains at approximately 80%. It seems that the model can correct the building's label when the mislabel rate is not more than 46.8%, and the correction accuracy decreases when the mislabel rate increases.

In summary, with the evaluation cases, we proved that our model can recognize the mislabeled building in the dataset and can provide important suggestions on which PSU label the mislabeled building should belong to. The results in the evaluation cases also indicate that model have uncertainty from input data. Specifically, it means that when the mislabeling rate in a certain original labeled building is too high, it makes the uncertainty of the model increase and the recognition accuracy decrease. The mislabeled buildings determined by the model may not be mislabeled in reality. However, the recognition is still logical because it is more important to complete the search in the mislabel recognition task. Further analysis of buildings identified as mislabeled by experts can determine whether each building is mislabeled. Our model provides a set of possible solutions that significantly reduces the workload of expert analysis.

5. Result analysis and interpretation

In Section 3, we introduce the application of our model to a dataset that is collected from the real world, and we prove in Section 4 that our model can recognize and correct the mislabeled building. Thus, the buildings that were suggested to be mislabeled by the model in the case study in Section 3 are worth attending to and will be discussed in this section. Meanwhile, as our model is structurally transparent and easy to interpret, some interpretations of the differences among different PSU labels in the dataset in the case study will also be presented.

5.1. Mislabel buildings in the real world

As shown in Fig. 6 and Figure 7, 22.4% of the buildings were predicted to be mislabeled by the model, and buildings with all PSU types had some mislabeled cases. There is a certain amount of mislabeling of buildings under each PSU label and may cause incorrect energy analysis based on such mislabels. Thus, the identification of mislabeled buildings is essential before building performance analysis.

The correction of building PSU labels shows the common mislabeling situations in reality. We may determine the mislabeled buildings with the Education label are predicted by the public service and office, but not entertainment. This indicates that some buildings built with education usage may be used for public service and offices in reality but seldom used for entertainment. Such results are in line with our perceptions because we seldom see that education buildings are used for entertainment. Similar results were also found in entertainment/public assembly buildings. Buildings labeled as entertainment may be used as a public service but seldom used as education or office. Mislabeled public service buildings may mainly be used for education and entertainment.

The number of buildings with Office labels, predicted to be mislabeled stands at 71%., which indicates that the actual office buildings have numerous uses that do not match the label. Although the mislabel rate, here, is high and thus the correction results are not sufficiently convincing, the high mislabel rate is still important in suggesting that the Office label may not be accurate to describing the building usage. The Office labels may need further refinement, to describe the building usage more accurately.

Meanwhile, some buildings may have mixed usages, according to our results. As shown in Fig. 6, some buildings are predicted as different labels from the real label in every basic classifier, such as building 126 in Fig. 6(b), building 370 in Fig. 6(c), and building 182 in Fig. 6(d). Such buildings may have multiple use types simultaneously, and thus their energy time series have similar features to the multi-type buildings. With the high transparency of the ensemble classifier, we can calculate the number of buildings that may have mixed usages by analyzing the predicted label given by every single classifier. There are 32 buildings among 215 buildings that may have mixed usages, which indicates that buildings with mixed usages are common in the current state. Thus, more studies on the description of mixed-usage buildings are needed.

Table 3
Evaluation results.

Mislabel rate in a single label	Recognition/Recall (mean)	Accuracy (mean)	Precision (mean)
(0.0131, 0.059]	0.9608	0.9147	0.4415
(0.059, 0.104]	0.9359	0.8545	0.5154
(0.104, 0.15]	0.9517	0.8664	0.5799
(0.15, 0.195]	0.9795	0.9412	0.3578
(0.195, 0.241]	0.9338	0.8300	0.6200
(0.241, 0.286]	0.9249	0.8153	0.6266
(0.286, 0.332]	0.9384	0.8132	0.6294
(0.332, 0.377]	0.9317	0.7947	0.6122
(0.377, 0.423]	0.9400	0.7791	0.6511
(0.423, 0.468]	0.9692	0.7933	0.6150
Total	0.9572	0.8909	0.4748

5.2. The temporal difference among PSU labels

To study the influence of building usage in different PSU labels on the energy time series of the building, we analyzed the influence of the individual input features on classification (Miller, 2019). Moreover, to remove the influence of the high correlative features, we use hierarchical clustering to determine the high correlative features and choose the cluster centers as a representation of clusters (Fulcher and Jones, 2017). To reduce the amount of calculation, we use 500 features with high prediction accuracy in a single feature classifier for clustering.

For each binary classifier, we analyzed the top 10 cluster center features and the results are listed in Table 4 (in appendix) and Fig. 11. The major difference between education buildings and entertainment buildings is the different energy consumption in a week. Education buildings may have higher energy consumption during weekdays than weekends, whereas the reverse is true for entertainment buildings. The energy used in education buildings for several months also differs from that in entertainment buildings. Education buildings have different occupancy schedules in a year, which differs significantly from that of entertainment buildings. Meanwhile, more equipment in the education buildings leads to a difference in the cost of cooling energy compared to entertainment buildings.

To the best of our knowledge, education buildings and office buildings may have similar occupancy schedules during a week. In this study, we found that the energy consumption of education buildings on Friday differ from that of office buildings. Meanwhile, the energy intensity also differs between the education buildings and the office buildings.

The major difference between education buildings and public service buildings is the energy consumption in different day types during a week. The features in Error! Reference source not found. indicate these differences. This difference is easy to be interpreted with our common knowledge. The occupancy schedule of the education buildings differs from that of public service buildings and thus causes different energy patterns.

The energy consumption on Monday, Friday and weekends of the entertainment buildings differ from those of the office buildings because of their different occupancy schedules. In addition, the energy intensity differs between entertainment buildings and office buildings.

Entertainment buildings and public service buildings have a

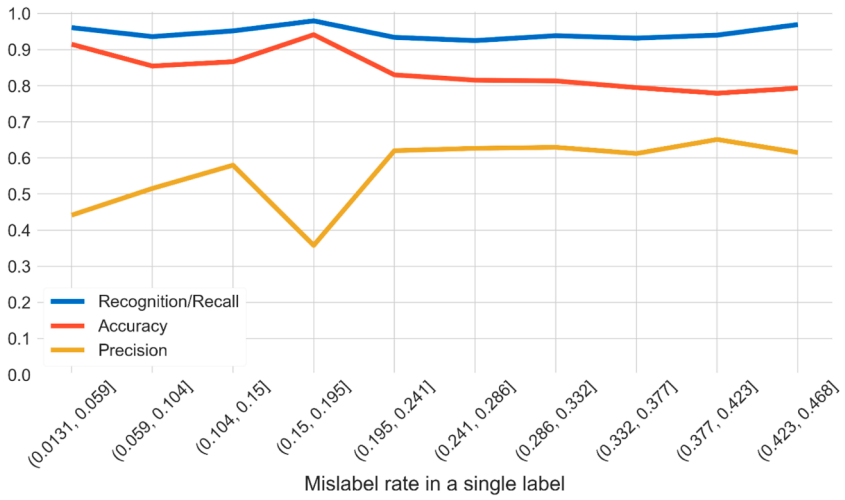


Fig. 10. . Evaluation results.

difference in energy consumption without the influence of the weather. Such energy consumption is contributed by equipment and occupancy in the entertainment buildings and public service buildings. In addition, the occupancy schedules of the entertainment buildings and public service buildings are indifferent and thus cause different energy consumption in specific day types during a week.

The major difference between office buildings and public service buildings is the energy consumption in different day types during a week. This is easy to interpret because two types of buildings have different occupancy schedules.

With the analysis of the features, we studied the temporal differences among buildings of different PSU types. We proved that the difference is not only in the energy patterns but also in the energy intensity; for example, education buildings and office buildings have different energy intensities owing to the different equipment and occupancy intensity. This may indicate that pattern analysis based on time series normalization and simple clustering is not sufficient for building categorization. Energy intensity plays an important role when benchmarking the energy efficiency of buildings among a group of similar buildings; thus, it should not be ignored in building categorization and related studies.5.3 Building PSU label revisited

With the method we developed in this study, we studied the current state of mislabeling. Thus, another question arises: should our categorization be renewed? How can we improve current building categorization? Although the PSU label has been widely used, it may require modifications to meet the current state. As discussed in Section 5.1, the rate of mislabeling of the office building is high in our case study and indicates that the office buildings, in reality, have different energy performances from each other. Office buildings in the real world may have multiple semi-types as the work types in the office building vary and may have different work schedules and equipment needs. Moreover, we found that 32 of the 215 buildings may have mixed usage in our case study. These results indicate that mixed usages is common in reality. Thus, further refinement for building labels is required and thus can describe the building usage more accurately.

The analysis of the temporal features provides temporal evidence for PSU building characterization. Buildings in different PSU labels have a temporal difference in their energy performance, and the PSU labels can be ascertained by temporal analysis. With the analysis, we also found that building PSU labels can indicate the main occupancy schedule types and energy intensity levels of the building, and thus can be used for building benchmarking.

In conclusion, the current PSU labels are mostly suitable for many cases; however, they require supplementation in the current state of mislabeling. Supplementation includes the definition of some mix-use types and the refined definition of office buildings.

6. Discussion

As shown in the results, several buildings were identified by the mislabel identification method as mislabel buildings, and their PSU labels were corrected in the case study. In the evaluation section, we proved the capability of mislabeling identification and label correction of the method. In our proposed method, the temporal features retained sufficient information of the time series in the analysis. The framework based on supervised learning makes the results closer to the existing classification situation. Thus, the method is more straightforward for misclassification recognition tasks.

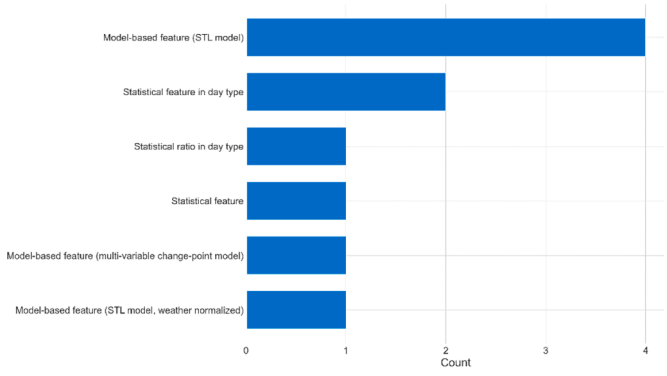
The method proposed in this paper can identify mislabeled buildings in the dataset and thus help improve the categorization accuracy. Using the method before building benchmarking can ensure that buildings of similar use types are analyzed together. Thus, we can understand the energy efficiency state of the building more accurately by comparing similar buildings. The proposed method can be used as an essential process before building energy benchmarking to help study the real state of energy efficiency. As shown in Section 4.2, the model performance mainly depends on the input data. When the mislabeling rate of the input data is too high, the uncertainty of the model will increase and the recognition accuracy will decrease. Further analysis of buildings identified as mislabeled by experts is needed at that time to help determine whether each building is mislabeled. With the set of possible solutions provided by our model, the workload of expert analysis will significantly reduces.

Furthermore, as shown in Section 5.2, the method has high model transparency and is easy to interpret. Our method opens the black box of the building PSU label and helps us understand how buildings with different PSU labels behave differently in their energy time series. When we apply the method to a larger dataset, we can analyze the difference among more PSU labels and thus determine the representative temporal features and their distributions for each type. The representative features and distributions can broaden our understanding of energy consumption in the real world and thus will be helpful for future building performance analysis and simulation studies.

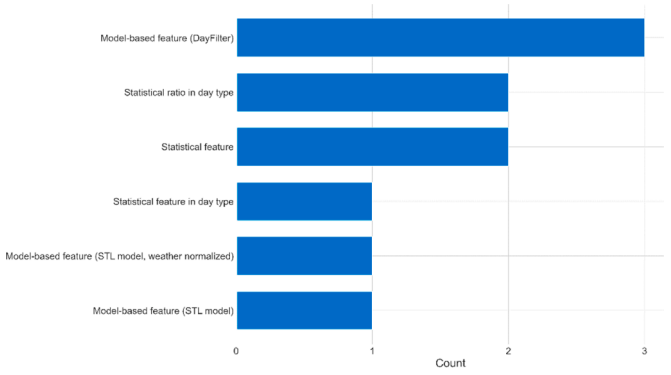
The temporal features extracted using this method may have wider usage in energy time-series analysis. As an increasing amount of energy meter data from buildings become available, energy time series analysis is becoming an important method to study the real state of the energy cost in buildings. Analysis with temporal features reduces the complexity of time-series analysis. The temporal features extracted in this study are easier to understand with our domain knowledge than the

Table 4
The top 10 cluster center features in binary classifiers.

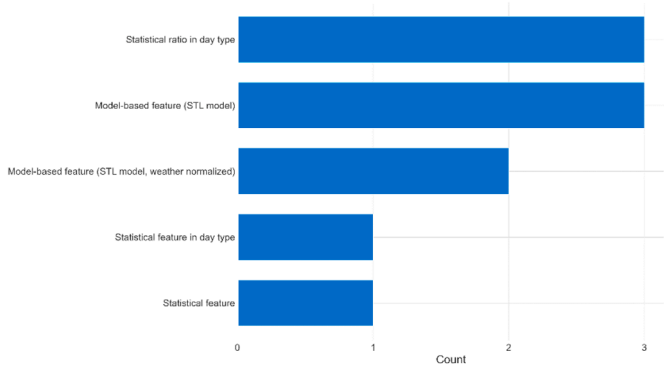
Classifier name	Feature type	Name of the cluster center	Features in the cluster	Descriptions
Education Vs Entertainment	Statistical feature in day type	Weekdays_50%	5	The energy consumption in weekdays
	Statistical feature in day type	Weekend_mean	12	The energy consumption in weekends
	Statistical feature	Top 10% energy consumption	2	The energy intensity
	Model-based feature (STL model)	Remainder_July_std	2	The energy variation from the usual situation in July
	Model-based feature (STL model)	Remainder_May_std	8	The energy variation from the usual situation in May
	Statistical ratio in day type	Weekend_innerratio_meduimvs95_mean	1	The energy consumption shape in weekends
	Model-based feature (multi-variable change-point model)	Coolingmax	3	The energy used for cooling
	Model-based feature (STL model, weather normalized)	Seasonal_weekly_std	2	The weekly seasonal energy consumption without the influence of weather
	Model-based feature (STL model)	Trend_Jun_std	2	The energy consumption trend in June
	Model-based feature (STL model)	Seasonal_weekly_Sat_mean	2	The weekly seasonal energy consumption on special day type
Education Vs Office	Statistical ratio in day type	Ratio_Fri.all_25%	1	The energy ratio between Friday and all days
	Model-based feature (STL model, weather normalized)	Trend_Feb_std	1	The energy consumption trend in February
	Statistical feature	mean	67	The energy intensity
	Statistical ratio in day type	Ratio_Fri.weekdays_25%	4	The energy ratio between Friday and weekdays
	Model-based feature (STL model)	Remainder_Apr_std	1	The energy variation from the usual situation in April
	Model-based feature (STL model, weather normalized)	Seasonal_weekly_Fri_mean	1	The weekly seasonal energy consumption without the influence of weather on Friday
	Statistical ratio in day type	Ratio_Thur.Wed_75%	1	The energy ratio between Thursday and Wednesday
	Statistical feature in day type	Weekdays_25%	5	The energy consumption in weekdays
	Model-based feature (STL model)	Remainder_Jun_mean	1	The energy variation from the usual situation in June
	Model-based feature (STL model)	Trend_May_mean	1	The energy consumption trend in May
Education Vs Public service	Statistical feature in day type	Weekend_mean	17	The energy consumption in weekends
	Statistical feature in day type	Mon_mean	1	The energy consumption in Monday
	Statistical ratio in day type	Ratio_Sat.Thur_mean	13	The energy ratio between Saturday and Thursday
	Statistical ratio in day type	Ratio_Sun.Sat_25%	1	The energy ratio between Sunday and Saturday
	Model-based feature (STL model, weather normalized)	Seasonal_weekly_Sat_mean	2	The weekly seasonal energy consumption without the influence of weather on special day type
	Statistical ratio in day type	Ratio_weekend.weekdays_75%	22	The energy ratio between weekend and weekdays
	Statistical ratio in day type	Ratio_weekend.all_50%	63	The energy ratio between weekend and all days
	Statistical ratio in day type	Ratio_Mon.all_mean	2	The energy ratio between Monday and all days
	Statistical feature in day type	Mon_50%	1	The energy consumption in Monday
	Statistical feature in day type	Fri_mean	2	The energy consumption in Friday
Entertainment Vs Office	Statistical feature	Max hour of day	2	Max energy consumption hour of day
	Model-based feature (STL model, weather normalized)	Seasonal_weekly_std	1	The weekly seasonal energy consumption without the influence of weather on special day type
	Model-based feature (DayFilter)	3_4h_Sun_freqover_0.2_groupscount	5	The daily pattern in Sunday
	Model-based feature (STL model)	Seasonal_weekly_std	1	The weekly seasonal energy consumption without the influence of weather
	Statistical feature in day type	Weekend_mean	18	The energy consumption in weekends
	Statistical feature	Top 10% energy consumption	2	The energy intensity
	Model-based feature (DayFilter)	7_8h_Sun_freqover_0.1_groupscount	2	The daily pattern in Sunday
	Statistical ratio in day type	Ratio_Mon.all_25%	1	The energy ratio between Monday and all days
	Statistical ratio in day type	Ratio_Fri.all_25%	1	The energy ratio between Friday and all days
	Model-based feature (DayFilter)	3_4h_Sat_freqover_0.4_groupscount	2	The daily pattern in Saturday
Entertainment Vs Public service	Model-based feature (multi-variable change-point model)	Total_CVRMSE	10	The energy used without the influence of weather
	Statistical feature	Max hour of day	2	Max energy consumption hour of day
	Model-based feature (STL model, weather normalized)	Trend_May_std	1	The energy consumption trend in May without the influence of weather
	Statistical ratio in day type	Weekend_meduimvsmax_std	16	The energy consumption shape in weekends
	Model-based feature (STL model)	Remainder_May_std	2	The energy variation from the usual situation in May
	Statistical ratio in day type	Ratio_Sat.weekend_std	5	The energy ratio between Saturday and weekend
	Statistical ratio in day type	Ratio_Fri.Wed_std	32	The energy ratio between Friday and Wednesday
	Model-based feature (STL model, weather normalized)	Trend_Aug_std	1	The energy consumption trend in August without the influence of weather
	Statistical ratio in day type	Ratio_Thur.weekend_std	10	The energy ratio between Thursday and weekend
	Model-based feature (STL model, weather normalized)	Remainder_Sep_std	2	The energy variation from the usual situation in September without the influence of weather
Office Vs Public service	Statistical ratio in day type	Ratio_Sat.Fri_mean	3	The energy ratio between Saturday and Friday
	Statistical feature in day type	Sat_mean	2	The energy consumption in Saturday
	Statistical ratio in day type	Ratio_Mon.weekend_25%	11	The energy ratio between Monday and weekend
	Statistical feature in day type	Fri_50%	2	The energy consumption in Friday
	Statistical feature in day type	Sun_mean	15	The energy consumption in Sunday
	Model-based feature (STL model)	Seasonal_weekly_std	1	The weekly seasonal energy consumption
	Statistical feature in day type	Mon_mean	1	The energy consumption in Monday
	Statistical ratio in day type	Ratio_Thur.Wed_25%	1	The energy ratio between Thursday and Wednesday
	Model-based feature (STL model, weather normalized)	Seasonal_weekly_std	1	The weekly seasonal energy consumption without the influence of weather
	Statistical feature in day type	Mon_50%	1	The energy consumption in Monday



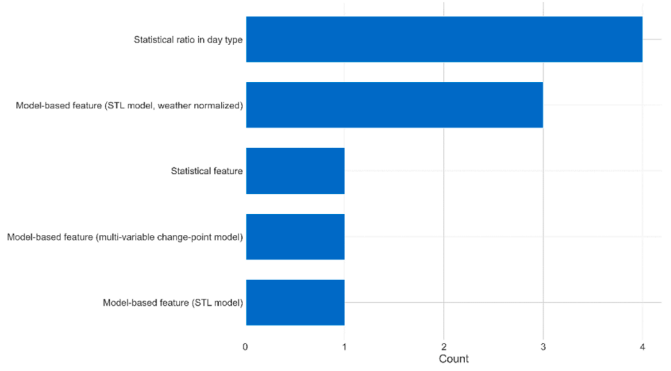
(a) Education Vs Entertainment



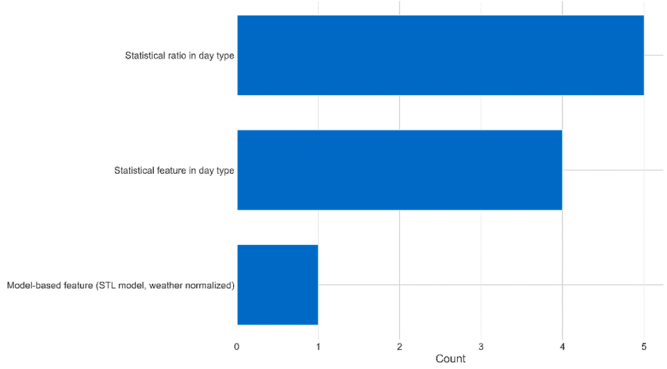
(d) Entertainment Vs Office



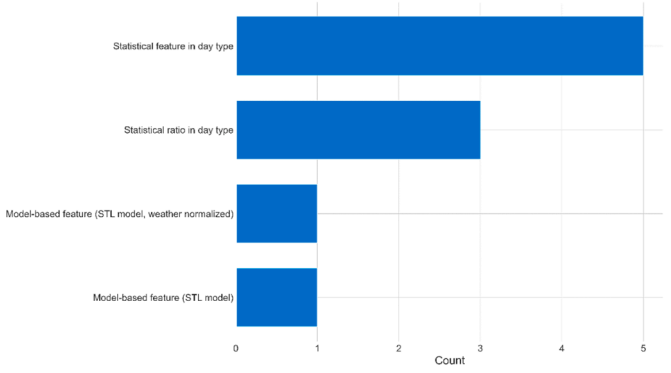
(b) Education Vs Office



(e) Entertainment Vs Public service



(c) Education Vs Public service



(f) Office Vs Public service

Fig. 11. . The top 10 cluster center features in binary classifiers.

features extracted with some data-driven methods or deep learning methods, such as primary component analysis (PCA) (Miller et al., 2018) (and autoencoder (Fan et al., 2019a). Thus, they are easier to accept by domain experts and have wider usage in engineering practice.

However, our method still has some shortcomings. Our limitation of this work is the further assessment of the extent to which the mislabel building is far from its PSU label. We can find some hints that the mislabel buildings identified by the ensemble classifier have different single-classifier vote results, as shown in Fig. 6. For example, building 165 and 174 in Fig. 6(b) are corrected as education buildings, whereas they have different classification results in the classifier that includes the PSU types education, entertainment and public service. The second limitation of our work is that our proposed model for classification is SVM because of its good accuracy in related work reported previously. More models are earning further work and attempt to improve the model identification accuracy. Third, more temporal features can be extracted from the model, such as the features based on the Fourier decomposition model. Although numerous features have been extracted from the time series and the method has good identification accuracy, we cannot prove that the features we extracted include all the temporal information needed for categorization. Finally, the method has not been applied to a larger dataset that includes more PSU types. Although we evaluated our model capability in some cases in this study, more cases are needed to analyze the performance of the method when facing wider usage.

7 Conclusion

In this study, we present a method for identify mislabeled

commercial buildings based on an analysis of the building energy time series collected by electric meters. Our focused label is a widely used categorization metric, primary space usage (PSU). The method enabled the comparison of buildings with different PSU labels by comparing the temporal features extracted from the energy time series. The main purpose of the method, which is the identification of the mislabeled buildings, is achieved by the prediction of the ensemble classifier. The predicted labels indicate the suggested labels of the buildings, and thus can help identify the mislabeling and correct the labels of the mislabel buildings. Our method was applied to a public dataset that comes from the Department of General Services from Washington, D.C. as a case study, and 1000 cases were designed to evaluate the identification accuracy and correction accuracy. The results of the case study showed that 22.4% of the buildings were mislabeled, and there were high mislabel rates in the office buildings. The results also indicated that the mix-use type is common in the real world and needs more attention. Meanwhile, the high transparency of the method assisted us in further studies on the temporal differences among buildings with different PSU labels. We learned that the current PSU labels are mostly suitable for many cases but required supplementation in the current state of mislabeling, especially the supplementation for the sub-label of the office buildings and mix-use buildings types. In conclusion, our proposed method can be used as an essential process before building energy benchmarking to help study the real state of energy efficiency and can assist us in further studies on building categorization in the real world.

Declaration of Competing Interest

None

Appendix

In Section 5.2, we analyzed the top 10 cluster center features for each binary classifier. The results are listed in Table 4.

References

Aggarwal, C.C. (.2015). Outlier Analysis, in: C. C. Aggarwal (Ed.), Data mining: The textbook. Springer International Publishing, Cham, pp. 237–263. 10.1007/978-3-319-14142-8_8.

Amber, K. P., Ahmad, R., Aslam, M. W., Kousar, A., Usman, M., & Khan, M. S. (2018). Intelligent techniques for forecasting electricity consumption of buildings. *Energy, 157*, 886–893. https://doi.org/10.1016/j.energy.2018.05.155

API Reference, tsod 0.1.2 documentation [WWW Document], n.d. URL https://dhi.github.io/tsod/api.html?highlight=constantvalue#tsod.ConstantValueDetector (accessed 6.1.21). (2021).

Benítez, I., Quijano, A., Díez, J.-. L., & Delgado, I. (2014). Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers. *International Journal of Electrical Power & Energy Systems, 55*, 437–448. https://doi.org/10.1016/j.ijepes.2013.09.022

Brodley, C. E., & Friedl, M. A. (1999). Identifying Mislabeled Training Data. *Journal of Artificial Intelligence Research, 11*, 131–167, 10.1613/jair.606.

Chen, Y., Xu, P., Chu, Y., Li, W., Wu, Y., Ni, L., & Wang, K. (2017). Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings. *Applied Energy, 195*, 659–670. https://doi.org/10.1016/j.apenergy.2017.03.034

CLEVELAND, R. B. (1990). STL : A seasonal-trend decomposition procedure based on loess. *Journal of Office Statistics, 6*, 3–73.

Commercial Buildings Energy Consumption Survey (CBECS), U.S. Energy Information Administration (EIA) [WWW Document], n.d. URL https://www.eia.gov/consumption/commercial/building-type-definitions.php (accessed 5.23.21). (2021).

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [cs, stat].

Energy Efficiency (2020), Analysis [WWW document], n.d.. IEA. URL https://www.iea.org/reports/energy-efficiency-2020 (accessed 5.23.21).

Fan, C., Sun, Y., Zhao, Y., Song, M., & Wang, J. (2019a). Deep learning-based feature engineering methods for improved building energy prediction. *Applied Energy, 240*, 35–45. https://doi.org/10.1016/j.apenergy.2019.02.052

Fan, C., Xiao, F., & Yan, C. (2015). A framework for knowledge discovery in massive building automation data and its application in building diagnostics. *Automation in Construction, 50*, 81–90. https://doi.org/10.1016/j.autcon.2014.12.006

Fan, C., Xiao, F., Yan, C., Liu, C., Li, Z., & Wang, J. (2019b). A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. *Applied Energy, 235*, 1551–1560. https://doi.org/10.1016/j.apenergy.2018.11.081

Fan, C., Yan, D., Xiao, F., Li, A., An, J., & Kang, X. (2021). Advanced data analytics for enhancing building performances: From data-driven to big data-driven approaches. *Building Simulation, 14*, 3–24. https://doi.org/10.1007/s12273-020-0723-1

Feng, W., Quan, Y., & Dauphin, G. (2020). Label noise cleaning with an adaptive ensemble method based on noise detection metric. *Sensors*20, 6718. 10.3390/s20236718.

Foucquier, A., Robert, S., Suard, F., Stephan, L., & Jay, A. (2013). State of the art in building modelling and energy performances prediction: A review. *Renewable & Sustainable Energy Reviews, 23*, 272–288. https://doi.org/10.1016/j.rser.2013.03.004

Fulcher, B. D., & Jones, N. S. (2017). hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell Systems, 5*, 527–531. https://doi.org/10.1016/j.cels.2017.10.001. e3.

Grolinger, K., L'Heureux, A., Capretz, M. A. M., & Seewald, L. (2016). Energy forecasting for event venues: Big data and prediction accuracy. *Energy and Buildings, 112*, 222–233. https://doi.org/10.1016/j.enbuild.2015.12.010

Karasu, S., Altan, A., Bekiros, S., & Ahmad, W. (2020). A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series. *Energy, 212*, Article 118750. https://doi.org/10.1016/j.energy.2020.118750

Kelly Kissock, J., & Eger, C. (2008). Measuring industrial energy savings. *Applied Energy, 85*, 347–361. https://doi.org/10.1016/j.apenergy.2007.06.020

Lavin, A., & Klabjan, D. (2015). Clustering time-series energy data from smart meters. *Energy Efficiency, 8*, 681–689. https://doi.org/10.1007/s12053-014-9316-0

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery, DMKD '0* (pp. 2–11). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/882082.882086.

Lipton, Z. (2016). The mythos of model interpretability. *Communications of the ACM, 61*. https://doi.org/10.1145/3233231

Luengo, J., Sánchez-Tarragó, D., Prati, R. C., & Herrera, F. (2021). Multiple instance classification: Bag noise filtering for negative instance noise cleaning. *Information Sciences, 579*, 388–400. https://doi.org/10.1016/j.ins.2021.07.076

Mathieu, J. L., Price, P. N., Kiliccote, S., & Piette, M. A. (2011). Quantifying changes in building electricity use, with application to demand response. *IEEE Transactions on Smart Grid, 2*, 507–518. https://doi.org/10.1109/TSG.2011.2145010

Miller, C. (2019). What's in the box?! Towards explainable machine learning applied to non-residential building smart meter classification. *Energy and Buildings, 199*, 523–536. https://doi.org/10.1016/j.enbuild.2019.07.019

Miller, C., & Meggers, F. (2017). Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings. *Energy and Buildings, 156*, 360–373. https://doi.org/10.1016/j.enbuild.2017.09.056

Miller, C., Nagy, Z., & Schlueter, A. (2015). Automated daily pattern filtering of measured building performance data. *Automation in Construction, 49*, 1–17. https://doi.org/10.1016/j.autcon.2014.09.004

Miller, C., Nagy, Z., & Schlueter, A. (2018). A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. *Renewable and Sustainable Energy Reviews, 81*, 1365–1377. https://doi.org/10.1016/j.rser.2017.05.124

Molnar, Christoph (2019). Interpretable machine learning. A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/.

Najafi, B., Depalo, M., Rinaldi, F., & Arghandeh, R. (2021). Building characterization through smart meter data analytics: Determination of the most influential temporal and importance-in-prediction based features. *Energy and Buildings, 234*, Article 110671. https://doi.org/10.1016/j.enbuild.2020.110671

New Energy Outlook (2020). | BloombergNEF [WWW Document], n.d. URL https://about.bnef.com/new-energy-outlook/(accessed 5.23.21).

Park, J. Y., Yang, X., Miller, C., Arjunan, P., & Nagy, Z. (2019). Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. *Applied Energy, 236*, 1280–1295. https://doi.org/10.1016/j.apenergy.2018.12.025

Pickering, E. M., Hossain, M. A., French, R. H., & Abramson, A. R. (2018). Building electricity consumption: Data analytics of building operations with classical time series decomposition and case based subsetting. *Energy and Buildings, 177*, 184–196. https://doi.org/10.1016/j.enbuild.2018.07.056

Pickering, E. M., Hossain, M. A., Mousseau, J. P., Swanson, R. A., French, R. H., & Abramson, A. R. (2017). A cross-sectional study of the temporal evolution of electricity consumption of six commercial buildings. *PloS one, 12*, Article e0187129. https://doi.org/10.1371/journal.pone.0187129

Powers, D.M.W. (2010). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv:2010.16061 [cs, stat].

Quintana, M., Arjunan, P., & Miller, C. (2021). Islands of misfit buildings: Detecting uncharacteristic electricity use behavior using load shape clustering. *Buildings Simulation, 14*, 119–130. https://doi.org/10.1007/s12273-020-0626-1

Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. arXiv:1606.05386 [cs, stat].

Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). "Why Should i trust you?": Explaining the predictions of any classifier. arXiv:1602.04938 [cs, stat].

Rokach, L. (2010). Ensemble-based classifiers. *Artificial intelligence review, 33*, 1–39. https://doi.org/10.1007/s10462-009-9124-7

Sezer, A., & Altan, A. (2021a). Detection of solder paste defects with an optimization-based deep learning model using image processing techniques. *Soldering & Surface Mount Technology, 33*, 291–298. https://doi.org/10.1108/SSMT-04-2021-0013

Sezer, A., & Altan, A. (2021b). Optimization of deep learning model parameters in classification of solder paste defects. In *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). Presented at the 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1–6). https://doi.org/10.1109/HORA52670.2021.9461342

Sun, Y., Haghighat, F., & Fung, B. C. M. (2020). A review of the -state-of-the-art in data -driven approaches for building energy prediction. *Energy Buildings, 221*, Article 110022. https://doi.org/10.1016/j.enbuild.2020.110022

Vallis, O., Hochenbaum, J., & Kejariwal, A. (2014). A novel technique for long-term anomaly detection in the cloud. In *Presented at the 6th {USENIX} workshop on hot topics in cloud computing (HotCloud 14)*.

Wang, Y., Chen, Q., Hong, T., & Kang, C. (2019). Review of smart meter data analytics: applications, methodologies, and challenges. *IEEE Transactions on Smart Grid, 10*, 3125–3148. https://doi.org/10.1109/TSG.2018.2818167

World Energy Investment (2020). Analysis [WWW Document], n.d.. IEA. URL https://www.iea.org/reports/world-energy-investment-2020 (accessed 5.23.21).

Yu, Z., Fung, B. C. M., & Haghighat, F. (2013). Extracting knowledge from building-related data - A data mining framework. *Buildings Simulations, 6*, 207–222. https://doi.org/10.1007/s12273-013-0117-8

Zhan, S., Liu, Z., Chong, A., & Yan, D. (2020). Building categorization revisited: A clustering-based approach to using smart meter data for building energy benchmarking. *Applied Energy, 269*, Article 114920. https://doi.org/10.1016/j.apenergy.2020.114920