IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 12, 2022, accepted March 10, 2022, date of publication March 16, 2022, date of current version March 25, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3160170

Toward Delicate Anomaly Detection of Energy Consumption for Buildings: Enhance the Performance From Two Levels

DONG WANG[1], THERESE ENLUND[2], JOHAN TRYGG[1], MATS TYSKLIND[1], AND LILI JIANG[3]

[1]Department of Chemistry, Umeå University, 901 87 Umeå, Sweden
[2]Mestro AB, 111 43 Stockholm, Sweden
[3]Department of Computing Science, Umeå University, 901 87 Umeå, Sweden

Corresponding author: Lili Jiang (lili.jiang@cs.umu.se)

This work was supported by the Green Technology and Environmental Economics (GreenTEE) Platform at Umeå University.

ABSTRACT Buildings are highly energy-consuming and therefore are largely accountable for environmental degradation. Detecting anomalous energy consumption is one of the effective ways to reduce energy consumption. Besides, it can contribute to the safety and robustness of building systems since anomalies in the energy data are usually the reflection of malfunctions in building systems. As the most flexible and applicable type of anomaly detection approach, unsupervised anomaly detection has been implemented in several studies for building energy data. However, no studies have investigated the joint influence of data structures and algorithms' mechanisms on the performance of unsupervised anomaly detection for building energy data. Thus, we put forward a novel workflow based on two levels, data structure level and algorithm mechanism level, to effectively detect the imperceptible anomalies in the energy consumption profiles of buildings. The proposed workflow was implemented in a case study for identifying the anomalies in three real-world energy consumption datasets from two types of commercial buildings. Two aims were achieved through the case study. First, it precisely detected the contextual anomalies concealed beneath the time variation of the energy consumption profiles of the three buildings. The performance in terms of areas under the precision-recall curves (AUC_PR) for the three given datasets were 0.989, 0.941, and 0.957, respectively. Second, more broadly, the joint effect of the two levels was examined. On the data level, all four detectors on the contextualized data were superior to their counterparts on the original data. On the algorithm level, there was a consistent ranking of detectors regarding their detecting performances on the contextualized data. The consistent ranking suggests that local approaches outperform global approaches in the scenarios where the goal is to detect the instances deviating from their contextual neighbors rather than the rest of the entire data.

INDEX TERMS Buildings, energy consumption, anomaly detection, contextualization, unsupervised learning.

I. INTRODUCTION

Energy consumed in buildings accounts for one-third of the final energy and half of the electricity in the global economy, making buildings the most energy-consuming segment. The vast energy consumption also results in enormous environmental impact – depletion of non-renewable fossil fuels and release of CO_2 and other pollutants. It has been reported that more than one-third of the global carbon emissions are from buildings [1]–[3]. Encouragingly, buildings are assessed to possess a highly untapped potential for energy efficiency [4]. The optimization of buildings' energy consumption should be focused on the daily operational energy since operational energy usually makes up the vast majority of the building's life cycle energy consumption [5]. Typically, optimizations are based on analyzing the patterns in energy consumption profiles, for example, benchmark identification, peak load forecast, and anomaly detection [6]–[13]. Anomaly detection of energy load profile not only benefits energy and operational cost saving but also can contribute to the safety and robustness of building systems. The reason is that anomalies in the energy data usually reflect faults of building systems,

The associate editor coordinating the review of this manuscript and approving it for publication was Khursheed Aurangzeb.

for example, poor maintenance, negligent operation, errors in sensing and transmission system, abrupt malfunction of equipment, or operational strategies with minimal consideration of energy efficiency [9], [14], [15]. Therefore, there is a great significance for conducting anomaly detection of energy load for buildings to save operational cost, reduce carbon emissions, and keep building systems robust and safe.

Anomalies can be generally categorized into three types based on their nature – point anomalies, contextual anomalies, and collective anomalies [16]–[18]. If an instance in the data deviates from the rest, it is regarded as a point anomaly (global sense). If an instance is anomalous in a particular context with respect to the rest instances in the same context, it is called a contextual anomaly. This instance might be normal in another context or in the global sense. In the sense of contextual anomaly detection, there are two types of attributes for an instance – contextual attribute and behavioral attribute. A behavioral attribute is the record of the object's behaviors, such as energy or water consumption. A contextual attribute is (part of) the environment where the behavior happens, such as temporal or spatial information. Collective anomalies are the subset of data as a collection deviates from the rest, but the individual instances in the subset are not anomalous either in a global or contextual sense. Of these three types of anomalies, contextual anomalies are of the most interest to this study for the following two reasons. First, temporal contexts significantly influence energy consumption in buildings, especially for the objects of this study – commercial buildings. The demand for heating or cooling changes as the season changes, and the intensity of user activities varies from day to night and from weekdays to weekends. Second, compared to the other two types, contextual anomalies are usually much harder to be intuitively identified. They are typically hidden among many other instances that have similar behavioral attribute values but are normal in their own contexts.

There are generally three different types of approaches for detecting anomalies. The first one is supervised anomaly detection [19], [20], which requires the training and test data to be fully labeled. However, in practice, it is usually very expensive and demanding to correctly label an adequate amount of instances as anomalies. Also, supervised learning on extremely imbalanced data has been proved to be very difficult and tends to yield highly biased results [21], [22]. The second type is semi-supervised anomaly detection, also known as one-class classification [23], [24]. Semi-supervised anomaly detection also needs training and test sets, but the training set only comprises normal instances. The normal pattern is learned by the model in the training stage, and the anomalies are identified if they deviate from the normal pattern. For semi-supervised anomaly detection, the labeling of anomalies is not required for training data, but the prior knowledge of identifying normal instances is crucial. The third one is unsupervised anomaly detection [25]–[31]. It is the most flexible and applicable approach among the three since it does not require the training of models, meaning the labeling of normal instances and anomalies is not required. Besides, unsupervised anomaly detection makes it easy to keep the detection dynamic and updated without the restriction from the old training data. Unsupervised anomaly detection algorithms estimate instances' anomalousness according to the intrinsic properties of the data, such as the structure of data and the relationships between instances. The more distinct an instance is from others, the higher score it will be assigned. For the tasks of unsupervised anomaly detection, there are numerous options in terms of algorithms based on various mechanisms.

In spite of the various practice of unsupervised anomaly detection for building energy data (see Section II), to the best of the authors' knowledge, there have been no studies investigating the combined influence of data structures and algorithms' mechanisms on the detection performance. Data and algorithms are the two key factors, so their individual and joint influences on the detection performance should be researched to better understand what data form should be prepared and what algorithm should be adopted for a more accurate and robust anomaly detection result in practice.

In this paper, we bridge this gap by proposing a workflow to evaluate the difference in effect between the original data (with only the behavioral attribute) and contextualized data (with both behavioral and contextual attributes), and between the unsupervised algorithms with global, local, and global-local-hybrid perspectives. The details and applicability of the workflow are demonstrated through a case study on the three commercial buildings' data provided by an energy management company, Mestro AB, in Sweden. There are two significant contributions of this paper: 1) Locally, for the case study subjects, it aims for precise identification of the contextual anomalies concealed beneath the time variation of the energy consumption profiles; 2) More broadly, it investigates the joint influence pattern of data structures and algorithm mechanisms on the performance of unsupervised anomaly detection for buildings' energy data.

The paper outline is as follows: Section II presents the related work and the gap in the literature; Section III describes the source and properties of the data; Section IV demonstrates the details of the fundamental methods; Section V illustrates the workflow of this paper; The results are shown and discussed in Section VI; In Section VII, the conclusions are drawn, and the future perspectives are discussed.

II. RELATED WORK

There have been various studies featuring different types of techniques on anomaly detection for energy consumption or energy-consumption-related issues in buildings.

Many anomaly detection studies adopt a two-stage framework. First, use the model trained on the historical energy consumption data to predict the current energy consumption. Second, compare the observed energy consumption with the predicted one, and a significant difference indicates that the observed one is anomalous [32]–[37]. The prediction can be based on Autoregressive Integrated Moving

Average (ARIMA), Periodic Auto-regression with Exogenous Variables (PARX), Artificial Neural Networks (ANN), or Long Short-Term Memory (LSTM). And the anomaly determination approaches can vary from a simple two-sigma rule to an active adaptive threshold to some complex identification systems, for example, Negative Selection algorithm, and an independent module incorporating Support Vector Machine (SVM), k-Nearest Neighbors (kNN), and cross-entropy. All of the studies claimed good detecting performances on the test data. However, this type of anomaly detection depends heavily on the data used to train the prediction models. If the training data contain many imprecise energy consumption values or even anomalies, the accuracy of the prediction models will be low. This will significantly hinder the afterward identification of anomalies from being effective and accurate.

In order to avoid the critical drawback associated with the quality of training data, many unsupervised approaches have been explored. Weng *et al.* proposed an unsupervised method based on LSTM and autoencoder to detect anomalous energy consumption at a campus [38]. LSTM structure was leveraged to incorporate the contextual information in the modeling. However, the contextual information in their paper only referred to the intrinsic sequential connections between instances due to time variation. There was no consideration of explicit contextual attributes. Besides, there was no comparison between the scenarios with and without the sequential information considered. Additionally, even though the performance of their method was compared to other common methods on the well-known anomaly detection datasets, the reason behind the performance difference was not investigated. Yeckle *et al.* explored seven unsupervised anomaly detection algorithms to detect electricity theft for improving the security of the Advanced Metering Infrastructure, and most of them showed good effectiveness [39]. However, there was no discussion on whether/how algorithm mechanisms caused the difference in the results. The algorithms were applied to datasets with different dimensionalities, but there was no information about the attributes' meanings and no appropriate discussion on the possible reasons associated with dimensionality. Liu *et al.* used Density-based Spatial Clustering Application with Noise (DBSCAN) to identify anomalies in the process of extracting typical electricity load patterns (TELPs) [40]. Multiple contextual attributes were included in the data, and their association with TELPs was studied. However, their influence on anomaly detection was neglected. Furthermore, there was no comparison with other anomaly detection techniques. Fan *et al.* applied an ensemble of autoencoders with various architectures and training schemes to identify building energy data anomalies [41]. There were multiple contextual attributes in the data, but only one contextual attribute was considered when comparing the detection performances between the original and contextualized data. Nevertheless, different levels of masking noise in the data were examined to reveal the influence of noise. In terms of algorithms, only autoencoder was involved. Pereira *et al.* proposed an approach incorporating variational recurrent autoencoder with self-attention and probabilistic reconstruction scoring for anomaly detection of time series energy consumption data [42]. The bidirectional LSTM module with a self-attention mechanism was employed to capture the temporal context around every instance, but other contextual information was not examined. Also, only one detecting approach was studied. Wang *et al.* applied and compared four algorithms, Deep Neural Network Regression (DNNR), Autoencoder with reconstruction (AER), encoder of the Autoencoder (EAE), and Support Vector Regression (SVR), in a task of detecting electricity meter failure (point anomalies) and unusual electricity consumption (contextual anomalies) [43]. The unsupervised AER was the optimal model for detecting point anomalies, and the unsupervised EAE performed almost equally well as the optimal model DNNR in detecting contextual anomalies. Nevertheless, the correlation between the mechanisms of the algorithms and the results was not examined. Also, the effects of the multiple data attributes on the detection performances were not investigated.

III. DATA SOURCE AND CHARACTERISTICS

The three datasets used in the case study are the electricity consumption profiles of Mestro AB's clients monitored by electricity meters, with the monitoring frequency of one record per hour. They are from two typical types of commercial buildings in three different Swedish cities – Karlskoga, Göteborg, and Jönköping. All the three datasets are the records for the whole year of 2018, so all of them possess 8760 instances. The original datasets only contain one attribute, which is the behavioral attribute 'energy consumption.' Dataset *A* was retrieved from the main electricity meter installed in a property in Karlskoga of 1622 m^2 heated area. The property was used for retail business. Dataset *B* was fetched from the main electricity meter serving a property in Göteborg of 47,166 m^2 heated area. The property was a university building, in which most of the rooms were offices. Dataset *C* was retrieved from the main electricity meter serving a property in Jönköping of 28,046 m^2 heated area. The function of this property was retail. General seasonality can be observed in all the three datasets – colder months correspond to higher energy consumption, and warmer months correspond to lower energy consumption.

Initially, the instances in the raw datasets were without labels of being normal or abnormal. However, they were labeled for evaluating anomaly detection performances. The labeling was carried out by employing both empirical domain knowledge and statistical methods as follows. First, the raw datasets were divided into various temporal groups at multiple levels. The instances in the same group are supposed to possess similar energy consumption values because, theoretically, the energy consumption activities shall be relatively consistent within each temporal range. Subsequently, the 3-sigma criterion was used to identify the outliers in each temporal group. The 3-sigma rule was adopted because it

had been proved to be effective in the literature of a similar study [44]. The collection of all the outliers from these groups is the set of labeled anomalies used in this study for detection performance evaluation. Through this approach, 30 instances were labeled as anomalies in dataset *A*, 35 in dataset *B*, and 25 in dataset *C*. The corresponding anomaly rates are 0.34%, 0.40%, and 0.29% for datasets *A, B, and C*, respectively.

IV. METHODS

A. LOCAL OUTLIER FACTOR (LOF)

Local Outlier Factor (LOF) [45] is a density-based outlier/anomaly detection algorithm identifying anomalies by estimating the deviation of the instance compared to its neighbors within a local range. The deviation is quantified by LOF values associated with density differences between the neighborhoods of the instance and its neighbors. LOF values reflect the instances' degree of anomalousness – a larger value means a more anomalous instance.

Given a dataset D and an instance p in D, the major steps of calculating the LOF value of p are as follows:

1. Calculate the *k_distance* of p (*k_distance(p)*), which is the distance between p and the kth nearest neighbor of p.
2. Identify the *k_distance* neighborhood (k nearest neighbors) of instance p:
   k_distance neighborhood of p is composed of every instance (denoted by q) whose distance from p is not larger than the *k_distance(p)*:

$$N_{k_{distance(p)}}(p) = \left\{q \in D \backslash \{p\} \,|d(p,q) \leq k_{distance(p)}\right\} \tag{1}$$

3. Calculate the reachability distance of p from another instance o (denoted by $reach\_dist_k(p,o)$). $reach\_dist_k(p,o)$ is the true distance between p and o, but at least the k_distance of o, which can be expressed by the following equation:

$$reacch_{dist\,k}(p,o) = max\left\{k_{distance(o)}, d(p,o)\right\} \tag{2}$$

4. Calculate the *local reachability density* of p (denoted by $lrd_k(p)$). $lrd_k(p)$ is the inverse of the average reachability distance of p from its k nearest neighbors:

$$lrd_k(p) = \frac{1}{\frac{\sum_{o \in N_k(p)} reach_{dist\,k}(p,o)}{|N_k(p)|}} \tag{3}$$

   where $N_k(p)$ is the set of the k nearest neighbors of p.
5. Calculate the *local outlier factor* of p (denoted by $LOF_k(p)$). $LOF_k(p)$ is the average *local reachability density* of p's k nearest neighbors divided by p's own *local reachability density*:

$$LOF_k(p) = \left(\frac{\sum_{o \in N_k(p)} lrd_k(o)}{|N_k(p)|}\right) \Big/ lrd_k(p) \tag{4}$$

B. CONNECTIVITY-BASED OUTLIER FACTOR (COF)

Connectivity-based outlier factor (COF) [46] can be regarded as a variant of LOF, which also estimates the instance's anomalousness by comparing it with its k nearest neighbors. However, it is superior to LOF in identifying the anomalies deviating from a low-density neighborhood. The motivation for developing COF is that an anomaly is not always in a lower-density neighborhood – it can be isolated from a pattern where the instances are well connected. The goal of COF is to estimate a COF value reflecting the instance's degree of being isolated. A larger COF value of an instance indicates that the instance is more anomalous.

The major steps of calculating the COF value of an instance p are as follows:

1. Find the k nearest neighborhood for p (denoted by $N_k(p)$).
2. Find *set based nearest path* (*SBN_path*) and the corresponding *set based nearest trail* (*SBN_trail*) for p. The *SBN_path* from instance p_1 to $N_k(p_1)$ is a sequence of instances $s = \{p_1, p_2, \ldots, p_{k+1}\}$ such that for all $1 \leq i \leq k$, p_{i+1} is the nearest neighbor of set $\{p_1, \ldots, p_i\}$ in $\{p_{i+1}, \ldots, p_{k+1}\}$. *SBN_trail* is a sequence of edges $e = \{e_1, \ldots e_k\}$, and each edge is a pair of two consecutive neighbors from *SBN_path.* The distance between the two neighbors in one edge is denoted by $dist(e_i)$.

1) Calculate the average chaining distance from p to its k nearest neighbors $N_k(p)$, denoted by $ac\_dist_{N_k(p)}(p)$ and defined as:

$$ac_{dist\,N_k(p)}(p) = \sum_{i=1}^{k} \frac{2(k+1-i)}{k(k+1)} dist(e_i) \tag{5}$$

   Given $o \in N_k(p)$ as one of the k nearest neighbors of p, calculate $ac\_dist_{N_k(o)}(o)$ for every o.

2) Calculate COF value of p with respect to its k nearest neighbors $N_k(p)$, which is defined as:

$$COF_k(p) = \frac{ac_{dist\,N_k(p)}(p)}{\frac{1}{|N_k(p)|}\sum_{o \in N_k(p)} ac_{dist\,N_k(o)}(o)} \tag{6}$$

C. CLUSTER-BASED LOCAL OUTLIER FACTOR (CBLOF)

Cluster-Based Local Outlier Factor (CBLOF) [47] is developed based on the concept that the instances not lying in the large clusters should be regarded as outliers. After clustering the instances and defining large and small clusters, CBLOF calculates a final score (CBLOF value) for each instance to indicate how much this instance deviates from its 'local' large cluster, i.e., the anomalousness of it. The clustering algorithm used for partitioning the dataset into multiple clusters is not restricted. However, the critical issue is to define which clusters are large and which ones are small. Suppose $C = \{C_1, C_2, \ldots C_k\}$ is the set of clusters after the partitioning of dataset D, and the sizes of the clusters are in the order $|C_1| \geq |C_2| \geq \ldots \geq |C_k|$. With two numeric parameters α and β, C_b is regarded as the boundary of large clusters if

one of the following formulas holds.

$$|C_1| + |C_2| + \ldots + |C_b| \geq |D| \cdot \alpha \tag{7}$$

$$\frac{|C_b|}{|C_{b+1}|} \geq \beta \tag{8}$$

Thus, the set of large clusters is $LC = \{C_i | i \leq b\}$, and the set of small clusters is $SC = \{C_j | j > b\}$.

With the large clusters and small clusters defined, the CBLOF value of instance p in dataset D can be defined as follows:

$$CBLOF\ (p) = \begin{cases} |C_i| \cdot \left(dist\left(p, C_j\right)\right), & where\ p \in C_i, \\ & C_i \in SC, C_j \in LC \\ |C_i| \cdot dist\ (p, C_i), & where\ p \in C_i, \\ & C_i \in LC \end{cases} \tag{9}$$

This means CBLOF value of an instance is subject to the size of its cluster, and the distance between the instance and its closest large cluster (if this instance is in a small cluster), or the distance between the instance and its cluster (if this instance belongs to a large cluster).

D. ISOLATION FOREST (IF)

Isolation forest (IF) [48] is established based on the idea that anomalous instances tend to get isolated more easily under random partitioning than the normal ones in the dataset. IF is an ensemble of multiple isolation trees (iTrees). In each iTree, the dataset $D = \{p_1, p_2, \ldots p_n\}$ is recursively divided by randomly choosing an attribute and a value between the attribute's minimum and maximum values for the split. An iTree stops growing until: (1) the tree reaches a depth limit, (2) $|D| = 1$, or (3) all instances in D have the same values. After building the iTrees, to help quantify the degree of anomalousness, the path length $h(p)$ is calculated for each instance p in every iTree. $h(p)$ is measured as the number of partitions needed to isolate the instance p, from the root to the terminating node of the iTree. Ultimately, the anomaly score needs to be estimated for each instance after obtaining its path length. The average path length of unsuccessful search in Binary Search Tree (BST) is borrowed to normalize $h(p)$ because of the equivalence between the iTree and BST structures: a termination to an external node of the iTree corresponds to an unsuccessful search in the BST. For D with n instances, the average path length of unsuccessful search in BST is:

$$C\ (n) = 2H\ (n-1) - \left(\frac{2\ (n-1)}{n}\right) \tag{10}$$

where $H(i)$ is the harmonic number, and it can be estimated by $ln(i) + 0.5772156649$ (Euler's constant).

Thus, the anomaly score s of an instance p in D is defined as:

$$s\ (p, n) = 2^{-\frac{E(h(p))}{C(n)}} \tag{11}$$

where $E(h(p))$ is the average of $h(p)$ from the ensemble of iTrees in IF.

E. STACKED AUTOENCODER (SAE)

An autoencoder is an unsupervised method leveraging an artificial neural network to learn efficient data representation in a latent space. The representation is validated and refined by iteratively reconstructing the original input from the representation and increasing the similarity between the reconstruction and the original input. The module mapping original data to the latent space in an autoencoder structure is named encoder. In contrast, the one of reconstructing original data from the latent space is named decoder. For encoder and decoder, they are always symmetric to one another. Autoencoders possessing multiple hidden layers (the layers between the input and representation or the ones between representation and reconstruction) are called stacked autoencoders (SAE). Like a typical multilayer perceptron, in SAE, data are fed forward from input to reconstruction layer, but training is performed using the backpropagation method [49]. Based on the gradient descent of the loss function, the neurons' weights and biases are updated backward from the reconstruction layer. The activation function applied for the SAE in this study is ReLU [50]. The training is performed by minimizing the least-square loss between the input layer and the reconstruction layer. The low-dimensional representation layer with significant amounts of information retained is the desired result in this study.

F. PRECISION-RECALL (PR) CURVE AND ITS AREA UNDER CURVE (AUC_PR)

Precision-recall (PR) curve is a graph to evaluate classification models' performance at various classification thresholds. It is a variant of the well-known receiver operating characteristic (ROC) curve. However, it is more accurate than ROC curve with imbalanced data, where ROC curve tends to yield an overly optimistic result [51].

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

In a binary classification problem (for example, anomaly detection), a model classifies instances into either positive or negative. Thus, as the confusion matrix in Table 1 shows, there are four categories regarding the classification result: true positives (TP), false positives (FP), true negatives (TN), false negatives (FN). Based on the confusion matrix, *recall* is defined in Equation (12), and *precision* is defined in Equation (13). In a PR curve plot, the values on the x-axis are *recall* ranging between 0 and 1, and the ones on the y-axis are *precision* ranging between 0 and 1. PR curve is the result of connecting all the (*recall, precision*) scatters

TABLE 1. Confusion matrix for binary classification.

	Actual positive	Actual negative
Predicted positive	TP	FP
Predicted negative	FN	TN

corresponding to the set of selected thresholds, which shows the tradeoff between precision and recall for different classification thresholds. The area under the PR curve (AUC_PR) is calculated to represent the model's prediction performance. AUC_PR is an overall indicator of the models' performance across all possible classification thresholds. AUC_PR can be understood as the probability that the model ranks a random positive instance above a random negative instance. The AUC_PR value is always bounded between 0 and 1, and a larger AUC_PR indicates a better performance. AUC_PR is adopted for performance evaluation and comparison for its two advantages: 1) Scale-invariant. It evaluates the ranking of predictions rather than the absolute predicted values. 2) Classification-threshold-invariant. It assesses the model's prediction performance irrespective of what classification threshold is applied.

V. WORKFLOW

As stated in Introduction, the ultimate goal of this study is to investigate how different data structures and algorithms with different perspectives influence anomaly detection performance. Therefore, the workflow of this study revolves around the employment and comparison of different data structure schemes and different perspectives of algorithm mechanisms.

A. CONTEXTUALIZATION

As shown in Figure 1, the whole workflow starts from the original data and goes through two parallel paths. The green arrow shows the first path, while the red arrows show the second path. The black arrows represent the flow shared by both paths. The first path is straightforward, directly applying the four algorithms to the original data to acquire the anomaly detection results, followed by the result ensemble (details presented below in this section) and comparison. The second path is distinguished from the first by its contextualization followed by the dimension reduction. Contextualization refers to adding temporal attributes to the original data's behavioral attribute 'energy consumption.' The temporal information is extracted from the time series index of the original data. The motivation of contextualization is that commercial buildings' activities are highly subject to time variation. Thus, the temporal attributes *hour*, *day class*, and *month* can define the environment the energy consumption yielded from, providing more information for estimating correlations between the instances. For the three categorical contextual attributes, one-hot encoding [52] is used to convert them to binary attributes. The number of binary attributes varies from dataset to dataset, depending on how many day classes for this dataset. Weekdays are always within one class; Saturdays and Sundays can belong to the same class or two different classes. Public holidays are always within the class of Sundays. Thus, there are thirty-nine or forty binary attributes. Those attributes are sparse attributes with scattered information, and the high dimensionality hinders the efficient and effective application of anomaly detection algorithms. Thus, compression of the binary attributes is required. SAE is used in this study to compress the data because it helps retain much more complex information compared to the traditional linear reduction method, such as Principal Component Analysis (PCA). Based on the test, 4 turns out to be the optimal number of dimensions for the latent space in SAE. Thus, the final dimensionality of the data after contextualization and compression is 5. Subsequently, the final data are fed to the four algorithms – LOF, COF, CBLOF, and IF.

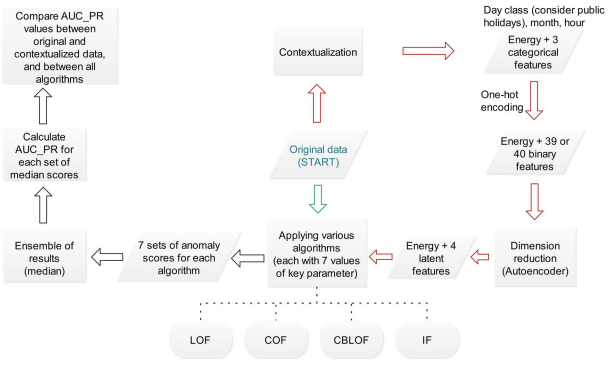

FIGURE 1. Workflow of study.

B. VALUE SETS OF KEY PARAMETERS

From the procedure where the four algorithms are applied to the final data, the two paths start to share the same procedures until the end of the workflow. For both original and contextualized data, seven distinct values of the key parameter (LOF: *number of neighbors*, COF: *number of neighbors*, CBLOF: *number of clusters*, IF: *random state*) are passed to each algorithm. LOF and COF share the same set of *number of neighbors*: {8, 16, 24, 32, 48, 64, 80}. The set of *number of clusters* for CBLOF is {160, 200, 240, 280, 320, 360, 400}. The set of *random state* for IF is {0, 2, 4, 5, 7, 9, 11}. The values of the parameters are chosen to cover a wide range of scenarios in consideration of the characteristics of the algorithms and datasets. For example, when *number of clusters* is 8, LOF and COF will assign a high anomaly score to the instance deviating from its eight nearest neighbors. This is the scenario of tight context. When *number of neighbors* is 80, the instance of interest will be compared with eighty nearest neighbors to estimate its anomalousness. This is the scenario of loose context. Similarly, *number of clusters* $= 400$ will set off a sensitive detecting strategy for CBLOF since more small clusters will get separated from others. In contrast, the CBLOF detector with *number of clusters* $= 160$ will have a much higher tolerance for grouping instances far from each other in the same cluster. This leads to the situation where many slightly deviating instances are not assigned high scores because they are not in small clusters. The seven distinct values of *random state* do not possess any numerical meaning. Instead, they can be regarded as seven different signs denoting seven distinct random seeds, corresponding to seven different random partition schemes.

C. MEDIAN ENSEMBLE

There are seven detectors for each algorithm corresponding to seven values of each key parameter. Thus, there are seven sets of anomaly scores for each algorithm for both the original and contextualized data. In this study, the median value of the seven anomaly scores for each instance is chosen to form a new set of anomaly scores. One advantage of doing so is that instances' degree of being anomalous will be much less subject to the specific values assigned to the key parameters. This benefit is of great significance in real-world application scenarios. In real-world applications, without feedback (labels), it is almost impossible to know the optimal value of a key parameter. Thus, trying with multiple values and taking the median score can be a very effective way to obtain robust results. Additionally, the median anomaly scores are a much more unbiased reflection of the algorithms' performances. This is crucial for this study since the goal is to examine the independent and joint effects of algorithms and data on anomaly detecting performance.

The new set of anomaly scores (median) is regarded as the result of the virtual ensemble detector. The PR curves of each ensemble detector are plotted based on their anomaly scores. With the PR curves and the AUC_PR values, the results of the ensemble detectors are evaluated and compared. Within the same dataset, the comparison is between the original and the contextualized data, and also between four algorithms. The optimal detector is selected based on the comparison, and the patterns presented in the results are discussed.

VI. RESULTS AND DISCUSSION

The PR curves of each ensemble detector on both original and contextualized data were plotted based on their anomaly scores. The corresponding AUC_PR values of these PR curves were calculated. For each dataset, the comparison was carried out between original and contextualized data and also between the ensemble detectors.

As shown in Table S1 to Table S12 in Supplementary Material, the virtual ensemble detectors' performances successfully represent the algorithms' overall performances on the data without being biased by the extreme anomaly scores. This property is advantageous in the typical real-world application scenarios, where we cannot know what value of the parameter contributes to a better detector since it is impossible to assess the detectors' performances without labels. However, this study clearly shows that taking the median of the anomaly scores under a set of parameter values can reflect the overall performance unbiasedly, regardless of specific parameter values adopted.

A. PERFORMANCES OF ANOMALY DETECTORS

1) RESULTS FOR DATASET *A*

As shown in Figure 2, it is evident that most of the detectors on the contextualized data cover more area (higher AUC_PR value) than the ones on original data do. It is not clear for CBLOF from the plots, but from Table 2, a slight improvement on contextualized data (0.784 *vs.* 0.764) can be observed. The difference shown by all detectors indicates that contextualization enhances the anomaly detection for dataset *A*, regardless of the specific algorithms. One thing worth noting in Figure 2 is that the curve trend of OG_CBLOF is opposite to the typical PR curve trend shown by other curves. This also happens with OG_COF, OG_IF, and CT_CBLOF in certain ranges. Since this unusual pattern also appears in Figure 3 and Figure 4, the likely reason behind it will be discussed in the summary of the three datasets' results in Section VI-B. Besides, in terms of performance, all the ensemble detectors are in the same sequence for both original data and contextualized data, which is LOF > COF > CBLOF > IF. Additionally, LOF, COF, and CBLOF present decent performances with the respective AUC_PR values of 0.989, 0.876, and 0.784 for the contextualized data, and 0.838, 0.817, and 0.764 for the original data. The overview of the results on dataset *A* is as follows: 1) All the detectors on the contextualized data are slightly superior to their counterparts on the original data; 2) LOF > COF > CBLOF > IF in terms of performance, and the AUC_PR value of LOF is 0.989 on the contextualized data.

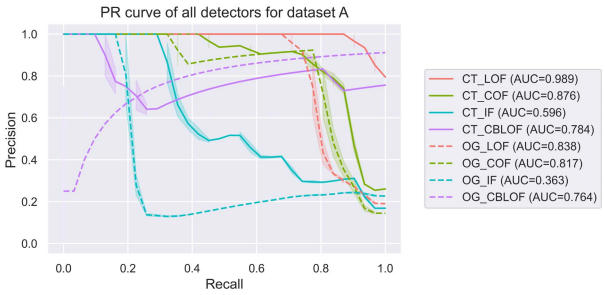

FIGURE 2. PR curves of detectors' median anomaly scores for dataset *A* (CT stands for the contextualized data, OG stands for the original data).

TABLE 2. AUC_PR values of ensemble detectors for dataset *A*.

	LOF	COF	IF	CBLOF
Original data	0.838	0.817	0.363	0.764
Contextualized data	0.989	0.876	0.596	0.784

2) RESULTS FOR DATASET *B*

The unique and most obvious feature of Figure 3 is that the curve OG_COF is absent because there was a numerical error while calculating the COF values on the original data. The possible reason is as follows. When a neighborhood is composed of instances with the same or very similar energy consumption values, the average chaining distances of those instances could be zero or very small. This means the denominator in Equation (6) will be zero or close to zero, which leads the final COF value to be null or extremely large (meaningless). This problem is likely to happen when the following two conditions are met: 1) There are many instances with the

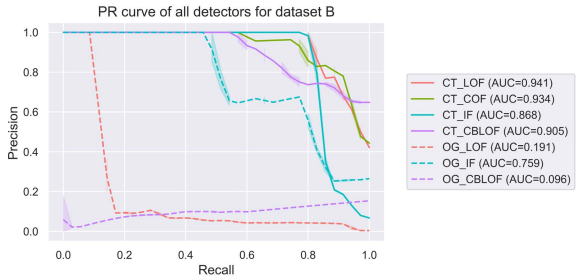

FIGURE 3. PR curves of detectors' median anomaly scores for dataset *B* (CT stands for the contextualized data, OG stands for the original data).

TABLE 3. AUC_PR values of ensemble detectors for dataset *B*.

	LOF	COF	IF	CBLOF
Original data	0.191	N/A	0.759	0.096
Contextualized data	0.941	0.934	0.868	0.905

same or very similar values in the original data; 2) The COF algorithm parameter 'number of neighbors' is assigned with a small value.

For the other three algorithms other than COF, it is apparent that the detectors on the contextualized data cover more area than those on the original data do. This difference indicates that contextualization enhances anomaly detection for dataset *B*, regardless of the specific algorithms. It is worth noting in Figure 3 that the curve trend of OG_CBLOF is opposite to the typical PR curve trend shown by other curves, which also happens to OG_IF in certain ranges. Besides, the detectors follow the sequence LOF > COF > CBLOF > IF in terms of their performances on the contextualized data. Additionally, as shown in Table 3, all the detectors present decent performances on the contextualized data with the respective AUC_PR values of 0.941, 0.934, 0.905, and 0.868. In contrast, except for IF, all the detectors' performances on the original data are significantly inferior. IF detectors possess the minimum gap of performances, implying that the severely worse performances of other detectors may result from the distance estimation in the algorithms' mechanisms since IF is the only algorithm not incorporating the estimation of distances between instances. The overview of the results on dataset *B* is as follows: 1) Almost all the detectors on the contextualized data are significantly superior to their counterparts on the original data – the exception is IF with a moderate gap of 0.109; 2) LOF > COF > CBLOF > IF in terms of performance on the contextualized data, and the AUC_PR value of LOF is 0.941.

3) RESULTS FOR DATASET C

As shown in Figure 4, it is apparent that the detectors on the contextualized data cover considerably more area than the ones on the original data do. This difference indicates that contextualization enhances anomaly detection for dataset *C*, regardless of the specific algorithms. It is worth noting in Figure 4 that the PR curves' trends of OG_CBLOF and CT_CBLOF are opposite to the typical PR curve trend shown by other curves, which also happens to OG_IF in

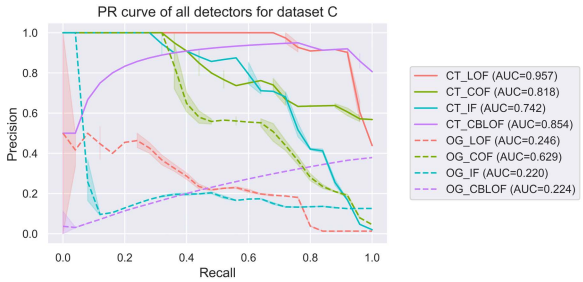

FIGURE 4. PR curves of detectors' median anomaly scores for dataset *C* (CT stands for the contextualized data, OG stands for the original data).

TABLE 4. AUC_PR values of ensemble detectors for dataset *C*.

	LOF	COF	IF	CBLOF
Original data	0.246	0.629	0.220	0.224
Contextualized data	0.957	0.818	0.742	0.854

a certain range. Besides, in terms of performance on the contextualized data, the sequence of detectors is LOF > CBLOF > COF > IF. Additionally, as shown in Table 4, all the detectors present decent performances on the contextualized data with the respective AUC_PR values of 0.957, 0.854, 0.818, and 0.742. For the original data, except for COF with mediocre performance, all the detectors' performances are poor and significantly inferior to their counterparts on the contextualized data. The overview of the results on dataset *C* is as follows: 1) All the detectors on the contextualized data are significantly superior to their counterparts on the original data; 2) LOF > CBLOF > COF > IF in terms of performance on the contextualized data, and the AUC_PR value of LOF is 0.957.

B. DISCUSSION ON OVERALL RESULTS

1) CURVE PATTERN DEVIATION

As mentioned above, in Figure 2, Figure 3, and Figure 4, some curves show the opposite trend of the majority. Specifically, for these curves, precision increases as recall increases across a partial range or the whole range of the threshold set. According to Equation (12) and (13), if the labeled anomalous instances get distinctive anomalous scores from the detectors, the increase of recall will typically come with the cost of decreasing precision. However, if multiple labeled anomalous instances are assigned very similar (or even the same) anomalous scores, the change of threshold from a higher value to a lower one will most probably increase TP along with minimal (or even zero) increases in FN and FP. This will eventually cause an increase in both recall and precision. As observed from Figure 2, Figure 3, and Figure 4, this pattern mainly happens to CBLOF and IF, especially on the original data. For CBLOF, the instances located very close to (or overlapping) each other will have very similar (or the same) distances to their nearest large cluster, leading to very similar (or the same) anomalous scores. For IF, the instances in the original data with the same energy consumption value will be isolated collectively, resulting in the same anomalous score for them.

2) COMMON RESULTS AND THE EXPLANATION FOR THEM

The first common result shared by all the three datasets is that all the detectors on the contextualized data are superior to their counterparts on the original data. The reason is that the contextual attributes (*month*, *day class*, and *hour*) added in the contextualization process help define the environment of the behavioral attribute (energy consumption) for each instance. This kind of environment information redefines the correlations between the instances. In contrast, the behavioral attribute is the only information for estimating the correlations between the instances in the original data. Without the crucial temporal information being considered and added to the original data, bias may be introduced to the estimation of the k nearest neighborhoods and the neighborhoods' densities for LOF, k nearest neighborhoods and the chaining distances for COF, the path length calculation for IF, and the cluster identification for CBLOF.

The second common result shared by all the three datasets is the ranking of algorithms in terms of their performance on the contextualized data. For datasets A and B, it is LOF $>$ COF $>$ CBLOF $>$ IF. For dataset C, it is just slightly different with the positions of COF and CBLOF switched: LOF $>$ CBLOF $>$ COF $>$ IF. This is likely related to the perspectives from which these algorithms estimate anomalousness. LOF quantifies the anomalousness of instances from a local perspective since the anomaly score is based on the difference in densities between the instance's neighborhood and its neighbors' neighborhoods. COF follows the same approach except for using chaining distance rather than density. Unlike the sheer local perspective of LOF and COF, CBLOF estimates the anomaly score in a hybrid fashion of global and local perspectives. Initial k-means clustering and the following determination of large and small clusters are on the global track – all instances in the dataset are scanned for generating the cluster centroids, and all the clusters' sizes are examined to define 'large' and 'small.' On the other hand, the final calculation of CBLOF values is directly subject to the size of the instance's cluster and the distance to its nearest large cluster. For IF, the isolation mechanism of it is established from a merely global perspective. First, the random sub-sampling for each iTree is conducted throughout the whole input data. Second, every random split in an iTree based on a particular attribute can be at any value between the attribute's minimum and maximum values. Based on the performance sequence and the mechanisms mentioned above, it implies that the local approaches are superior to the global ones, at least for the datasets in this paper, in which the labeled anomalies are the anomalies within their neighborhoods, rather than extreme instances with regard to the rest of the whole data space.

To summarize, the superior performance of LOF, COF, and CBLOF on the contextualized data can be attributed to the factors at two levels – data structure level and algorithm mechanism level. At the data structure level, contextualization reconstructs the original mono-dimensional data space to a multi-dimensional one and redefines the location of each instance, leading to a much less biased estimation of distances and similarities between the instances. At the algorithm mechanism level, the algorithms with local perspectives tend to amplify the role of contextual attributes since defining 'local' and 'nonlocal' is an essential procedure for them. Thus, those algorithms can make the most of the information brought by contextualization to identify the imperceptible contextual anomalies. Furthermore, LOF is the best performing algorithm for all the contextualized data, meaning the density comparison mechanism well fits the datasets in this study. Since these three datasets are representative of the commercial buildings' energy consumption profiles, we believe that LOF will perform well on other energy data of commercial buildings. However, it is always worth examining COF because the deviation of instances can sometimes be reflected by pattern distinctness rather than density difference.

VII. CONCLUSION AND FUTURE WORK

To keep building systems efficient, robust, and safe, we proposed a novel workflow to effectively detect the imperceptible anomalies in the energy consumption profiles of buildings. The workflow was developed on two levels – data structure level and algorithm mechanism level. The focus of the data structure level was the difference in detecting effectiveness between the original and contextualized data. At the algorithm mechanism level, detecting algorithms with different perspectives of estimating anomalousness were traversed to compare their performances. The workflow was employed in a case study to detect the anomalies in three energy consumption datasets from two types of commercial buildings in three different cities. The case study demonstrated full details of the workflow, and it fulfilled two objectives. First, it accurately identified the contextual anomalies concealed beneath the time variation of the energy consumption profiles of the three buildings. The best performances were all from LOF on the contextualized datasets, and its AUC_PR values for datasets A, B, and C were 0.989, 0.941, and 0.957, respectively. Second, more broadly, it examined the joint effect of data structures and algorithm mechanisms on the performance of unsupervised anomaly detection for buildings' energy data. On the data level, all detectors on the contextualized data showed superior detecting capacity to their counterparts on the original data. On the algorithm level, there was a constant ranking of detectors concerning their detecting performances on the contextualized data. For datasets A and B, it was LOF $>$ COF $>$ CBLOF $>$ IF. For dataset C, it was unchanged except that the positions of COF and CBLOF were switched – LOF $>$ CBLOF $>$ COF $>$ IF. This pattern implies that local approaches will outperform global approaches in the cases where the aim is to detect the instances deviating from their contextual neighbors rather than the rest of the whole data. In the future, we might explore more in the data monitoring and retrieving procedure to ensure a higher resolution of the data. Therefore more granular information could be yielded from the anomaly detection. We might also identify the connections between anomalous

energy consumption patterns and specific malfunctions. This would enable us to develop prediction workflows to directly predict particular malfunctions according to the energy consumption data.

ACKNOWLEDGMENT

The authors acknowledge Amanda Fors, the Engagement Manager of Mestro AB, Sweden, for offering this collaboration opportunity. The presented work was performed as part of the Green Technology and Environmental Economics (GreenTEE) Initiative at Umeå University. This involves collaborations with companies to develop technologies and promote policy-making studies directed towards improving cities' sustainability. The authors acknowledge support from the GreenTEE platform for funding this platform.

REFERENCES

[1] *2019 Global Status Report for Buildings and Construction: Towards a Zero-Emission, Efficient and Resilient Buildings and Construction Sector*, Global Alliance for Buildings and Construction, International Energy Agency, United Nations Environment Programme, 2019.
[2] *2020 Global Status Report for Buildings and Construction: Towards a Zero-Emission, Efficient and Resilient Buildings and Construction Sector*, United Nations Environment Programme, Nairobi, Kenya, 2020.
[3] *2021 Global Status Report for Buildings and Construction: Towards a Zero-Emission, Efficient and Resilient Buildings and Construction Sector*, United Nations Environment Programme, Nairobi, Kenya, 2021.
[4] D. B. Araya, K. Grolinger, H. F. ElYamany, M. A. M. Capretz, and G. Bitsuamlak, "Collective contextual anomaly detection framework for smart buildings," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 511–518.
[5] M. Bahramian and K. Yetilmezsoy, "Life cycle assessment of the building industry: An overview of two decades of research (1995–2018)," *Energy Buildings*, vol. 219, Jul. 2020, Art. no. 109917.
[6] Y. Chen, H. Tan, J. Wu, and X. Song, "Resilient regional energy benchmarking of classified public buildings," *Energy Proc.*, vol. 142, pp. 2365–2370, Dec. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876610217358964
[7] X. Luo, T. Hong, Y. Chen, and M. A. Piette, "Electric load shape benchmarking for small- and medium-sized commercial buildings," *Appl. Energy*, vol. 204, pp. 715–725, Oct. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261917309819
[8] Z. A. Khan and D. Jayaweera, "Smart meter data based load forecasting and demand side management in distribution networks with embedded PV systems," *IEEE Access*, vol. 8, pp. 2631–2644, 2020.
[9] E. Samani, P. Khaledian, A. Aligholian, E. Papalexakis, S. Cun, M. H. Nazari, and H. Mohsenian-Rad, "Anomaly detection in IoT-based PIR occupancy sensors to improve building energy efficiency," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Feb. 2020, pp. 1–5.
[10] U. Ali, M. H. Shamsi, C. Hoare, E. Mangina, and J. O'Donnell, "Review of urban building energy modeling (UBEM) approaches, methods and tools using qualitative and quantitative analysis," *Energy Buildings*, vol. 246, Sep. 2021, Art. no. 111073.
[11] H. Fu, J.-C. Baltazar, and D. E. Claridge, "Review of developments in whole-building statistical energy consumption models for commercial buildings," *Renew. Sustain. Energy Rev.*, vol. 147, Sep. 2021, Art. no. 111248.
[12] F. Johari, G. Peronato, P. Sadeghian, X. Zhao, and J. Widén, "Urban building energy modeling: State of the art and future prospects," *Renew. Sustain. Energy Rev.*, vol. 128, Aug. 2020, Art. no. 109902.
[13] Y. Q. Ang, Z. M. Berzolla, and C. F. Reinhart, "From concept to application: A review of use cases in urban building energy modeling," *Appl. Energy*, vol. 279, Dec. 2020, Art. no. 115738.
[14] R. Yan, Z. Ma, G. Kokogiannakis, and Y. Zhao, "A sensor fault detection strategy for air handling units using cluster analysis," *Autom. Construct.*, vol. 70, pp. 77–88, Oct. 2016.
[15] R. Zhang and T. Hong, "Modeling of HVAC operational faults in building performance simulation," *Appl. Energy*, vol. 202, pp. 178–188, Sep. 2017.
[16] M. A. Hayes and M. A. Capretz, "Contextual anomaly detection framework for big sensor data," *J. Big Data*, vol. 2, no. 1, pp. 1–22, Dec. 2015.
[17] M. Ahmed and A. N. Mahmood, "Novel approach for network traffic pattern analysis using clustering-based collective anomaly detection," *Ann. Data Sci.*, vol. 2, no. 1, pp. 111–130, 2015.
[18] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller, "A unifying review of deep and shallow anomaly detection," *Proc. IEEE*, vol. 109, no. 5, pp. 756–795, May 2021.
[19] Y. Himeur, A. Alsalemi, F. Bensaali, and A. Amira, "A novel approach for detecting anomalous energy consumption based on micro-moments and deep neural networks," *Cognit. Comput.*, vol. 12, no. 6, pp. 1381–1401, Nov. 2020.
[20] G. Pang, A. van den Hengel, C. Shen, and L. Cao, "Toward deep supervised anomaly detection: Reinforcement learning from partially labeled anomaly data," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 1298–1308.
[21] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning From Imbalanced Data Sets*. Berlin, Germany: Springer, 2018.
[22] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016.
[23] M. E. Villa-Pérez, M. Á. Álvarez-Carmona, O. Loyola-González, M. A. Medina-Pérez, J. C. Velazco-Rossell, and K.-K.-R. Choo, "Semi-supervised anomaly detection algorithms: A comparative summary and future research directions," *Knowl.-Based Syst.*, vol. 218, Apr. 2021, Art. no. 106878.
[24] V. Vercruyssen, W. Meert, G. Verbruggen, K. Maes, R. Baumer, and J. Davis, "Semi-supervised anomaly detection with an application to water analytics," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 527–536.
[25] C. Fan, F. Xiao, Z. Li, and J. Wang, "Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review," *Energy Buildings*, vol. 159, pp. 296–308, Jan. 2018.
[26] T. Amarbayasgalan, V. H. Pham, N. Theera-Umpon, and K. H. Ryu, "Unsupervised anomaly detection approach for time-series in multi-domains using deep reconstruction error," *Symmetry*, vol. 12, no. 8, p. 1251, Jul. 2020.
[27] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "USAD: UnSupervised anomaly detection on multivariate time series," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 3395–3404.
[28] C. Aytekin, X. Ni, F. Cricri, and E. Aksu, "Clustering and unsupervised anomaly detection with l_2 normalized deep auto-encoder representations," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–6.
[29] T. Finke, M. Krämer, A. Morandini, A. Mück, and I. Oleksiyuk, "Autoencoders for unsupervised anomaly detection in high energy physics," *J. High Energy Phys.*, vol. 2021, no. 6, pp. 1–32, Jun. 2021.
[30] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Med. Image Anal.*, vol. 54, pp. 30–44, May 2019.
[31] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding Gaussian mixture model for unsupervised anomaly detection," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–19.
[32] J. Chou and A. S. Telaga, "Real-time detection of anomalous power consumption," *Renew. Sustain. Energy Rev.*, vol. 33, pp. 400–411, May 2014.
[33] A. da Silva, I. S. Guarany, B. Arruda, E. C. Gurjão, and R. S. Freire, "A method for anomaly prediction in power consumption using long short-term memory and negative selection," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2019, pp. 1–5.
[34] M. Gaur, S. Makonin, I. V. Bajić, and A. Majumdar, "Performance evaluation of techniques for identifying abnormal energy consumption in buildings," *IEEE Access*, vol. 7, pp. 62721–62733, 2019.
[35] X. Liu, N. Iftikhar, P. S. Nielsen, and A. Heller, "Online anomaly energy consumption detection using Lambda architecture," in *Proc. Int. Conf. Big Data Anal. Knowl. Discovery*. Cham, Switzerland: Springer, 2016, pp. 193–209.
[36] J. Luo, T. Hong, and M. Yue, "Real-time anomaly detection for very short-term load forecasting," *J. Mod. Power Syst. Clean Energy*, vol. 6, no. 2, pp. 235–243, Mar. 2018.
[37] X. Wang and S.-H. Ahn, "Real-time prediction and anomaly detection of electrical load in a residential community," *Appl. Energy*, vol. 259, Feb. 2020, Art. no. 114145.
[38] Y. Weng, N. Zhang, and C. Xia, "Multi-agent-based unsupervised detection of energy consumption anomalies on smart campus," *IEEE Access*, vol. 7, pp. 2169–2178, 2019.

[39] J. Yeckle and B. Tang, "Detection of electricity theft in customer consumption using outlier detection algorithms," in *Proc. 1st Int. Conf. Data Intell. Secur. (ICDIS)*, Apr. 2018, pp. 135–140.
[40] X. Liu, Y. Ding, H. Tang, and F. Xiao, "A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data," *Energy Buildings*, vol. 231, Jan. 2021, Art. no. 110601.
[41] C. Fan, F. Xiao, Y. Zhao, and J. Wang, "Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data," *Appl. Energy*, vol. 211, pp. 1123–1135, Feb. 2018.
[42] J. Pereira and M. Silveira, "Unsupervised anomaly detection in energy time series data using variational recurrent autoencoders with attention," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 1275–1282.
[43] L. Wang, M. Turowski, M. Zhang, T. Riedel, M. Beigl, R. Mikut, and V. Hagenmeyer, "Point and contextual anomaly detection in building load profiles of a university campus," in *Proc. IEEE PES Innov. Smart Grid Technol. Eur. (ISGT-Europe)*, Oct. 2020, pp. 11–15.
[44] X. Zhou, T. Yang, L. Liang, X. Zi, J. Yan, and D. Pan, "Anomaly detection method of daily energy consumption patterns for central air conditioning systems," *J. Building Eng.*, vol. 38, Jun. 2021, Art. no. 102179.
[45] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 93–104.
[46] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2002, pp. 535–548.
[47] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern Recognit. Lett.*, vol. 24, nos. 9–10, pp. 1641–1650, Jun. 2003.
[48] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422.
[49] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, Nov. 2018, Art. no. e00938.
[50] A. F. Agarap, "Deep learning using rectified linear units (ReLU)," 2018, *arXiv:1803.08375*.
[51] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*.
[52] L. Yu, R. Zhou, R. Chen, and K. K. Lai, "Missing data preprocessing in credit classification: One-hot encoding or imputation?" *Emerg. Markets Finance Trade*, vol. 58, no. 2, pp. 472–482, 2022.

DONG WANG received the Ph.D. degree in chemometrics from Umeå University, Sweden, in 2022. He is a data science practitioner in various industries. He has been working with different companies in Sweden to deal with their practical problems through the application of machine learning (ML) methods and data mining (DM) approaches. For example, uncovering cause-and-effect relationships for improving process control in wastewater treatment plants; identifying culprits in boiler failures in waste-to-energy plants for process safety and saving running costs; and (this study) investigating how to effectively detect imperceptible energy consumption anomalies for buildings—to save energy and keep the building systems safe and robust.

THERESE ENLUND received degrees in biology, science journalism, and computer science, turning her focus towards interpretability and the users need to turn data into actions that saves energy and lowers carbon dioxide footprint. She is a Developer and an Analyst at Mestro AB, Sweden. She is working with analysis of real estate owners energy data and implementing tools for efficient energy usage in buildings. Her research interests include machine learning, big data, visualization, and communication.

JOHAN TRYGG is a Professor in chemometrics with Umeå University, Sweden. He's a Visiting Professor in computation and systems medicine with Imperial College London, U.K.; and the Chair of chemometrics within the Swedish Chemical Society. Over the last 25 years, he has built an extensive national and international networks within advanced data analytics, high-throughput omics platforms, and computational biology. His entrepreneurial activities included AcureOmics, that focused on metabolic profiling for precision medicine that was also partner in three EU projects, including FP7 and Horizon2020 (HUMAN, BatCure, and BOLD). He has a strong academic track record with over 200 scientific publications, over 21000 citations, and over ten patents as well as having graduated ten Ph.D. students as a Main Supervisor. His research interests include advanced data analytics in life science and the use of modern data science and engineering tools to develop computational models to understand, simulate, and predict behavior of complex biological systems research. He has been an Associate Editor of *Journal of Proteome Research* (ACS).

MATS TYSKLIND received the doctoral degree in environmental chemistry from Umeå University, Sweden, in 1993. He was the Chair Professor in environmental chemistry, in 1999. He is a Professor in environmental chemistry with the Department of Chemistry, Umeå University. He has published more than 200 peer-reviewed original articles in international scientific journals. His research interests include fate and transport processes of anthropogenic pollutants, novel environmental technologies, multivariate structure-activity modeling, machine learning, and environmental systems analysis. His research has been focused on complex environmental and process modelling combining fundamental process understanding and real environmental and process applications. During recent years, he has been focusing on the complexity of processes in a systems perspective in order to obtain more sustainable and resource efficient solutions meeting future societal demands. This has also been the objectives of the research and collaboration platform Green Technology and Environmental Economics at Umeå University coordinated by Mats Tysklind.

LILI JIANG received the doctoral degree in computer science from Lanzhou University, China, in 2012. She is an Associate Professor with the Department of Computing Science, Umeå University, Sweden, and leading the Deep Data Mining Research Group. Before joining Umeå University, she was a Research Scientist at NEC Laboratories Europe, Germany, and previously a Postdoctoral Researcher at the Department of Databases and Information Systems, Max-Planck-Institut für Informatik, Saarbrücken, Germany. She has been dedicating to address academic challenges motivated from real applications by applying the state-of-the-art data science techniques and exploring novel solutions. Especially in recent years, she has been focusing on AI-enhanced knowledge harvesting by applying her data science and artificial intelligence expertise to address the real-world challenges motivated by the pressing need of sustainability and responsibility. Her research interests include text mining, information retrieval, natural language processing, machine learning, and privacy preservation.