IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 26, 2019, accepted September 22, 2019, date of publication September 25, 2019, date of current version October 7, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2943667*

# Sample Efficient Home Power Anomaly Detection in Real Time Using Semi-Supervised Learning

**XINLIN WANG**[1], **INSOON YANG**[2,3], **(Member, IEEE), AND SUNG-HOON AHN**[1,4,5]
[1] Department of Mechanical and Aerospace Engineering, Seoul National University, Seoul 08826, South Korea
[2] Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea
[3] Automation and Systems Research Institute, Seoul National University, Seoul 08826, South Korea
[4] Innovative Technology and Energy Center, Arusha, Tanzania
[5] Institute of Advanced Machines and Design, Seoul National University, Seoul 08826, South Korea

Corresponding authors: Insoon Yang (insoonyang@snu.ac.kr) and Sung-Hoon Ahn (ahnsh@snu.ac.kr)

This work was supported in part by the International S&T Cooperation Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (MSIP) under Grant NRF-2017K1A3A9A04013801, in part by the Basic Research Laboratory Program through NRF funded by the MSIT under Grant 2018R1A4A1059976, and in part by the Creative-Pioneering Researchers Program through Seoul National University.

**ABSTRACT** Anomaly detection in home power monitoring can be categorized into two main types: detection of electrical theft, leakage, or nontechnical loss and monitoring anomalies in the daily activities of residents. Focusing on the application and practicality of anomaly detection, we propose sample efficient home power anomaly detection (SEPAD) with improved monitoring performance in terms of electricity usage as well as changes in the daily living activities of residents via provision of detailed feedback. SEPAD consists of two classifiers: an appliance pattern matching classifier (APMC) and an energy consumption habit classifier (ECHC). The APMC uses a single-source separation framework based on a semi-supervised support vector machine (semi-SVM) model. This semi-supervised learning method requires only a small amount of labeled data to achieve high accuracy in near real time and is a sample efficient detection method. The hidden Markov model (HMM)-based ECHC improves the rationality of SEPAD by providing anomaly detection functionality with respect to the daily activities of householders, especially the elderly and residents in developing areas. When SEPAD detects the appearance of an unknown pattern or known patterns contrary to the household's electricity usage habits, it triggers an alarm. SEPAD was applied to monitor power consumption data from Mkalama, a rural area in Tanzania with 52 households containing nearly 150 occupants connected to a solar powered off-grid network. The results of the practical test demonstrate the high accuracy and practicality of the proposed method.

**INDEX TERMS** Anomaly detection, power monitoring, support vector machine, semi-supervised learning.

## I. INTRODUCTION

Currently, an emphasis on environmentally friendly practices has prompted society to seek more sustainable energy practices, with ambitious targets being set by many countries in an effort to achieve significant energy savings [1]–[4]. Home power usage, along with that of the industrial and commercial sectors, is one of the major routes to reaching these targets. According to a 2018 study by the Energy Information Administration, home electricity consumption accounted for 38.5% (1.46 trillion kWh) of the total annual electricity consumption in the United States [5]. Therefore, monitoring home power consumption is a critical step toward lowering carbon emissions and reducing anthropogenic climate change effects, as well as a viable first step in curbing unnecessary electricity use [6]–[8]. Anomaly detection can monitor energy waste through load usage. Additionally, a good home power anomaly detector can improve the quality of life of residents by checking for anomalies related to health and well-being.

In this paper, we propose sample efficient home power anomaly detection (SEPAD) that has two main applications: detecting anomalies in electricity usage attributable to theft, leakage, or nontechnical loss and monitoring resident daily activities. In terms of the first application, previous research has shown that the detection performance of a supervised classification-based anomaly detection system

The associate editor coordinating the review of this manuscript and approving it for publication was Chin-Feng Lai.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

depends mainly on labeled high-quality training data, and this dependence limits the scalability and efficiency of the system [9]. An unsupervised classification-based anomaly detection system has the disadvantage of lacking detection references. A regression model-based anomaly detection system detects anomalies based on the prediction results of a forecasting model; however, it is difficult to find an appropriate prediction model to monitor the data dynamically. In addition, deep learning is an important class of machine learning methods and is widely used in many areas such as computer vision, energy consumption estimation, and health monitoring [10], [11]. However, the reasons that deep learning is unsuitable for our work are as follows: 1) Interpretability. In the research field of home power anomaly detection, interpretability is a very important aspect. Customers should be given reasonable explanations for why/how we detect that their power use is abnormal. However, it is well known that the results of deep learning are difficult to interpret for why/how a specific outcome is justified; 2) Practicality. Our work employs existing classification/clustering algorithms to overcome the limitations of each algorithm and proposes an efficient anomaly detector. The proposed data clustering scheme not only saves training cost, but also simplifies calculation process. Meanwhile, the semi-SVM-based pattern matching proposes a new pattern matching approach which has low computational complexity and computational cost. After an actual test on a desktop computer with a relatively low hardware configuration, both the detection accuracy and computing speed of our work are demonstrated to be efficient. Thus, advanced computing infrastructures are not required for using the proposed method.[1]

Regarding the second application of the anomaly detector, i.e., detecting anomalies in the daily activities of residents, prior research has focused mostly on static analysis of the daily routines of householders as opposed to using a ‘smart’ design with real-time feedback that addresses changes in health status. The purpose of this study is to fill these knowledge gaps by introducing SEPAD, a hybrid learning-based anomaly detection method equipped with an appliance pattern matching classifier (APMC) and an energy consumption habit classifier (ECHC). The APMC is used to test whether the monitoring data belong to a learned pattern, which can directly monitor power consumption data from a smart meter; this classifier is based on a single-source separation framework that uses a semi-supervised support vector machine (semi-SVM) model, in which only a small number of labeled samples are needed to achieve higher accuracy. The ECHC is based on a hidden Markov model (HMM) that uses data binning technology to process the data; the ECHC depends on residents’ electricity usage habits/readings to detect the home power usage. In terms of home electricity usage, any results that are contrary to the usage habits of residents may be indicated as anomalies or even health emergencies. In detection, either of these two classifiers is detected as anomalous, SEPAD triggers an alarm. The contributions of this work can be summarized as follows:

1) We focus on developing an anomaly detector for home energy usage and highlight its flexibility and practicality. SEPAD has low requirements regarding computer hardware and can be applied to any device with a visual interface. Moreover, a series of tests demonstrate that SEPAD can achieve real-time monitoring with high detection accuracy.
2) The proposed SEPAD is based on a semi-SVM model, which is a sample efficient classification method, that provides a simple but efficient way to reduce training costs and uses a small quantity of labeled data to perform classification with high accuracy.
3) SEPAD employs a two-dimensional monitoring method, that checks for anomalies in the daily living activities of residents according to the electricity usage habits of a household. Moreover, the results can be further applied to the field of health monitoring to provide residents with additional information about their lifestyle and health.
4) SEPAD carefully combines classification and clustering algorithms to overcome the limitations of each algorithm and to improve the overall detection performance and sample efficiency. Our work provides useful guidelines for the application of machine learning technology to home power anomaly detection, particularly when a computing infrastructure with high specifications is unavailable.
5) SEPAD was applied in Mkalama (longitude and latitude: 37.2608067, -3.4613795), a rural area in Tanzania with 52 households containing nearly 150 occupants. The households lacked a sustainable energy source and did not have access to local healthcare. SEPAD is an off-grid power source-based anomaly detector. Its successful operation addresses the need for improving quality of life in remote areas through better power monitoring and anomaly detection accuracy.

The rest of this paper is organized as follows. Section 2 introduces related work, and Section 3 presents a system overview. Section 4 describes the data selection phase. Sections 5 and 6 examine the APMC and ECHC, respectively, with comparative results presented in Section 7. Section 8 summarizes our work.

## II. RELATED WORK

Based on extensive research in the literature, current anomaly detectors applied for power usage can be categorized into three main types: 1) classification-based, 2) regression-based, and 3) others.

Classification-based anomaly detection can be subcategorized as supervised classification, unsupervised classification and semi-supervised classification. Nagi *et al.* [12] presented an SVM-based nontechnical loss detection system.

[1] On the other hand, the computational cost of deep learning methods is high in general. Furthermore, a large amount of data are needed in the training phase of deep learning, unlike the proposed method.

Using feature selection and extraction function results to train the SVM classifier, abnormal load patterns could be identified with relatively high precision. However, the time cost of offline training and whether this detector could be applied to real-time detection, were not mentioned in this work. Depuru *et al.* [13] presented a novel SVM-based electricity theft detection algorithm that combines an SVM model and a rule engine to classify customers as genuine or illegal customers; the algorithm operates at a relatively fast operating speed. Additionally, the rule engine improved the efficiency of the detector; however, numerous parameters must be set in advance. The practicability of supervised learning-based anomaly detection was further improved by Makonin *et al.* [14]; they proposed an HMM-based nonintrusive load monitoring system with the ability to disaggregate appliances with complex multistate power signatures, preserving dependency between loads in real time. To address the data imbalance caused by insufficient training data in supervised learning algorithms, Jokar *et al.* [15] proposed a hybrid learning model-based electricity theft detector, which used a cascade classification and clustering method; in the training phase, k-means clustering and silhouette plots were applied to determine the number of clusters, from which an SVM-based classifier was built. Finally, oversampling could be used to equalize the number of benign and attack samples. Although the effectiveness of the supervised classification-based anomaly detector has been verified, its practical value is limited by the amount of time required to obtain high-quality training data and the amount of resources required to label all of the training data [16]. For these reasons, Fan *et al.* [9] proposed an unsupervised classification-based building power consumption data anomaly detection system that uses an autoencoder to classify the data. In the data exploration phase, dominant periods and influential exogenous variables are identified by spectral density estimation and a decision tree. A neural network (NN)-based autoencoder is selected based on the data exploration results. Finally, the autoencoder is used to calculate the anomaly score of each observation. Scores 5%–10% higher than the threshold are identified as anomaly candidates. The unsupervised classification-based system provides a new method for identifying anomalous data in real-time. However, due to the lack of anomaly references, this approach is limited because the detection results cannot be easily evaluated, and the method has weak interpretability. During the anomaly detection process, the results should be strengthened with explanations about the reasons that the data are detected as abnormal [9], [16].

Semi-supervised classification aims to achieve a balance between supervised classification and unsupervised classification and addresses the training cost problem of supervised classification and improves the interpretability of the model. Through observations, some patterns are labeled, and the algorithm seeks to identify the unlabeled clusters associated with the labeled patterns to determine whether the unlabeled clusters belong to these labeled patterns [17]–[19]. Iwayemi and Zhou [16] invented a semi-supervised learning based residential appliance annotator. Dynamic time warping (DTW) is utilized to calculate the distance between the unlabeled data and all the labeled instances, and the Mahalanobis distance is employed to identify the boundaries of appliance clusters. After labeling all the unlearned data, a k-nearest neighbor (kNN) model is used to learn all the labeled data. Yan *et al.* [20] developed a semi-SVM-based anomaly detector for recognizing air handling unit faults. To address the insufficient fault samples of air handling units in real-world industrial applications, the proposed method divides the original dataset into training and testing sets, which contain different fault samples. By iteratively inserting new testing samples to train the SVM model, and comparing the classification results of each interaction with a preset threshold, the training pool can be enriched.

A regression-based anomaly detector estimates anomalies by comparing differences between historical predicted data and actual data. Zhang *et al.* [21] proposed a linear regression anomaly detector that accounted for the effect of temperature on home power consumption, from which different linear models were built; the linear regression detector used an F-test to determine the most suitable number of linear pieces in the regression model, and then applied the prediction result from the model as a baseline for comparison with actual power consumption data. If the real power consumption is far below the baseline, it is classified as an anomaly. The regression detector provides a new method for predicting the electricity usage of residents, taking into consideration the environmental conditions. However, this detector may not be suitable in regions that show little environmental change. Using a simple linear regression model, Chou and Telaga [22] presented a real-time building power anomaly detector based on a hybrid regression forecasting model, an artificial neural network (ANN) and autoregressive integrated moving average (ARIMA) model. The main contribution of this work was that the proposed system could monitor smart meter data in real time. Here, the detection result depends mainly on predictions based on previous data; thus, the result is limited by the selection of a suitable prediction model, a long-standing issue in this field. Regarding power consumption data prediction, autoregression may not be appropriate given that the electricity usage of residents mainly depends on their usage habits as opposed to previous usage status.

There are several other types of anomaly detectors. Cabrera and Zareipour [23] proposed an association rule learning-based anomaly detector to identify power waste patterns in educational institutions. Data binning is used initially to smooth noisy data and reduce the data size. Five determination rules are used to identify an energy waste pattern. This static analysis approach detects anomalies with full consideration of the surrounding environment. Hu *et al.* [24] presented a meta-feature based anomaly detector that targets anomalies in time series. Compared with detecting the anomalies in complex original time series directly, the proposed method first locates the data in the meta-feature space and then uses the simplified results to detect the anomalies.

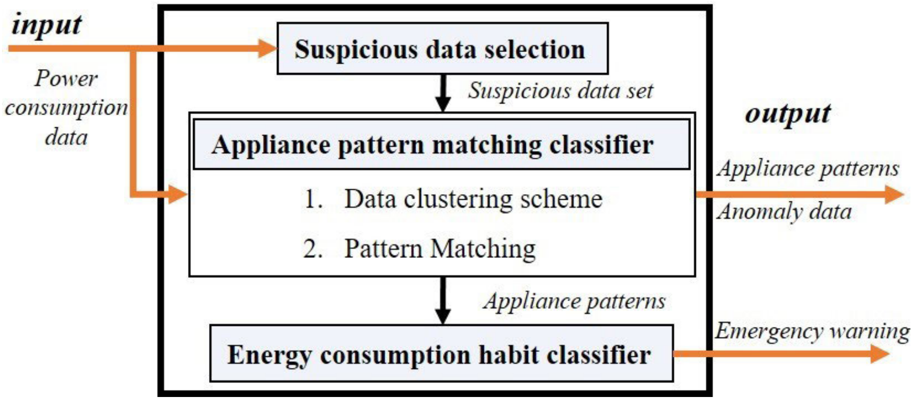


**FIGURE 1.** Overview of the proposed method.

Beyond general anomaly detection, which only detects electrical theft, leakage, or nontechnical loss, another important application of home power anomaly detection is identifying anomalies in the daily living activities of residents, e.g., the residents' unusual health conditions deduced from their atypical energy consumption patterns. Health monitoring is important given the growing proportion of single-occupancy households and the poor medical infrastructure in developing countries. Compared to traditional smart home monitoring systems, which require the installation of additional sensors that tend to invade the user's privacy, power consumption data-based anomaly detectors provide nonintrusive monitoring with improved scalability. By monitoring the meter load, the daily living activities of residents in their own homes can be detected, which can be further used as a proxy to the residents' health conditions [25]. Rahimi *et al.* [26] developed a kNN model-based nonintrusive load monitor. By identifying different electrical appliances at home, the proposed method can efficiently monitor the occupant's activities of daily living. However, as discussed above, the limitation of this work may be the training cost. To achieve high recognition accuracy, all devices need to be operated and measured separately in the training phase. Alcalá *et al.* [25] presented an energy disaggregation based nonintrusive health monitoring system. First, an HMM-based appliance detector identifies appliance usage from smart meter data, and then use frequencies are calculated using a log Gaussian Cox process model. Finally, the proposed algorithm learns the usage pattern of each household and issues a warning when any deviation from the learned pattern is detected. This work presents a new approach to improve the research on home health monitoring. However, the detecting speed and training cost are not discussed in this work. Hori *et al.* [27] introduced a power consumption-based home health monitoring system in which a kNN model was used to monitor anomalies in real time. By dividing time into different time zones, a larger anomaly score in a specific time zone may indicate the occurrence of abnormal events. However, the proposed system requires a number of parameters be set in advance, which may limit its application.

## III. SYSTEM OVERVIEW

We designed an anomaly detector to imitate human decision-making in the event of abnormal data. The framework of SEPAD is illustrated in Figure 1. First, monitored data with large deviations from the other observed samples are identified as candidate suspicious data. This stage is used for preliminary screening to identify suspicious data for later verification. Second, the suspicious data are then analyzed to determine the reasons for the differences. Here, the suspicious data are assessed via a two-step process involving an APMC and ECHC. The APMC is employed for classification and matching of processed data. The matched patterns in the APMC cycle are binned and sent to the ECHC, where power usage is detected depending on the electricity usage habits of residents. Figure 2 illustrates the proposed anomaly detection process. The input data set (Figure 2 (a)) is the power consumption data of one household in the test interval. After the suspicious data selection step, the candidates of suspicious data have been screened out as shown in Figure 2 (b). Then, the input data set is assessed by APMC and ECHC. In APMC, the data set is divided into three groups, and according to the results of semi-SVM-based pattern matching, the three groups belong to two learned patterns. The final detection result (Figure 2 (c)) shows that there is no anomaly.

## IV. SUSPICIOUS DATA SELECTION

As previously mentioned, we focus on improving the flexibility and practicality of SEPAD. Therefore, to minimize manual intervention while ensuring accuracy, k-means clustering and z-score analyses are used in this phase. k-means clustering assigns $K$ data samples into $k$ disjoint clusters, and $k$ is the target number of clusters, a preset parameter [28], [29]. Compared to other clustering methods, the computation speed of k-means is relatively faster; additionally, k-means clustering is easy to implement. The z-score is a standardization tool

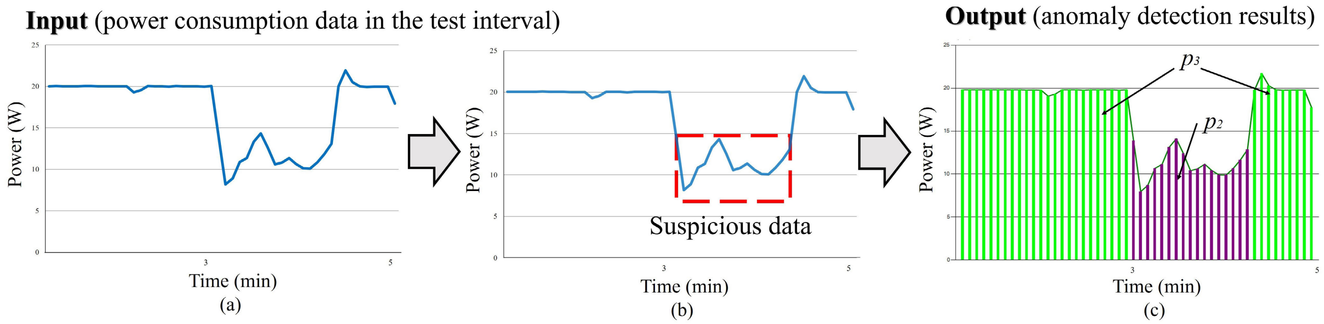


**FIGURE 2.** Illustration of the proposed anomaly detection process.

widely used to compare observations to a theoretical deviate. The z-score function is defined as [30]:

$$Z_{score}(x_{j_n}, \bar{x}, \sigma) = \frac{x_{j_n} - \bar{x}}{\sigma}, \tag{1}$$

where $\bar{x}$ and $\sigma$ are the mean and standard deviation of $C_N$.

```
Algorithm 1 Suspicious Data Selection
1  Input: Monitored data set $X \in \{x_1, \ldots, x_K\}$
2  Output: Suspicious data set $C_{suspicious}$;
3  Variables and Functions: Normal data set $C_N \subseteq X$;
4  Uncertain data set $C_U \subseteq X$;
5  Kmeans($A, k$): k-means clustering, where $A$ is the data
   set to be clustered and $k$ is the target number of clusters;
6  $Z_{score}(x_j, \bar{x}, \sigma)$: z-score function (1);
7  Suspicious Data Selection:
8  $(C_N, C_U) =$ Kmeans$(X, 2)$;
9  $\bar{x} =$ average$(C_N)$; $\sigma =$ standard deviation$(C_N)$;
10 for $n = 1 : N$ do
11     $x_{j_n} \in C_U$,
12     if $Z_{score}(x_{j_n}, \bar{x}, \sigma)) > threshold_1$ then
13         $x_{j_n} \in C_{suspicious}$
14     end
15 end
```

The detailed steps for selecting suspicious data are described in Algorithm 1. Here, $X$ denotes the monitored data set. First, Kmeans() is used to group the monitored data set $X$ into two disjoint clusters (line 8). The cluster with the most data is selected as the normal class $C_N$. Second, the mean value $\bar{x}$ and standard variance $\sigma$ of $C_N$ are calculated (line 9). Third, the z-score function is used to calculate the z-scores of all items in $C_U$, which are then compared to $threshold_1$. If the z-score of $x_{j_n}$ is larger than $threshold_1$, $x_{j_n}$ is selected as suspicious data (lines 10–15). $Threshold_1$ is a parameter, used for preliminary screening the suspicious data. It varies with $\sigma_X$ (the standard deviation of $X$). If $\sigma_X$ is large, the value of $threshold_1$ should be set to a large value, which avoids the high computational cost in the following steps. Conversely, if $\sigma_X$ is small, the value of $threshold_1$ should be set to a small value to ensure the suspicious data with large deviations from the other monitored data can be identified. In this study, the value of $threshold_1$ is selected as follows:

$$threshold_1 = \begin{cases} 0.01\sigma_X + 0.5 & \text{if } 0 < \sigma_X < 10 \\ 0.01 & \text{otherwise}, \end{cases} \tag{2}$$

where the parameters are determined based on practical consumption data from an investigation. Figure 3 shows an example of suspicious data selection. After using Kmeans() to group the processed data set $X$ into two clusters, the cluster that has more data is labeled as $C_N$ (green), and the other is labeled as $C_U$. The red points in Figure 3 represent suspicious data.

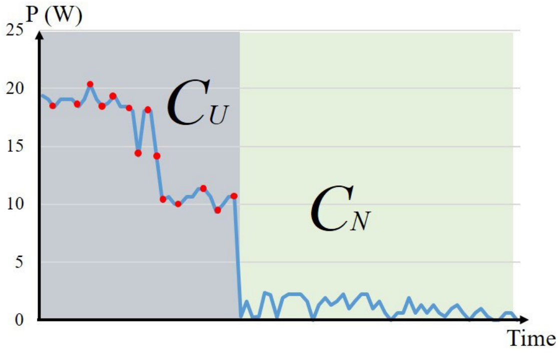


**FIGURE 3.** Example of suspicious data selection.

## V. APPLIANCE PATTERN MATCHING CLASSIFIER USING SEMI-SUPERVISED LEARNING

To improve the accuracy of SEPAD, it is important to determine the status of the target load at any given time. Suspicious data stand out due to their high deviations from the average of most of the other data. It is necessary to know whether the causes of these deviations are electricity theft, leakage, or some other reasons. The APMC is based on a semi-SVM model, which separates single power consumption data obtained from the smart meter into $k$ disjoint clusters [30]. Here, we introduce the APMC as an optimal clustering scheme followed by semi-SVM-based labeled pattern matching.

### A. DATA CLUSTERING SCHEME USING SILHOUETTE COEFFICIENT

We begin by introducing the silhouette coefficient (SC) as a measure of how similar an object is to its own cluster compared to other clusters [31]. The SC is applied in this stage to overcome the weakness of k-means clustering in setting parameter $k$ in advance. Additionally, this coefficient provides single-source separation by effectively reducing the computation associated with pattern matching. The SC function is defined as

$$SC(k) = \frac{b(k) - a(k)}{\max\left(a(k), b(k)\right)}, \tag{3}$$

where $a(k)$ denotes the mean intra-cluster distance and $b(k)$ denotes the mean nearest-cluster distance for $k$, Figure 4 illustrates an example of $a(k)$ and $b(k)$. The value of $SC(k)$ ranges from $-1$ to $+1$, with $+1$ corresponding to the best value. The detailed steps of the optimal clustering scheme are presented in Algorithm 2.

**Algorithm 2** Data Clustering Scheme

```
1  Input: Monitored data set $X \in \{x_1, \ldots, x_K\}$;
2  Suspicious data set $C_{suspicious}$;
3  Output: Clustered data sets $C_1, \ldots, C_n$, where $n$ is the
   number of clusters;
4  Variables and Functions: Kmeans$(A, k)$;
5  $Z_{score}(x_j, \bar{x}, \sigma)$: z-score function (1);
6  $SC(k)$: silhouette coefficient (3), where $k$ is the number
   of target clusters;
7  Clustering Algorithm:
8  for $i = 2 : m$ do
9      Compute $SC(i)$;
10 end
11 Set $n :=$ argmax$_i SC(i)$;
12 if $n >$ $threshold_2$ then
13     Compute Kmeans$(X, n)$, and divide $X$ into $n$ cluster;
14 else
15     Take $X$ as a single cluster, and set $n := 1$;
16 end
```

Because we do not know how many patterns appear in this monitoring, we assume that there may be at most $m$ groups (with no prior knowledge, $m$ is set to 9). Using $i$ to represent each scenario, $i$ ranges from 2 to $m$. In every scenario, Kmeans() is used to cluster the data, and the SC function is used to evaluate the results of the clustering scenario. After $m$ cycles of clustering, we obtain $m$ SC scores, and then select the maximum value $SC(m)$ among these scores (lines 8–11). $SC(n)$ is then compared to $threshold_2$,[2] if $SC(n)$ is larger than $threshold_2$, the number $n$ of clusters is accepted and k-means clustering Kmeans$(X, n)$ is performed to obtain $n$ clusters. Else, all of the data belong to a single cluster, and $n$ is set to 1 (lines 12–16).

### B. SEMI-SVM-BASED PATTERN MATCHING

After grouping the data into $n$ clusters, the next step is to determine if these clusters have been learned before. Because the semi-SVM model presented in this study is a sample efficient classification method that does not need a large number of training data, only a small amount of data are labeled as the training set. Additionally, to study the effect of different kernel functions on the classification results, this paper proposes multiclass SVM models with two kernel functions: a Gaussian function and a linear function. Algorithm 3 describes the detailed steps of semi-SVM-based pattern matching.

After dividing the data into $n$ clusters, the algorithm cycles $n$ times and executes the same script in each cluster. In each cluster, first, the mean value and standard variance of the cluster are calculated (lines 12–13); second, the z-score is used to select the center point $D_{c_i}$ and the edge point $D_{e_i}$; and third, the test point generation function Test() is used to calculate the test point for classification, where

$$\text{Test}(\bar{x}_i, D_{e_i}, D_{c_i}) := \alpha \bar{x}_i + \beta D_{e_i} + \delta D_{c_i} \tag{4}$$

Here, $\alpha, \beta$ and $\delta$ are constants, and $\alpha + \beta + \delta = 1$.[3] The test point designed in this phase is used to improve the classification efficiency. Different from other works using distance measurements to find the similarity/dissimilarity between two data sets, in our work, we calculate a test point to represent all of the data in the cluster. The accuracy and practicability of this method are discussed for comparison results in Section VII. The kNN model is then used to find the closest group to the test point in the training set. kNN is a simple classification method that is widely used due to its simplicity and practicability. Here, $k$ is a positive integer, corresponding to the number of nearest neighbors to be used in the decision. In this study, we used the simplest kNN model 1-NN [15], [32].

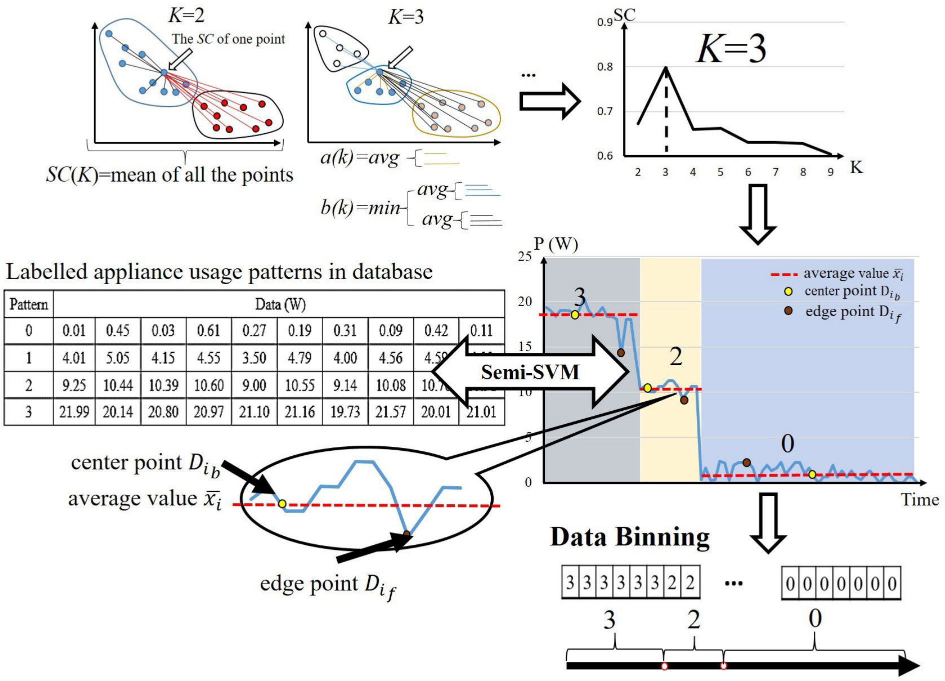
The SVM algorithm is used to classify the training set. The cross-entropy loss function Loss() is used to calculate the classification loss, $loss_1$. The cross-entropy loss function is widely applied in evaluating classification and clustering performance by virtue of its simplicity, accuracy and adaptability in global optimization. This function is defined as follows [33]:

$$\text{Loss}(y, \widehat{y}) = \sum_{i=1}^{T} y_i \log(\widehat{y}_i), \tag{5}$$

where $y$ represents the real labeling results in the training dataset, $\widehat{y}$ denotes the classification results of SVM, and $T$ is

[2] $threshold_2$ is a preset parameter, ranges from 0 to 1. The determination of $threshold_2$ is related to the variance of the training set. In general, to avoid clustering error, the smaller the variance is, the higher the value of $threshold_2$. In this study, $threshold_2$ is set to 0.7.

[3] Test() is a weight function, which calculates the test point by considering the mean value of the cluster, the point with maximum deviation and the point with minimum deviation. In this study, through investigating the influence of the value of $\alpha$, $\beta$ and $\delta$ on the classification results, $\alpha$ is set to 0.4, $\beta$ is set to 0.3, and $\delta$ is set to 0.3.



Labelled appliance usage patterns in database

| Pattern | Data (W) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.01 | 0.45 | 0.03 | 0.61 | 0.27 | 0.19 | 0.31 | 0.09 | 0.42 | 0.11 |
| 1 | 4.01 | 5.05 | 4.15 | 4.55 | 3.50 | 4.79 | 4.00 | 4.56 | 4.59 | [illegible] |
| 2 | 9.25 | 10.44 | 10.39 | 10.60 | 9.00 | 10.55 | 9.14 | 10.08 | 10.7[illegible] | [illegible] |
| 3 | 21.99 | 20.14 | 20.80 | 20.97 | 21.10 | 21.16 | 19.73 | 21.57 | 20.01 | 21.01 |

**FIGURE 4.** Appliance pattern matching classifier (APMC).

the size of the training dataset. Lower values of cross-entropy loss indicate better classification results. In this study, two kinds of kernel functions can be selected; their effect on classification accuracy is discussed later.

The test point is then labeled by using the kNN result and placed into the training set. With the updated training set, the SVM algorithm is again used and $loss_2$ is calculated. If $loss_2$ minus $loss_1$ is smaller than $threshold_G$, it proves that the test point belongs to the labeled set and the other data in this cluster also belong to this labeled set. If the cluster and $C_{candidate}$ have common data, the common data in $C_{candidate}$ are removed and SEPAD places the test point into the training set. Else if $loss_2$ minus $loss_1$ is larger than $threshold_G$, then the cluster has an unknown pattern. If the cluster and $C_{candidate}$ have common data, the common data are placed into $C_{anomaly}$, and the system issues a warning.[4]

Figure 4 illustrates the key steps in the APMC. After the optimal grouping phase, the processed data set is divided into three clusters. Among the three clusters, the classifier calculates the mean of each cluster, screens out the center point and edge point in every cluster, and generates the test point of the cluster. In the next step, the kNN model is used to select the closest cluster to each of the three clusters among the training set. For ease of discussion, the different appliance patterns are denoted as numbers. The result of this step shows that the first cluster is closest to pattern 3, the second cluster is closest to pattern 2, and the last cluster is closest to pattern 0. Finally, the semi-SVM model is used to verify the results of the aforementioned stages; the model indicates that all three clusters match a labeled pattern (3, 2, and 0, respectively).

## VI. ENERGY CONSUMPTION HABIT CLASSIFIER

In the APMC, the monitored data set is divided into different groups, and each group is verified by a semi-SVM-based pattern matching classifier. The outputs are the unknown power usage patterns and the anomalous data set. In the ECHC, even suspicious data can be classified into a learned appliance usage pattern; however, the anomaly detector must still determine whether the matched pattern depends on the household's electricity usage habits [25].

### A. DATA BINNING

Data binning is widely applied in denoising and concept hierarchy generation, and this process divides data into different buckets or bins [23]. In this study, the bin value is the labeling of matched patterns in the APMC. Figure 4 illustrates an example of data binning.

### B. HMM

Many classifiers can be used for classification, such as SVMs, decision trees, and NNs. However, these classification approaches do not satisfy the characteristics of home power consumption data, which ignore the temporal relations that exist in a time series [34]. An HMM is a type of Markov chain in which the target states are hidden and can only be predicted. The Markov chain itself is a mathematical model consisting of a number of states and computable probabilities. The Markov chain has the property that the present state

[4] $threshold_G$ is a preset parameter, which is determined based on the loss of classification from an investigation. In this study, $threshold_G$ is set to 0.2.

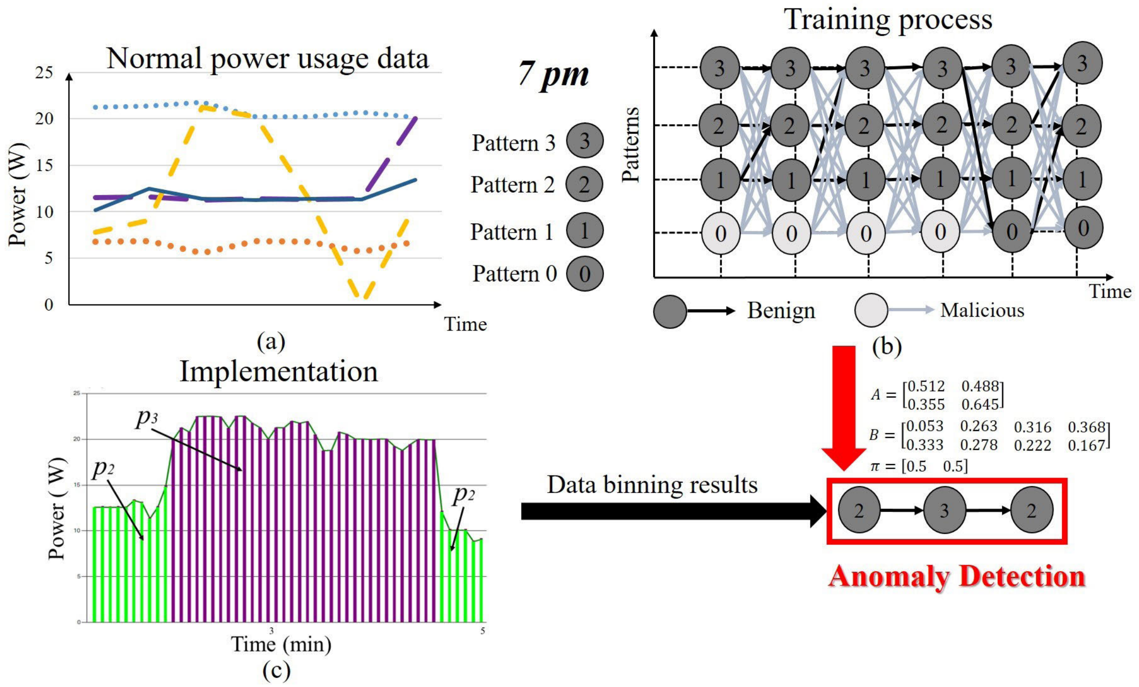


**FIGURE 5.** Energy consumption habit classifier (ECHC).

depends only on the previous state, as opposed to a sequence of states [25], [35], [36].

### 1) DATA PROCESSING

To balance training dataset quality with system accuracy, the training data in this phase are processed separately (in hours). Therefore, there are 24 training datasets, and each training dataset consists of a benign dataset and a malicious dataset. Because the normal power consumption data series reflect the power usage habits of residents, these series are regular. Therefore, we select the most representative power consumption data series as the normal data series. The selected data series are first divided into 24 training sets according to chronological order. Then, in each training dataset, the APMC is used to classy the series into different patterns. At last, the data binning is employed to process the results of APMC, and the obtained chains are used as the benign dataset. Because the algorithm is trained to estimate abnormal power usage based on the normal usage habits of householders, all other chains different from the benign dataset can be considered as the malicious dataset. In the training phase, to prevent the classification error caused by imbalanced data and reduce the training cost, the malicious dataset usually contains the same amount of samples as the benign dataset.

### 2) TRAINING

To describe the HMM, we define the following:

1) Set of possible states:
$S = \{s_1, \ldots, s_N\}$, where $s_i$ is a possible state, and $N$ is the number of states. Because this work focuses only on identifying anomalies, there are two states, i.e., $N = 2$, and $S = \{0, 1\}$, where 0 means normal and 1 means abnormal.
2) Set of observations:
$P = \left\{p_0, p_1, p_2, \text{âĂę}, p_{M-1}\right\}$. The observation variables in this work are the appliance usage patterns in training dataset $L$. Additionally, $M$ is the number of patterns.
3) Observation sequence:
$O = \left\{Dataset_{benign}, Dataset_{malicious}\right\}$.
Where, $Dataset_{benign}$ and $Dataset_{malicious}$ denote the benign dataset and malicious dataset. The length of the observation sequence in this study is the length of the results of data binning in the training phase.

The training data are used to build the following matrices:

1) Transition probability matrix:
$A = [a_{i,j}]_{n \times n}$, where $a_{i,j} = \text{Prob}(s_i|s_j)$ is the probability that the next state is $s_i$ given that the current state is $s_j$.
2) Measurement output probability matrix:
$B = [b_{i,j}]_{n \times m}$, where $b_{i,j} = \text{Prob}(P_j|s_i)$ is the probability of observing $P_j$ given that the current state is $s_i$.
3) Initial state probability vector:
$\pi = (\pi_1, \ldots, \pi_N)$, where $\pi_i = \text{Prob}(Y_1 = s_i)$, where $Y_1$ denotes the state at stage 1.

Therefore, the HMM is defined as a 3-tuple, $(A, B, \pi)$. Figure 5 shows the training process of HMM. Five data chains (Figure 5 (a)) were selected as the normal data series for 7 PM, and the corresponding data binning results were used as the benign dataset as shown in Figure 5 (b). After selecting five chains which are different from the benign dataset as the malicious dataset, $(A, B, \pi)$ were obtained. Figure 5 (c) shows the implementation of ECHC, the trained model is used to detect the results of APMC from 7 PM to 8 PM.

### 3) IMPLEMENTATION

1) Test sequence:
$T = \{Z_{previous}, Z_{now}\}$, where, $Z_{previous}$ and $Z_{now}$ denote the previous and monitored (current) patterns, respectively. The length of the observation sequence in this study is two. Every matched pattern in Algorithm 3 is composed of an observation sequence with the matched pattern in previous step, and $\{Z_{previous}, Z_{now}\} \in P$.
2) Detection result:
$Y_t$ = the final result of the ECHC. If $Y_t$ is 0, the test sequence depends on the learned energy consumption habits, and this power use is under the normal condition. Conversely, if $Y_t$ is 1, the ECHC judges that the test sequence does not depend on the learned habits. Eventually, SEPAD issues a warning to the manager to take further action. The ECHC selects the appropriate training set to train the HMM based on the current time $t$.

As shown in Figure 5 (c), in this test, the matched pattern chain in Algorithm 3 was $\{p_2, p_3, p_2\}$; therefore, the system executed three times, and the inputs were $\{p_2, p_2\}$ at 7 PM, $\{p_2, p_3\}$ at 7 PM, and $\{p_3, p_2\}$ at 7 PM. It should be noted that, in the first input $\{p_2, p_2\}$, the previous pattern was the most recently matched pattern. The results of the first two tests were 0. However, in the last time, $Y_t$ was 1. Finally, the results of ECHC showed that the power usage is abnormal, which does not depend on the learned usage habits.

## VII. EXPERIMENTAL RESULTS

The data used in our experiments are obtained from Mkalama. The wiring diagram of Mkalama is shown in Figure 6. A radio frequency-based wireless energy monitoring system was built to monitor the power consumption of each household in the village, and the time resolution of measurement is 1 second. This system consists of two parts: a local energy monitoring part that resides in Tanzania and a remote part, i.e. the proposed SEPAD, located in South Korea. The home power consumption data are collected in Tanzania and transmitted to South Korea at a controllable test interval. Considering the underdeveloped communication infrastructure of Mkalama and ensuring the timeliness of detection, the test interval is usually set from 5 to 30 minutes. In South Korea, after receiving the data, SEPAD identifies anomalies in real time. SEPAD has a low computer hardware requirement; it is implemented using a desktop computer with a 3.4GHz Intel Core i5 processor, 4 GB RAM, and the Windows 10 operating system. To further simplify data mining, the detector design is based on C# and Python. In addition to processing the data, C# is used to build the visual interface, providing more options to users. Figure 7 shows the overall power consumption of Mkalama from July 2018 to March 2019. Benefiting from the detailed power consumption feedback and the inference of energy consumption habits, with the application of SEPAD, the electricity usage of Mkalama has increased by 238.10 %. An increasing number of residents in villages have begun to use electricity. This increased access to electricity may contribute to improving the quality of life. To further demonstrate the performance of SEPAD, two experimental cases are presented: a normal monitoring case and a test case.

**Algorithm 3** Semi-SVM Based Pattern Matching

```
Input: Suspicious data set $C_{suspicious}$ obtained from Algorithm 1;
Training data set $L$;
Grouped clusters $C_1, \dots, C_n$;
$y$: labels of training data set
Output: Anomalous data set $C_{anomaly}$;
Variables and Functions:
SVM(): support vector machine;
Test(): test point generation function (4);
Loss(): classification loss function (5);
kNN($k, x, L$): $k$-Nearest Neighbor, where $x$ is the test point and $L$ is the training data set;
Pattern Matching:
Initialize $C_{candidate}$ as $C_{suspicious}$;
for $i = 1 : n$ do
    $\bar{x}_i$ :=average($C_i$);
    $\sigma_i$ :=standard deviation($C_i$);
    $D_{e_i}$ := $\text{argmax}_{x_c \in C_i} Z_{score}(x_c, \bar{x}_i, \sigma_i)$;
    $D_{c_i}$ := $\text{argmin}_{x_c \in C_i} Z_{score}(x_c, \bar{x}_i, \sigma_i)$;
    $D_{t_i}$ :=Test($\bar{x}_i, D_{e_i}, D_{c_i}$); (test point generation)
    $L_i$ :=kNN($1, D_{t_i}, L$);
    $\widehat{y}$ : prediction results of SVM($L$);
    $loss_1$ :=Loss($y, \widehat{y}$);
    Label $D_{t_i}$ as $L_i$, and place it into the training set;
    $\widehat{y}$ : prediction results of SVM($L$);
    $loss_2$ :=$Loss(y, \widehat{y})$;
    if $(loss_2 - loss_1) < threshold_G$ then
        Place $D_{t_i}$ into the training set;
        $C_{candidate} := C_{candidate} \setminus C_i$;
    else
        $C_{anomaly} := C_{anomaly} \cup (C_{candidate} \cap C_i)$;
    end
end
```

### A. CASE 1

Case 1 is a normal monitoring case in which the same consumption data are tested by three functions of SEPAD. Test A uses only suspicious selection to screen out suspicious data; Test B uses the APMC to match the appliance patterns; and Test C estimates anomalies by using both the APMC and ECHC. In monitoring, users can choose the functions according to their own specific needs. The kernel function of the semi-SVM model is chosen to be a linear function. The parameters in the tests are listed in Table 1.

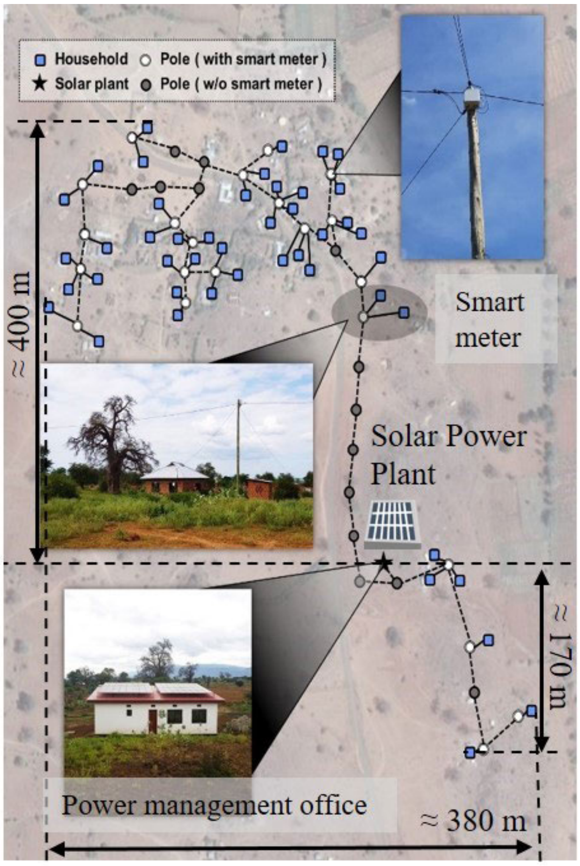


**FIGURE 6.** Wiring diagram of Mkalama.

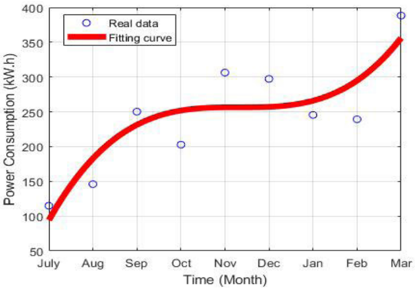


**FIGURE 7.** The overall power consumption of Mkalama from July 2018 to March 2019.

**TABLE 1.** Parameter of case I.

| Items | Value | Items | Value |
|---|---|---|---|
| $threshold_2$ | 0.7 | $threshold_G$ | 0.2 |
| $\alpha$ | 0.4 | $\beta$ | 0.3 |
| $\varepsilon$ | 0.3 | test interval | 5 min |

#### 1) TEST A

The results of Test A are shown in Figure 8 (a). The average power consumption from the monitoring data was 10.22 W, and $threshold_1$ in this case was set to 0.56. The 0th to the 29th data entries were identified as suspicious.

#### 2) TEST B

In Test B, APMC verification identified two clusters corresponding to $p_0$ and $p_2$; suspicious data belong to $p_0$, whereas the other data are categorized as $p_2$, as shown in Figure 8 (b). Therefore, SEPAD reported no anomalous data.

#### 3) TEST C

After the APMC, the ECHC was used to determine whether the matched patterns depend on any power usage habit. SEPAD showed a monitoring time of 5 PM (Tanzania time); the previous pattern matched from the last iteration was $p_0$. Therefore, the first input to the HMM was $\{p_0, p_0\}$ at 5 PM, and the output ''0'' indicated that this input was dependent on the power usage habit of the household. The second input to the HMM was $\{p_0, p_2\}$ at 5 PM; again, the output was normal. Eventually, after all phases of anomaly detection, these data were proven to be normal. Thus, the result of Test C is the same as that of Test B, as shown in Figure 8 (b).

### B. CASE 2

To test the sensitivity of SEPAD, several new electrical appliances were used during monitoring. To verify the complex data processing capabilities of the system, the test interval was 30 minute, and the data source did not change.

#### 1) TEST A

In test A, SEPAD employs both the APMC and ECHC. The kernel function of the semi-SVM model is chosen to be a Gaussian function. The results of Test A are shown in Figure 9. According to the APMC results, three appliance usage patterns $p_1, p_2$, and $p_3$ were matched in this test. SEPAD also detected and warned of the existence of an unlearned cluster. In the ECHC, after data binning, the Markov chain was $\{p_2, p_1, p_3, p_2, unknown, p_2, p_1\}$. When SEPAD monitored the third matched pattern $p_3$, the test time was 6 PM. The input of the HMM was $\{p_1, p_3\}$ at 6 PM. The prediction result of the HMM was 1, indicating an anomaly; therefore, a warning was issued by the system.

#### 2) TEST B

Test B is a comparison test that shows the detection results of the APMC using different kernel functions for the same process. The testing environment of Test B is the same as that of Test A. Figure 10 (a) shows the results obtained with a Gaussian kernel function, and Figure 10 (b) presents the results obtained with a linear kernel function. According to the Gaussian results, four clusters were identified. Three clusters were matched using the APMC $\{p_2, p_1, p_2\}$. The last cluster corresponded to a new electrical appliance. The detection results are in accordance with the actual test situation. However, using the linear kernel function, not only the first three clusters matched, but also the last cluster with the labeled pattern $p_3$. Misclassification can be avoided by modifying the threshold; however, after several actual tests,

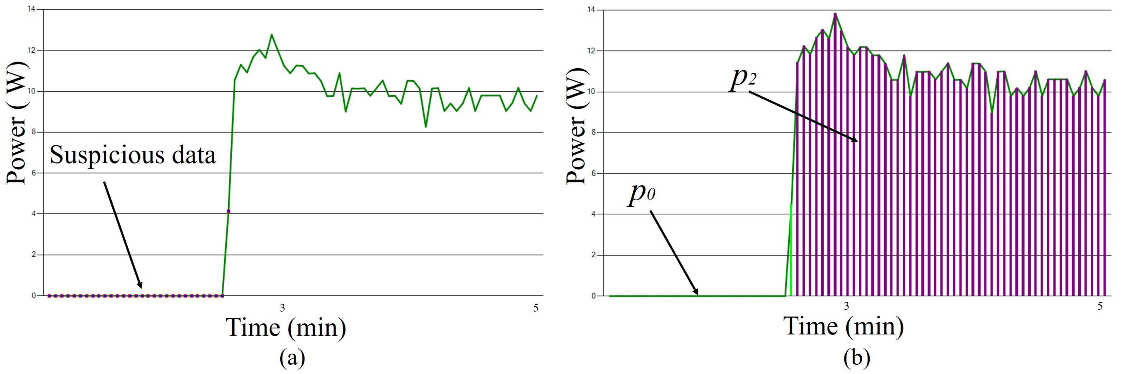


**FIGURE 8.** The result of case 1. (a) Suspicious data selection results: Some data were screened out as suspicious data. (b) Results of APMC and ECHC: There were two appliance usage patterns matched in APMC, and ECHC determined the matched patterns depending on the usage habits of this household.

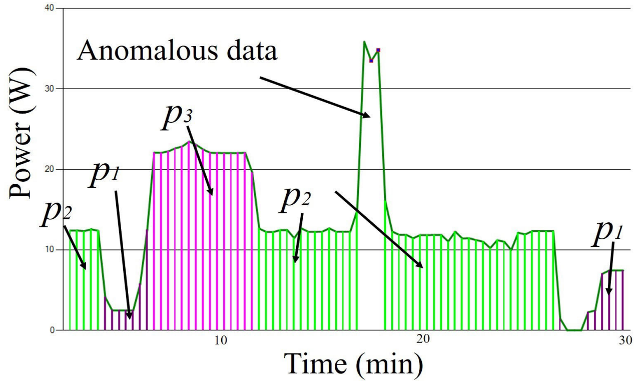


**FIGURE 9.** The result of Test A. According to the APMC results, there are three labeled appliance usage patterns and an unknown pattern in this test. Moreover, according to the ECHC results, these matched patterns do not depend on the usage habits of the household.

the Gaussian kernel function showed higher accuracy in the complex process than the linear kernel function. In addition, it should be noted that the pattern $p_2$ appeared twice at different time points in this test. However, the results based on Gaussian kernel function and linear kernel function identified them correctly, which benefits from the accurate classification results of the data clustering scheme and semi-SVM based pattern matching proposed in this study.

### C. COMPARISON RESULTS AND EVALUATION

To demonstrate how SEPAD achieves a high degree of anomaly detection in monitoring the power consumption data and detects anomalies according to the electricity usage habits of residents, we compared SEPAD to three other detectors, based on two parameters: accuracy and training cost and speed. We compared SEPAD against the results of Nagi *et al.* [12], Chou and Telaga [22], and Hori *et al.* [27].

The Nagi *et al.* method is a SVM-based anomaly detector (SVM-AD), the Hori *et al.* method is a kNN-based anomaly detector (kNN-AD), and the Chou *et al.* method is an ARIMA and ANN-based anomaly detector (ARIMA-ANN-AD). Because both the SVM-AD and kNN-AD are supervised classifiers, to further prove the sample efficiency of SEPAD, we used the same one month smart meter data as a training set to train the two supervised classifiers, and only 7 days of data to train SEPAD. The training set of the ANN model in the ARIMA-ANN-AD is same as the other supervised classifiers, and both the autoregressive terms and lagged forecast errors of the ARIMA model in the ARIMA-ANN-AD are chosen to be 1. The test data are taken from one household in Mkalama from July 2018 to March 2019. The comparison results are reported in Table 2.

**TABLE 2.** Comparison results between SEPAD and others.

| Detectors | Accuracy | $F_1$ | Cross-Validation |
|---|---|---|---|
| SEPAD (Gaussian) | 0.8730 | 0.9279 | 0.8800 (APMC) |
| SEPAD (Linear) | 0.7381 | 0.8612 | 0.8600 (APMC) |
| SVM-AD | 0.8373 | 0.9091 | 0.7800 |
| ARIMA-ANN-AD | 0.5833 | 0.7126 | N/A |
| kNN-AD | 0.8455 | 0.9140 | 0.8200 |

#### 1) ACCURACY

For a more comprehensive explanation of detection accuracy, two accuracy measurement methods, an accuracy function and $F_1$ score, are employed to verify the comparison results. The accuracy is defined as follows [37]:

$$\text{Accuracy} := \frac{TN + TP}{TN + FP + FN + TP}, \tag{6}$$

where $TP$ represents the number of time points at which normal data are correctly classified, $FP$ represents the number of time points at which anomalous data are classified as normal, $TN$ represents the number of time points at which anomalous data are detected as abnormal, and $FN$ represents the number of time points at which normal data are detected as abnormal. $F_1$ score is a measure, that provides a balanced evaluation of the overall performance of a classifier. The $F_1$ score consists of precision $p$ and recall $r$, which were developed by the information retrieval community. The precision $p$, recall $r$ and

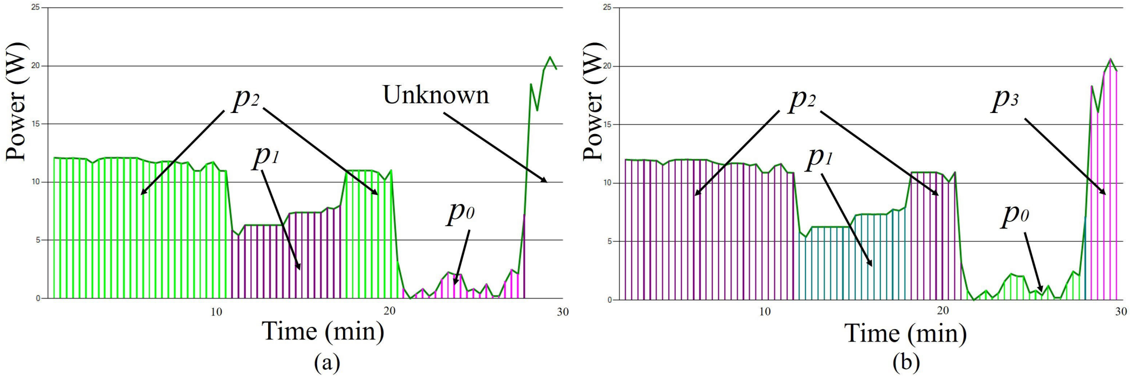


**FIGURE 10.** The result of Test B. (a) Results of the APMC with the Gaussian kernel function: The last cluster was determined to be an unknown pattern. (b) Results of the APMC with the linear kernel function: The last cluster was misclassified into a labeled pattern, P3.

$F_1$ score are defined as follows: [16]

$$p := \frac{TP}{FP + TP}, \tag{7}$$

$$r := \frac{TP}{FN + TP}, \tag{8}$$

$F_1$ score is the harmonic mean of the precision $p$ and recall $r$ as follows:

$$F_1 := \frac{2 \times p \times r}{p + r}, \tag{9}$$

As shown in Table 2, the ARIMA-ANN-AD shows both the lowest accuracy and $F_1$ score among all the detectors. After actual verification, the occurrence of random events is not easy to predict. ARIMA is an autoregression model that is not suited for nonlinear prediction. Another reason for the low accuracy of the ARIMA-ANN-AD may be the insufficient training data for the ANN, which may also explain the low accuracy of the SVM-AD. The kNN-AD show high accuracy and $F_1$ score in anomaly detection and can effectively detect abnormal changes in the data; however, for real-time monitoring, we recommend setting parameter a in the nonactivity level function to less than 10. Finally, the Gaussian kernel-based SEPAD shows both the highest accuracy and $F_1$ score among all classifiers. As opposed to the other approaches that use distance measurements to determine the degree of similarity between two data sets, the APMC, after clustering data into different groups, calculates the test point of each group for classification. The kNN model is used to find the closest pattern in the training set, and then, the semi-SVM model is used to verify the kNN result and update the training set. This method of using test points to classify and combine the semi-SVM model greatly improves the detection efficiency and reduces unnecessary calculations.

In our study, to verify the ability of the ''kNN + semi-SVM'' set to improve the accuracy and speed of SEPAD, we also compared the accuracy and running speed results with those of two other models. First, we replaced the ''kNN + semi-SVM'' set with DTW. After clustering using Algorithm 2, the DTWs between every clustered group and pattern in the training set were calculated. If the minimum DTW was less than a set threshold, the group was determined to belong to this pattern. We used this method because DTW is a well-known similarity measure for time series data and addresses the limitation of equal-length alignment. However, in actual testing, the computation speed is too slow for real-time monitoring.

The second method is to replace the kNN model with the mean value in this step. After calculating the test point, the mean values of all patterns in the training set were computed and compared to the test point. The pattern with the smallest difference between the test point and the mean was selected as the input of the semi-SVM. However, although this method showed a faster running speed, its accuracy dropped sharply. In real tests using the same data, the accuracy of this method was 0.7540 and its $F_1$ score was 0.8559. In terms of a comparison between the two kernel functions, it can be seen that the kernel function directly affects the classification result. In normal states, the accuracy and $F_1$ score of the Gaussian kernel function-based SEPAD are higher.

### 2) TRAINING COST AND SPEED

SEPAD consists of two parts: the APMC and ECHC. Because the APMC is based on a semi-SVM model, the training cost is significantly lower than that of the other classifiers, and this effect provides a sample efficient classification method and improves the application availability. The ECHC is an HMM-based classifier. Although we created the training set in hours to reduce the training cost, the training cost of the ECHC was still high. To further evaluate the semi-SVM-based APMC, the K-fold cross-validation was employed to test all the classifiers with 50 data samples containing different patterns with $K = 5$. The last column in Table 2 exhibits the mean accuracy of each classifier. Because the ARIMA-ANN-AD identifies anomalies based on the difference between the regression result and real data, the K-fold cross-validation is not suitable for evaluating this model. Finally, the semi-SVM based APMC shows the highest accuracy among all the methods. In terms of calculating speed, all of the compared anomaly detectors showed high running speeds. SEPAD can realize near real-time monitoring of less than 1 second.

## VIII. CONCLUSION

In general, it is a challenging task to monitor appliance usage patterns and anomalies in the daily activities of residents using only power readings at a main connection. In this work, we have proposed a real-time anomaly detector, SEPAD, with improved monitoring performance with respect to electricity usage, as well as the usage habits of householders via the provision of detailed feedback. The semi-SVM based APMC uses a small amount of labeled data to disaggregate power data with high accuracy and high computation speed. The ECHC monitors home power, using an HMM to optimize the anomaly detector to monitor and warn residents of any anomalies in their daily living activities. In addition, this component can be further used to provide additional information about their lifestyle and health condition to identify potential emergencies, as well as the early stages of disease. Additionally, it is worth mentioning that the proposed method is applied in Tanzania, which is located near the equator. Due to its special geographical location, the external environment has no obvious seasonal changes during a year. However, if this method is applied in other regions, the impact of different seasons on power consumption can also be considered to improve the detection performance. SEPAD is designed based on the programming languages C# and Python, providing a high computation speed for real-time home monitoring. Notably, this is an off-grid power source-based anomaly detector. It is our hope that the demonstrated accuracy and practicality of this detector can be applied widely, particularly in remote areas of developing countries with a large proportion of single-occupancy households and/or poor medical infrastructures.

## ACKNOWLEDGMENT

The authors would like to thank all the members in Innovative Technology and Energy Center, Tanzania, especially Dr. Herb Rhee, for their kind cooperation and help in this project. Also, they would like to express sincere thanks to Jiyoung Rhee of E3 Empower and Institute of Engineering Research at Seoul National University.

## REFERENCES

[1] J. Fletcher and W. Malalasekera, "Development of a user-friendly, low-cost home energy monitoring and recording system," *Energy*, vol. 111, pp. 32–46, Sep. 2016.

[2] C. Li, D. Zhou, and Y. Zheng, "Techno-economic comparative study of grid-connected PV power systems in five climate zones, China," *Energy*, vol. 165, pp. 1352–1369, Dec. 2018.

[3] Y. Iwafune, Y. Mori, T. Kawai, and Y. Yagita, "Energy-saving effect of automatic home energy report utilizing home energy management system data in Japan," *Energy*, vol. 125, pp. 382–392, Apr. 2017.

[4] H.-T. Lee, J.-H. Song, S.-H. Min, H.-S. Lee, K. Y. Song, C. N. Chu, and S.-H. Ahn, "Research trends in sustainable manufacturing: A review and future perspective based on research databases," *Int. J. Precis. Eng. Manuf.-Green Technol.*, vol. 6, no. 4, pp. 809–819, 2019.

[5] (2018). *Electricity Consumption in the United States was About 3.95 Trillion Kilowatthours (KWH)*. Accessed: 2018. [Online]. Available: https://www.eia.gov/energyexplained/index.php?page=electricity_use/

[6] A. Rotimi, A. Bahadori-Jahromi, A. Mylona, P. Godfrey, and D. Cook, "Estimation and validation of energy consumption in UK existing hotel building using dynamic simulation software," *Sustainability*, vol. 9, no. 8, p. 1391, 2017.

[7] A. Spagnuolo, A. Petraglia, C. Vetromile, R. Formosi, and C. Lubritto, "Monitoring and optimization of energy consumption of base transceiver stations," *Energy*, vol. 81, pp. 286–293, Mar. 2015.

[8] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, "Deep structured energy based models for anomaly detection," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1–10.

[9] 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.

[10] E. Mocanu, P. H. Nguyen, M. Gibescu, and L. Wil Kling, "Deep learning for estimating building energy consumption," *Sustain. Energy, Grids Netw.*, vol. 6, pp. 91–99, Jun. 2016.

[11] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mech. Syst. Signal Process.*, vol. 115, pp. 213–237, Jan. 2019.

[12] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE Trans. Power Del.*, vol. 25, no. 2, pp. 1162–1171, Apr. 2010.

[13] S. S. S. R. Depuru, L. Wang, V. Devabhaktuni, and R. C. Green, "High performance computing for detection of electricity theft," *Int. J. Electr. Power Energy Syst.*, vol. 47, pp. 21–30, May 2013.

[14] S. Makonin, F. Popowich, I. V. Bajić, B. Gill, and L. Bartram, "Exploiting HMM sparsity to perform online real-time nonintrusive load monitoring," *IEEE Trans. Smart Grid*, vol. 7, no. 6, pp. 2575–2585, Nov. 2016.

[15] P. Jokar, N. Arianpoo, and V. C. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216–226, 2016.

[16] A. Iwayemi and C. Zhou, "SARAA: Semi-supervised learning for automated residential appliance annotation," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 779–786, Mar. 2017.

[17] E. Bair, "Semi-supervised clustering methods," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 5, no. 5, pp. 349–361, 2013.

[18] M. Wytock and J. Z. Kolter, "Contextually supervised source separation with application to energy disaggregation," in *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 1–7.

[19] E. Elhamifar and S. Sastry, "Energy disaggregation via learning powerlets and sparse coding," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 1–7.

[20] K. Yan, C. Zhong, Z. Ji, and J. Huang, "Semi-supervised learning for early detection and diagnosis of various air handling unit faults," *Energy Buildings*, vol. 181, pp. 75–83, Dec. 2018.

[21] Y. Zhang, W. Chen, and J. Black, "Anomaly detection in premise energy consumption data," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Jul. 2011, pp. 1–8.

[22] J. Chou and A. S. Telaga, "Real-time detection of anomalous power consumption," *Renew. Sustain. Energy Rev.*, vol. 33, pp. 400–411, May 2014.

[23] D. F. M. Cabrera and H. Zareipour, "Data association mining for identifying lighting energy waste patterns in educational institutes," *Energy Buildings*, vol. 62, pp. 210–216, Jul. 2013.

[24] M. Hu, Z. Ji, K. Yan, Y. Guo, X. Feng, J. Gong, X. Zhao, and L. Dong, "Detecting anomalies in time series data via a meta-feature based approach," *IEEE Access*, vol. 6, pp. 27760–27776, 2018.

[25] J. Alcalá, O. Parson, and A. Rogers, "Detecting anomalies in activities of daily living of elderly residents via energy disaggregation and Cox processes," in *Proc. ACM Int. Conf. Embedded Syst. Energy-Efficient Built Environ.*, 2015, pp. 225–234.

[26] S. Rahimi, A. D. C. Chan, and R. A. Goubran, "Nonintrusive load monitoring of electrical devices in health smart homes," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf.*, May 2012, pp. 2313–2316.

[27] M. Hori, T. Harada, and R.-I. Taniguchi, "Anomaly detection for an elderly personwatching system using multiple power consumption models," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2017, pp. 669–675.

[28] S. R. Gaddam, V. V. Phoha, and K. S. Balagani, "K-Means+ID3: A novel method for supervised anomaly detection by cascading K-means clustering and ID3 decision tree learning methods," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 345–354, Mar. 2007.

[29] X.-D. Wang, R.-C. Chen, F. Yan, Z.-Q. Zeng, and C.-Q. Hong, "Fast adaptive K-means subspace clustering for high-dimensional data," *IEEE Access*, vol. 7, pp. 42639–42651, 2019.

[30] S. Welikala, C. Dinesh, M. P. B. Ekanayake, R. I. Godaliyadda, and J. Ekanayake, "Incorporating appliance usage patterns for non-intrusive load monitoring and load forecasting," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 448–461, Jan. 2019.

[31] R. C. de Amorim and C. Hennig, "Recovering the number of clusters in data sets with noise features using feature rescaling factors," *Inf. Sci.*, vol. 324, pp. 126–145, Dec. 2015.

[32] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018.

[33] D. P. Kroese, R. Y. Rubinstein, and T. Taimre, "Application of the cross-entropy method to clustering and vector quantization," *J. Global Optim.*, vol. 37, no. 1, pp. 137–157, 2007.

[34] B. Esmael, A. Arnaout, R. K. Fruhwirth, and G. Thonhauser, "Improving time series classification using hidden Markov models," in *Proc. Int. Conf. Hybrid Intell. Syst.*, Dec. 2012, pp. 502–507.

[35] R. Bonfigli, E. Principi, M. Fagiani, M. Severini, S. Squartini, and F. Piazza, "Non-intrusive load monitoring by using active and reactive power in additive factorial hidden Markov models," *Appl. Energy*, vol. 208, pp. 1590–1607, Dec. 2017.

[36] R. Raman, P. K. Sa, B. Majhi, and S. Bakshi, "Direction estimation for pedestrian monitoring system in smart cities: An HMM based approach," *IEEE Access*, vol. 4, pp. 5788–5808, 2016.

[37] A. Baratloo, M. Hosseini, A. Negida, and G. El Ashal, "Part 1: Simple definition and calculation of accuracy, sensitivity and specificity," *Emergency*, vol. 3, no. 2, pp. 48–49, 2015.

**XINLIN WANG** is currently pursuing the Ph.D. degree with the Department of Mechanical and Aerospace Engineering, Seoul National University (SNU). His current research interests include far-field wireless energy monitoring and energy consumption anomaly detection.

**INSOON YANG** received the B.S. degree *(summa cum laude)* in mathematics and MAE from Seoul National University (SNU), in 2009, the M.S. degree in EECS, the M.A. degree in mathematics, and the Ph.D. degree in EECS from UC Berkeley, in 2012, 2013, and 2015, respectively. He was an Assistant Professor with the ECE Department, USC, from 2016 to 2018, and a Postdoctoral Associate with the Laboratory for Information and Decision Systems, MIT, from 2015 to 2016. He is currently an Assistant Professor with the ECE Department, SNU. His research interests include stochastic control and optimization with application to cyber-physical systems and safe autonomy. He was a recipient of the 2015 Eli Jury Award and a finalist for the Best Student Paper Award at the 55th IEEE Conference on Decision and Control 2016.

**SUNG-HOON AHN** received the bachelor's degree in aerospace engineering from the University of Michigan, in 1992, and the master's and Ph.D. degrees in aeronautics and astronautics from Stanford University, in 1994 and 1997, respectively. Since 1997, he has held professional and visiting positions with Stanford University, the University of California at Berkeley, Gyeongsang National University, and the University of Washington. He joined Seoul National University (SNU), in 2003. He has served for the SNU Institute of Global Social Responsibility as the Director, from 2016 to 2017, the University Industry Technology Force (UNITEF), South Korea, as the President, from 2016 to 2017, and the SNU Graduate School of Engineering Practice as the Associate Dean, from 2017 to 2018, and has been serving for the Innovative Technology Energy Center (ITEC) as the Program Director, since 2017. He is currently a Full Professor with the Department of Mechanical and Aerospace Engineering and Institute of Advanced Machines and Design, SNU, where he is also the Director of the Innovative Design and Integrated Manufacturing Laboratory. His research interests include soft robotics, smart/composite materials, micro/nano fabrications, 3D/4D printing, smart factory, green manufacturing, renewable energy, smart grid, and appropriate technology. Since 2013, as the Editor-in-Chief, he has founded and managed the *International Journal of Precision Engineering and Manufacturing-Green Technology* (IJPEM-GT, 2018 impact factor 4.561).