Received: 1 February 2024 | Revised: 7 May 2024 | Accepted: 17 June 2024

DOI: 10.4218/etrij.2023-0155

ORIGINAL ARTICLE

ETRI Journal WILEY

Anomaly detection and prediction of energy consumption for smart homes using machine learning

Anitha Ambat | **Jayakrushna Sahoo**

Department of Computer Science and Engineering, Indian Institute of Information Technology, Kottayam, India

Correspondence
Jayakrushna Sahoo, Department of Computer Science and Engineering, Indian Institute of Information Technology, Kottayam, Kerala, India.
Email: jsahoo@iiitkottayam.ac.in

Abstract

As technology advances, smart homes are being increasingly adopted, thus generating massive data that pose new research challenges. We propose a machine learning framework for monitoring energy consumption in smart home devices. The proposed framework involves an anomaly detection module, followed by a predictive model to forecast energy consumption patterns in a typical smart home. We employ three outlier-based techniques for anomaly detection: (1) local outlier factor, (2) connectivity-based outlier factor, and (3) cluster-based local outlier factor. Furthermore, we apply random forest, linear regression, decision tree, and the ensemble techniques of adaptive, gradient, and extreme gradient boosting to anomaly free data to develop baseline models that predict the energy consumption patterns of smart home devices. The framework is evaluated on three publicly available energy datasets collected from various smart homes. The experimental results reveal that the cluster-based local outlier factor with extreme gradient boosting achieves promising results with high prediction accuracy.

KEYWORDS

anomaly detection, energy consumption pattern, machine learning, outlier identification, smart home

1 | INTRODUCTION

Over the past decade, the Internet of Things (IoT) [1] has become a technology that has substantially enhanced our lives. The IoT comprises objects, including sensors, electronic appliances, and actuators, endowed with Internet connectivity. Users can remotely control devices over the Internet through software tools. Remarkably, smart homes have become a widespread application of the IoT in recent years. Home devices, such as washing machines, fridges, air conditioners, lights, cameras, ovens, home theaters, and other appliances, can be connected to and communicate with users over the Internet [2]. Therefore, a smart home can be established by implementing automated operations, making it easy and convenient for users to live more comfortably and enhancing energy management [3]. Smart homes also provide a strong defense for ensuring the security and safety of living by automatically monitoring activities such as human activity identification, fall detection for the elderly, smoke, gas, and fire detection, and invasion and home surveillance [4]. While smart home appliances improve user safety, security, and quality of life, these devices also generate enormous amounts of data from operations and user–device communications. These data include audio, video, and user activity logs, which must be collected and stored in the background for security audits and analyses and other tasks.

However, IoT devices include smart home devices, face safety, and security threats [5]. Users face heightened privacy risks owing to unauthorized access, device hijacking, tampered protocols, and other security violations [6]. The occurrence of anomalies owing to the wasteful use of devices exacerbates these risks [7]. Nevertheless, because smart home data reflect all the activities, including usage data and status of the devices inside a smart home, analysis with efficient algorithms can be applied to identify anomalies. For instance, anomalous energy consumption concerns smart home energy management and is usually triggered by device faults, false data injection attacks [8], electricity stealing, and user inattention [9]. Energy consumption depends on the total electric power, usage patterns of the occupants, average indoor/outdoor temperature, and number of occupants in a smart home. Thus, by analyzing energy consumption data of smart homes, anomalies in devices can be identified and further necessary steps can be implemented to mitigate future anomalies. In addition, a reliable model for predicting daily energy consumption patterns in related smart home data can contribute to the proper use of devices in smart homes.

Anomaly detection involves the identification of observations that considerably deviate from the expected patterns in collected historic data. Anomaly detection is typically performed using energy consumption data at two levels. The first level consists of analyzing the total energy consumption of a smart home, and the second level involves precisely classifying and identifying modules or subsystems of a device that exhibit anomalous behaviors [10]. Subsequently, the estimation of the anomalous energy consumption can help reduce the overall energy and operation costs, and it can enhance energy efficiency.

Anomalies in energy consumption data of smart homes are usually identified using statistical and classification approaches [11]. Machine learning techniques have also been employed to detect anomalous energy consumption in related data [12]. To adequately manage energy, monitoring can be performed using early anomaly detection in energy consumption data. However, no reliable models are available for predicting energy consumption in smart homes incorporating early anomaly detection, and existing anomaly detection from energy consumption data can be further improved. Moreover, an efficient prediction model for energy consumption in smart homes can be developed. To devise an accurate prediction model, the detection and removal of anomalies from a dataset are essential. To this end, we combine anomaly detection and prediction frameworks by applying different machine learning techniques to energy consumption patterns from related data. Our proposal can contribute to enhancing security, overuse mitigation, faulty device detection, and energy waste reduction in smart homes.

We developed a model for smart home anomaly detection using energy consumption data that include anomalies resulting from device malfunctions or inefficient use. By employing three outlier detection methods, namely, local, connectivity, and cluster-based outlier detection (LOF, COF, and CBLOF, respectively), we accurately detected anomalies. The identified anomalous datapoints were removed. In addition, we devised a predictive model from anomaly free data by employing different machine learning techniques, namely, linear regression (LR), random forest (RF), decision tree (DT), support vector regression (SVR), *k*-nearest neighbors (KNN), adaptive boosting (AdaBoost), gradient boosting, and extreme gradient boosting (XGboost). These techniques were used to forecast energy consumption patterns in smart homes.

The contributions of this study are summarized as follows:

- We detect anomalies in the daily energy consumption patterns from smart home data based on three outlier detection techniques (that is, LOF, COF, and CBLOF).
- We propose a framework comprising different machine learning techniques for the prediction of daily energy consumption patterns in smart homes.
- We evaluate the efficiency of the proposed framework based on the coefficient of determination (R^2) and root mean square error (RMSE) when applying the framework to three representative energy datasets collected from smart homes.

The remainder of this paper is structured as follows. Section 2 outlines related work. The methods adopted for anomaly detection of smart home energy consumption data are described in Section 3. Section 4 describes the workflow of this study, and Section 5 presents the experimental evaluation and results. Finally, Section 6 presents conclusions including a direction of future work.

2 | RELATED WORK

Various studies have been conducted on approaches for recognizing daily energy consumption patterns in smart homes and detecting anomalies in such data. Diverse machine learning techniques have been applied to differentiate deviated patterns from typical events. Liu and others [13] introduced a data mining framework that extracts regular electrical load patterns and detects anomalies using the DBSCAN and *k*-means algorithms. Furthermore, the CART algorithm was used to determine the association between electricity load patterns and their influencing factors. Lazim Qaddoori and Ali [14] proposed a lightweight framework for smart meters. It leverages machine learning to identify irregularities in energy consumption data for individual buildings.

Wang and Ahn [15] introduced a framework for home electrical load anomaly detection that combined a hybrid one-step-ahead load predictor with a rule-engine-based load anomaly detector to improve the load prediction accuracy. They used the KNN algorithm and a support vector machine to enhance anomaly detection. Xu and Chen [16] presented a system for tracking irregular building energy usage and performing anomaly detection using hybrid data mining. They used a recurrent neural network to find the incorrect prediction interval, and unusual building energy consumption outcomes were evaluated using the quantile regression range. This method was applied to analyze energy consumption data from three different residences. An anomaly detection method that uses a federated learning approach to train a long short-term memory model and concurrently solve multiple tasks was investigated by Sater and Hamza [17]. In [18], a spectral dual convolutional neural network detected anomalies in a time-series data stream by setting a threshold to determine anomalous values from the entire energy consumption data.

Zainab and others [19] detected spam in smart home device readings from time-series data using machine learning. Their method estimates a spamicity score of each IoT device and determines the dependability of devices in the home network based on the feature importance and RMSE of machine learning techniques.

Bouabdallaoui and others [20] proposed a machine learning technique for the predictive maintenance of buildings. Data were collected from different sensors in a building, and a model was developed using an autoencoder, recurrent neural network, and long short-term memory model for fault detection. A real-world case study was considered to predict the maintenance of heating, ventilation, and air conditioning installations in sports buildings. Gaur and others [21] proposed a method to generate a ground truth based on available data for detecting anomalies in short- and long-range energy consumption using the z-score and LR models.

3 | METHODS

This section presents the machine learning techniques used to detect anomalies and predict energy consumption in smart homes. To detect anomalies in the energy consumption data, we used the LOF, COF, and CBLOF methods. Outliers in a dataset likely reflect inconsistencies. An outlier is a datapoint that substantially differs from the rest of available observations [22] and may indicate an anomaly. An irregular, uncommon, infrequent, or flawed observation depending on the context can be considered as an outlier [23].

3.1 | LOF

LOF [24] compares the density of a point with the densities of its neighbors and detects an outlier based on the local neighborhood outlier factor. A high score indicates that the evaluated point is an outlier.

For each datapoint a, let $d^k(a)$ represent the distance between datapoint a from its k-th neighbor and $N_k(a)$ represent a set of datapoints within distance $d^k(a)$. The reachability distance for each datapoint a is calculated as

$$rd_k(a,b) = \max(\mathrm{dist}(a,b), d^k(b)). \tag{1}$$

We then compute the local reachability distance, $lrd_k(a)$, of datapoint a as follows:

$$lrd_k(a) = \left[\frac{\mathrm{MEAN}_{b\in N_k(a)} rd_k(a,b)}{N_k(a)}\right]^{-1}. \tag{2}$$

Finally, the LOF score for each point a is calculated as

$$\mathrm{LOF}_k(a) = \frac{\sum_{b\in N_k(a)} lrd_k(a)}{lrd_k(b)}. \tag{3}$$

3.2 | COF

COF [25] compares the likelihood of an observation being an outlier compared with its neighbors. When the COF value of a datapoint is high, the datapoint is considered as an outlier.

The calculation of COF for a datapoint proceeds as follows. For each datapoint a, let $d^k(a)$ represent the distance of point a from its k-th neighbor and $N_k(a)$ represent a set of points within distance $d^k(a)$. We compute set-based nearest path $P=\{a_1, a_2, \ldots, a_k\}$ and find set-based nearest trail $E=\{e_1, e_2, \ldots, e_k\}$. The average chaining distance from a_1 to its KNN $N_k(a)$ is calculated as

$$\mathrm{ACD}-\mathrm{dist}_{N_{(k)}(a)}(a) = \sum_{i=1}^{k} \frac{2(k+1-i)}{k(k+1)} \mathrm{dist}(E_i). \tag{4}$$

The COF score of a with respect to its KNN $N_k(a)$ is then calculated as

$$\mathrm{COF}(a) = \frac{K(\mathrm{ACD} - \mathrm{dist}_{N_{(k)}(a)}(a))}{\sum\limits_{(O \in N_k(a)} \mathrm{ACD} - \mathrm{dist}_{N_{(k)}(O)}(O)}. \tag{5}$$

3.3 | CBLOF

In CBLOF [26], the clustering algorithm divides dataset D into multiple clusters $C = \{C_1, C_2, \ldots, C_k\}$. For each datapoint, the CBLOF score is computed according to the cluster size (large LC or small SC cluster), and the distance between the clusters closest to the datapoint as follows:

$$\mathrm{CBLOF}(a) = \begin{cases} (|C_i| * (\min(\mathrm{dist}(a, C_j))) \\ \quad \text{if } C_j \in LC, a \in C_i, C_i \in SC, \\ |C_i| * \mathrm{dist}(a, C_i) \\ \quad \text{if } a \in C_i, C_i \in LC. \end{cases} \tag{6}$$

3.4 | Machine learning techniques

In LR [27], the equation describing the relation between the dependent and independent variables is a straight line. Multiple LR models employ a set of linear functions to make predictions, thus extending conventional LR. RF is an ensemble method [28] associated with numerous DTs and combines them to obtain a more accurate and stable prediction. Let x be the observed input vectors with various features of the available samples and build the average of q regression trees as follows:

$$Y(x) = \frac{1}{q} \sum_{q=1}^{q} T(x). \tag{7}$$

DT regression is a tree-based structure used to predict the numerical outcomes of a dependent variable. SVR is a supervised learning algorithm used to build regression models. Its performance depends on the selection of an appropriate kernel and relevant parameters [29]. KNN is a nonparametric regression algorithm that demonstrates strong predictive capabilities by leveraging the proximity of datapoints in the feature space. It computes the average of the target values of k adjacent neighbors. The average distance between two n-feature occurrences represents their similarity. For $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$, the Euclidean distance is calculated as [30]

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{8}$$

Ensemble methods combine various learners to obtain a stronger one. AdaBoost assigns initial weights to samples, trains base learners using weighted data, adjusts the sample weights based on the learner performance, iteratively trains new learners, and finally combines them to create a strong learner. Gradient boosting, which often uses regression trees, is effective for both regression and classification, iteratively producing robust and interpretable models according to the functional gradient descent. It sequentially combines weak learners, fitting each learner against the residual error of the current ensemble and scaling the learning rate before adding it to the ensemble. XGBoost is a powerful ensemble learning algorithm that excels in handling complex datasets and capturing intricate relations between features. Its ability to sequentially optimize DTs coupled with regularization techniques allows XGBoost to mitigate overfitting and achieve superior predictive performance [31]. Let x and y be the observation and target, respectively. We have

$$F_0(x) = \mathrm{argmin}_{\lambda} \sum_{i=1}^{N} L(y_i, \lambda), \tag{9}$$

where $\mathrm{argmin}_{\lambda}$ provides the minimized sum, that is, the average across y values:

$$\tilde{y}_i = -\left[\frac{\partial L(y_{i^i} F(x_i))}{\partial F(x_i)}\right] F(x) = F_{m-1}(x), \tag{10}$$

with m ranging from 1 to M to build M trees. Using the actual and predicted values, we calculate the following loss function:

$$F_m(x) = F_{m-1}(x) + \lambda_m h(x;\, a_m), \tag{11}$$

where λ_m is the learning rate and $h(x; a_m)$ represents the base learner.

4 | PROPOSED FRAMEWORK

The proposed framework for detecting anomalies and predicting the energy consumption of smart homes mainly relies on machine learning techniques. Figure 1 illustrates the intermediate processes in the proposed framework, which involves a two-stage anomaly detection module and prediction module. The anomaly detection module identifies anomalous energy consumption from the corresponding smart home data using three outlier detection algorithms. Subsequently, from anomaly free data, a model that predicts the energy consumption

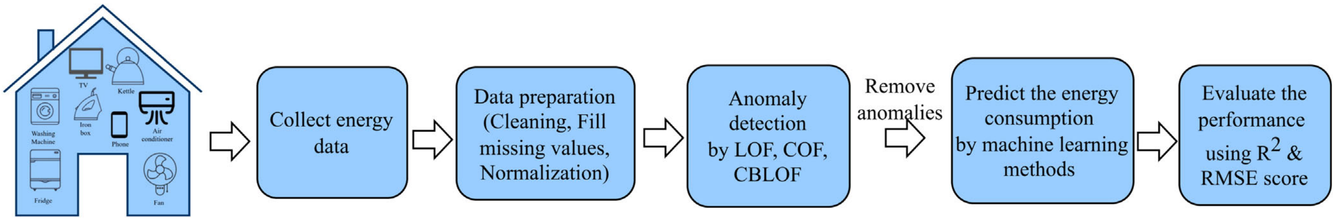

TV
Kettle
Washing Machine
Iron box
Phone
Air conditioner
Fridge
Fan
Collect energy data
Data preparation (Cleaning, Fill missing values, Normalization)
Anomaly detection by LOF, COF, CBLOF
Remove anomalies
Predict the energy consumption by machine learning methods
Evaluate the performance using R^2 & RMSE score

FIGURE 1 Overall diagram of proposed framework.

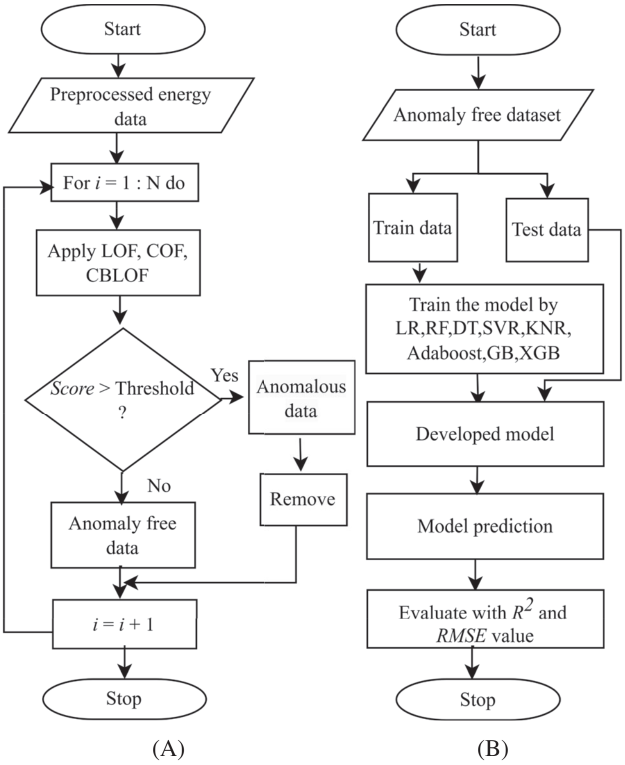


FIGURE 2 Flowchart of the proposed framework for (A) anomaly detection and (B) prediction of energy consumption in smart home.

patterns of smart home occupants is constructed using various machine learning techniques. The flowchart of the framework is shown in Figure 2.

4.1 | Data preparation

We used three publicly available energy consumption datasets to conduct performance evaluations in terms of two accuracy metrics.

Energy consumption data were gathered from three publicly available datasets, namely, REFIT (https://repository.lboro.ac.uk/articles/dataset/REFIT_Smart_Home_dataset/2070091, Hue https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/N3HGRN), and DRED https://www.st.ewi.tudelft.nl/~akshay/dred/. The REFIT dataset includes smart home energy consumption data from 20 households in the UK collected from 2013 to 2015. It encompasses the energy consumption data of devices including motion sensors, smart meters, door sensors, programmable thermostats, window sensors, and programmable radiator valves installed in 389 rooms. The complete energy consumption data of each smart home were recorded with timestamps at 30-min intervals and comprised 20767 tuples. The Hue dataset details the hourly energy consumption collected from 28 houses in different residential buildings in British Columbia over 3 years. It contains 26304 instances of date, time, hour, and energy consumption measured in kilowatt hours. DRED (Dutch residential energy dataset) includes the aggregated and appliance-level energy consumption of a single household in the Netherlands over 2 weeks from July 5 to 17, 2015 at 1-s intervals. We downsampled the data to 1-min intervals to obtain 17477 rows containing the time and energy consumption in watts.

4.2 | Data preprocessing

The datasets with a few anomalous values, such as missing, null, incomplete, and noisy entries, were substituted using the following procedure. The mean values of the energy consumption datasets were used to fill all missing datapoints from the same hour for a given weekday within a specific month. Alternatively, the values of same day from the previous year were used. Furthermore, the energy consumption data were normalized to the maximum energy consumption in the smart home as follows:

$$N(E(t)) = \frac{E(t)}{\max(E(t))}, \tag{12}$$

where $N(E(t))$ is the normalized energy consumption data, $E(t)$ is the actual energy consumption at different times $t = 1, 2, \ldots, 24$, and $\max(E(t))$ is the maximum daily energy consumption.

4.3 | Anomaly detection

Algorithm 1 describes the procedure to find anomalous datapoints.

Algorithm 1 Anomaly detection algorithm

```
$M$ ← LOF/COF/CBLOF
Threshold $T$
Input: Energy consumption dataset $E$
Output: Anomaly free dataset $Y$
for $i = 0$:N do
    Apply each model ($M$) and calculate $score(M)$ using 3, 5, and 6
    if score ($M$) > $T$ then
        $M_a$ ← Anomaly
    else
        $M_n$ ←Normal
    end if
end for
return $Y$
```

Let E represent the energy consumption dataset and M represent the model used for the LOF, COF, and CBLOF methods. We applied these models separately and calculated $score(M)$ per method. If the score exceeded threshold T, the data were labeled as anomalous. The LOF method was used to analyze the energy data and calculate the anomalous score per datapoint. The scores were compared with the maximum threshold, and any datapoint with a score above the threshold was considered anomalous. The same procedure was used for the COF and CBLOF methods.

4.4 | Prediction of energy consumption patterns

Algorithm 2 Prediction algorithm

```
$P$ represents LR, RF, DT, SVR, KNN, AdaBoost, gradient boost, XGBoost
Input: Anomaly free data of each model, $Y$
Output: RMSE and $R^2$
Define datasets train($Y$) and test($Y$)
Initialize $P$
Training: $M_P$ ← model($P$, train($Y$))
if Training complete then
    $S_P$ ← predict($M_P$, test($Y$))
end if
Calculate RMSE and $R^2$
if RMSE is low and $R^2$ is high then
    fitting model
else
    unsuitably fitting model
end if
```

Various algorithms have been used to identify and predict energy consumption in smart homes. We applied LR, RF, DT, SVR, KNN, AdaBoost, gradient boosting, and XGBoost to efficiently predict energy consumption patterns. We applied Algorithm 2 to anomaly free data derived from each of the abovementioned algorithms to identify and predict energy consumption patterns. In detail, 80% of the anomaly free data were used for model training, and the remaining 20% were used for testing. Each model was trained on a training set and tested for prediction.

We used two accuracy metrics, namely, RMSE and R^2, to evaluate the goodness of fit according to (13) and (14). RMSE is the square root of the average square of the difference between the actual and predicted values. A lower RMSE indicates a higher goodness of fit of the model. R^2 is the coefficient of determination with a value between 0 and 1. When R^2 is 1, the model perfectly fits the data. If it is above 0.5, the model is reasonably good, while lower values indicate an unreliable model.

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{m}(x_i - y_i)^2}, \tag{13}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(x_i - y_i)^2}{\sum_{i=1}^{m}(x_i - \overline{y_i})^2}, \tag{14}$$

where x_i is the actual value, y_i is the predicted value, $\overline{y_i}$ is the mean across actual values, and M is the number of datapoints.

5 | RESULTS AND DISCUSSION

We used Google Colab as the primary computing platform in this study. We predominantly used Python as the programming language because of its versatility and rich ecosystem of libraries suitable for data analysis, machine learning, and statistical modeling. Within Google Colab notebooks, we used libraries such as NumPy, Pandas, scikit-learn, and PyOD to implement the various components of our framework and evaluation experiment. The data were cleaned and prepared as described in Section 4, and experiments to detect anomalies were conducted at three contamination levels: 0.01, 0.05, and 0.1. Furthermore, anomaly free data were extracted for prediction. By applying the different algorithms mentioned in Section 3, the models were fitted to the extracted data, and the goodness of fit was evaluated. We present and discuss the experimental results in this section.

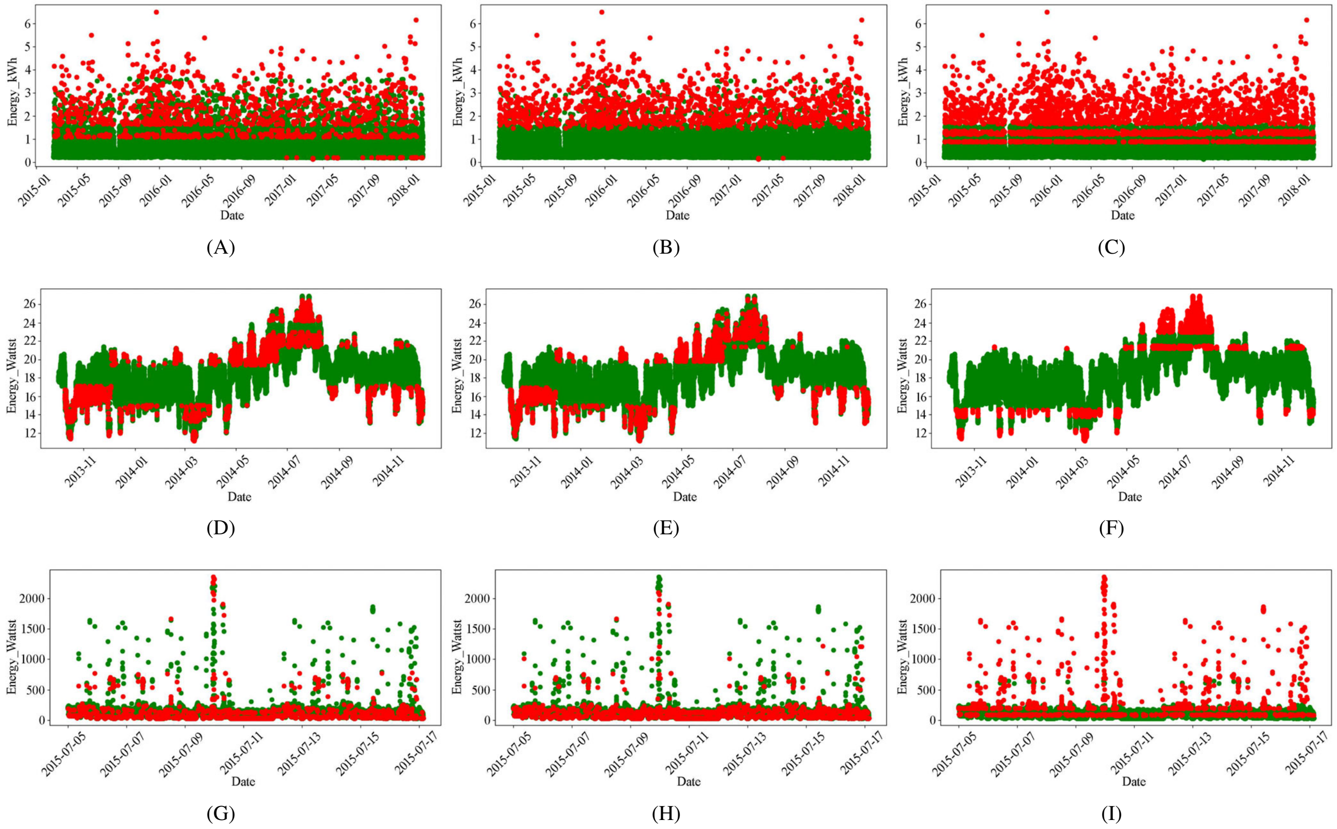
5.1 | Anomaly detection results

Table 1 lists the number of anomalous datapoints detected using each method with different parameter settings. The LOF method with contamination level 0.05 and 50 neighbors identified 1137 anomalies at threshold $T = 1.0$ on the Hue dataset. In contrast, on the REFIT dataset and DRED, the numbers of detected anomalous datapoints were 2076 and 1748, respectively, for the same parameter settings. Similarly, the COF method captured 2056 anomalies at a threshold of 0.0 on the Hue dataset, 2071 anomalies on the REFIT dataset at $T = 1.24$, and 1748 anomalies on DRED at $T = 1.36$. The CBLOF method provided 2605 anomalies at a threshold of 0.14, 2070 anomalies at 0.75, and 1748 anomalies at 25.81 for the three datasets. The highest number of anomalies at

TABLE 1 Anomaly detection using three outlier methods with different parameter settings on Hue dataset, REFIT dataset, and DRED.

Dataset		Hue		REFIT		DRED	
Method	Parameters	Anomaly	T	Anomaly	T	Anomaly	T
	$c = 0.01$, $nn = 50$	233	2.00	206	1.11	174	2.35
	$c = 0.05$, $nn = 50$	1137	1.0	1023	4.66	873	1.37
LOF	$c = 0.1$, $nn = 20$	682	1.0	2076	1.04	1748	1.21
	$c = 0.01$, $nn = 20$, method = fast	242	1.25	205	2.22	175	2.75
	$c = 0.05$, $nn = 20$, method = fast	1304	0.86	1039	1.53	873	1.58
COF	$c = 0.1$, $nn = 20$, method = fast	2056	0.0	2071	1.24	1748	1.36
	$c = 0.01$, $nc = 8$	264	1.90	208	2.54	175	170.15
	$c = 0.05$, $nc = 8$	1311	0.34	1020	1.05	874	32.44
CBLOF	$c = 0.1$, $nc = 8$	2605	0.14	2070	0.75	1748	25.81

Abbreviations: c, contamination level; CBLOF, cluster-based outlier detection; COF, connectivity-based outlier detection; LOF, local-based outlier detection; nc, number of clusters; nn, number of neighbors; RMSE, root mean square error; T, threshold.



(A) (B) (C)

(D) (E) (F)

(G) (H) (I)

FIGURE 3 Anomaly detection results using (A) LOF, (B) COF, and (C) CBLOF on Hue dataset, (D) LOF, (E) COF, and (F) CBLOF on REFIT dataset, and (G) LOF, (H) COF, and (I) CBLOF on DRED.

various contamination levels was selected per method. Subsequently, the detected anomalous datapoints were removed from the datasets, and the resulting anomaly free data were used for prediction.

Figure 3A–C shows the anomaly detection results on the Hue dataset, where the *x*-axis represents the date and time, and the *y*-axis represents the energy consumption in kilowatt hours. We used different parameter values to generate the anomaly scores, as listed in Table 1.

Figure 3D–F shows the anomaly detection results on the REFIT dataset, and Figure 3G–I shows the results on DRED for the parameter settings listed in Table 1. The CBLOF method provided high performance owing to its ability to accurately identify outliers within data clusters, thus enhancing data quality and decision-making. Sophisticated clustering techniques and local outlier factor analysis ensure unparalleled accuracy, making this a cost-effective and reliable solution for anomaly detection across various datasets.

TABLE 2 Parameter settings of machine learning techniques.

Method	LR	RF	DT	SVR	KNN	AdaBoost	Gradient boosting	XGBoost
LOF	fi = True	ne = 900, rs = 0 md = 10, mss = 30	md = 10 mid = 0.10 mss = 50	k = rbf	nn = 120	ne = 1 lr =1	ne = 800 md = 2 lr = 1.0	cb = 0.3, lr = 0.1 md = 10, a = 10 ne = 200
COF	fi = True	ne = 900, rs = 0 md = 10, mss = 30	md = 7 mss = 100	k = rbf	nn = 75	ne = 1 lr = 1	ne = 50 md = 2 lr = 1.0	cb = 0.3, lr = 0.1 md = 10, a = 10 ne = 200
CBLOF	fi = True	ne = 900, rs = 0 md = 10, mss = 30	md = 10 mid = 0.10 mss = 50	k = rbf	nn = 120	ne = 1 lr = 1	ne = 800 md = 2 lr = 1.0	cb = 0.3, lr = 0.1 md = 10, a = 10 ne = 200

Abbreviations: a, alpha; cb, column samples by tree; DT, decision tree; fi, fitting intercept; k, kernel; KNN, *k*-nearest neighbors; lr, learning rate; LR, linear regression; md, maximum depth; mid, minimum impurity decrease; mssm, minimum sample split; ne, number of estimators; nn, number of neighbors; RF, random forest; rs, random state; SVR, support vector regression.

TABLE 3 R^2 and RMSE of prediction models on Hue dataset.

Method	Metric	LR	RF	DT	SVR	KNN	AdaBoost	Gradient boosting	XGBoost
LOF	R^2	0.3931	0.4611	0.3104	0.4131	0.4541	0.4218	0.4551	**0.4809**
	RMSE	0.4429	0.4173	0.4721	0.4355	0.4200	0.4323	0.4196	**0.4096**
COF	R^2	0.3049	0.3782	0.3682	0.3203	0.3828	0.3414	0.3782	**0.3845**
	RMSE	0.47	0.44	0.451	0.4687	0.446	0.4614	0.4483	**0.4460**
CBLOF	R^2	0.6402	0.6937	0.5788	0.6646	0.6813	0.6642	0.6934	**0.6980**
	RMSE	0.3410	0.3146	0.3689		0.3209	0.329	0.3148	**0.3124**

Note: Values bold emphasis only present the acceptable outperformed values of the corresponding metric and methods.
Abbreviations: CBLOF, cluster-based outlier detection; COF, connectivity-based outlier detection; DT, decision tree; KNN, *k*-nearest neighbors; LOF, local-based outlier detection; LR, linear regression; RF, random forest; RMSE, root mean square error; SVR, support vector regression.

TABLE 4 R^2 and RMSE of prediction models on REFIT dataset.

Method	Metric	LR	RF	DT	SVR	KNN	AdaBoost	Gradient boosting	XGBoost
LOF	R^2	0.1722	0.886	0.5499	0.4735	0.9854	0.6205	0.9871	**0.9901**
	RMSE	2.339	0.8643	1.724	1.865	0.3099	1.583	0.2910	**0.255**
COF	R^2	0.1724	0.9398	0.5499	0.472	0.9850	0.617	0.9879	**0.990**
	RMSE	2.338	0.6304	1.724	1.867	0.3145	1.589	0.2827	**0.2558**
CBLOF	R^2	0.242	0.8741	0.600	0.547	**0.9925**	0.626	0.9899	0.9899
	RMSE	2.237	0.9122	1.624	1.728	**0.222**	1.570	0.2572	0.2575

Note: Values bold emphasis only present the acceptable outperformed values of the corresponding metric and methods.
Abbreviations: CBLOF, cluster-based outlier detection; COF, connectivity-based outlier detection; DT, decision tree; KNN, *k*-nearest neighbors; LOF, local-based outlier detection; LR, linear regression; RF, random forest; RMSE, root mean square error; SVR, support vector regression.

5.2 | Prediction results

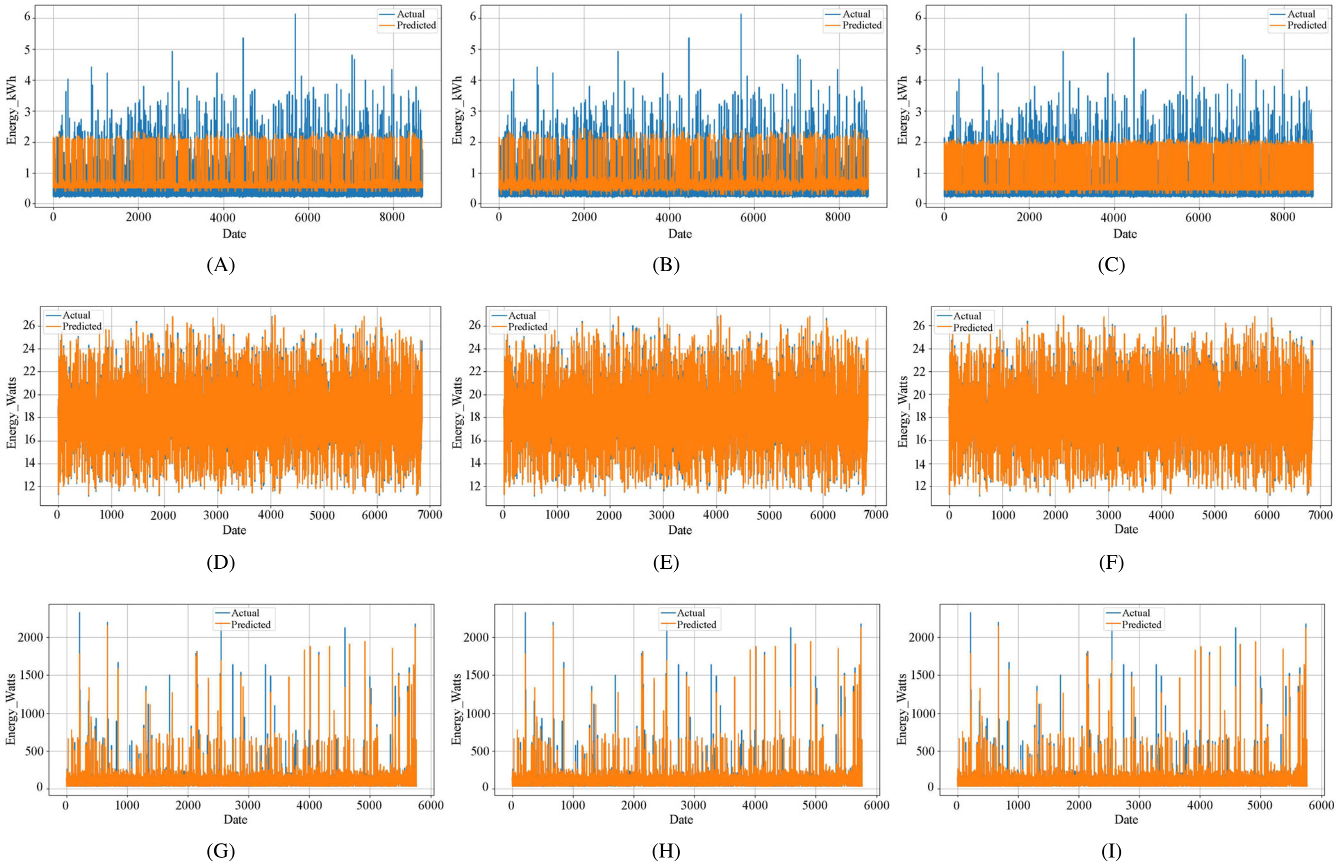
We also evaluated the prediction performance of the proposed framework using the parameters listed in Table 2. On the Hue dataset, XGBoost with CBLOF performed better than the other two methods, LOF and COF, given its R^2 of 0.6980 and RMSE of 0.3124, as listed in Table 3. Accordingly, we can conclude that XGBoost with CBLOF achieves the best energy prediction performance on the Hue dataset. We selected the minimized loss function as RMSE for XGBoost.

Table 4 shows that employing XGBoost on the anomaly free data produced by the LOF and COF methods yields R^2 values of 0.9901 and 0.990, respectively, and RMSE values of 0.255 and 0.2558, respectively, on the REFIT dataset. Hence, XGBoost with LOF or COF is suitable for predicting energy consumption on the REFIT dataset. In contrast, with CBLOF, KNN performed much

TABLE 5 R^2 and RMSE of prediction models on DRED.

Method	Metric	LR	RF	DT	SVR	KNN	AdaBoost	Gradient boosting	XGBoost
LOF	R^2	0.8353	0.894065	0.8667	0.2746	0.5307	0.4777	0.8721	**0.89563**
	RMSE	65.52	52.562	58.957	137.54	110.62	116.71	57.73	**52.170**
COF	R^2	0.823	0.894061	0.8111	0.2781	0.6409	0.4778	0.8926	**0.89562**
	RMSE	67.94	52.563	70.17	137.21	96.76	116.700	52.90	**52.175**
CBLOF	R^2	0.8627	0.897	0.8561	0.2853	0.5510	0.4624	0.8803	**0.8987**
	RMSE	59.8346	51.81	61.25	136.52	108.20	118.40	55.861	**51.39**

Note: Values bold emphasis only present the acceptable outperformed values of the corresponding metric and methods.
Abbreviations: CBLOF, cluster-based outlier detection; COF, connectivity-based outlier detection; DT, decision tree; KNN, *k*-nearest neighbors; LOF, local-based outlier detection; LR, linear regression; RF, random forest; RMSE, root mean square error; SVR, support vector regression.



(A) (B) (C)

(D) (E) (F)

(G) (H) (I)

FIGURE 4 Actual and predicted energy consumption for (A) LOF, (B) COF, and (C) CBLOF with XGBoost on Hue dataset, (D) LOF and (E) COF with XGBoost and (F) CBLOF with KNN on REFIT dataset, and (G) LOF, (H) COF, and (I) CBLOF with XGBoost on DRED.

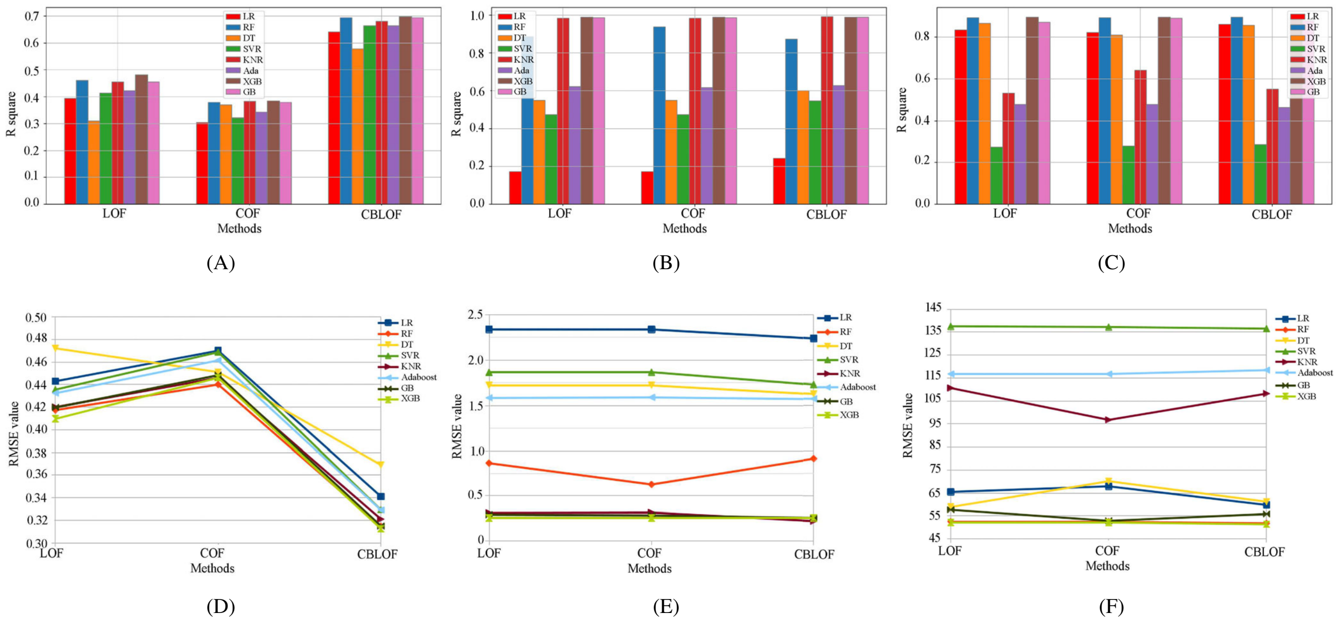


FIGURE 5 Performance of prediction methods in terms of R^2 (top row) and RMSE (bottom row) on (A,D) Hue dataset, (B,E) REFIT dataset, and (C,F) DRED.

better, which is evident from its R^2 and RMSE of 0.9925 and 0.222, respectively. On DRED, CBLOF with XGBoost achieved R^2 and RMSE of 0.8987 and 51.39, respectively, as listed in Table 5.

For the Hue dataset, the actual and predicted energy consumption values obtained using XGBoost are shown in Figure 4A–C, with the method achieving suitable fitting. However, on the REFIT dataset, XGBoost with LOF and COF and KNN with CBLOF outperformed the other methods in predicting energy consumption, as shown in Figure 4D–F. By estimating the output based on the neighbor average, KNN adapted well to various data distributions and was robust to outliers. On DRED, CBLOF with XGBoost fit well, as shown in Figure 4G–I. These findings indicate that the adopted techniques allow to identify daily patterns and detect anomalies in energy consumption data from smart home appliances.

The performance of the evaluated methods on the three datasets is shown in Figure 5. XGBoost obtained better results than the other methods on the test sets. Its scalability and efficiency render it suitable for large-scale datasets and real-world applications. The trained models can predict the energy consumption and secure smart homes with high confidence.

6 | CONCLUSION

Smart homes have recently emerged as a beneficial solution to deliver comfortable living conditions. The security and cost-effectiveness of smart homes must be ensured by continuously monitoring and optimizing the energy consumption of their devices. We propose an anomaly detection method for smart home energy consumption data. The proposed framework involves a prediction model that employs machine learning techniques to predict the energy consumption of a smart home and detect any deviation as anomalous usage. Thus, the framework helps users take the necessary countermeasures by detecting energy consumption anomalies. The framework was tested on three representative smart home energy consumption datasets and achieved promising results. The experimental results show that XGBoost and KNN consistently outperform the other techniques across the R^2 and RMSE metrics. The high performance is due to their robustness to outliers and ability to capture complex patterns from diverse datasets. A limitation of the proposed framework is that we focused on the total energy consumption without considering the consumption of each device. In future work, we will use device-level energy consumption data to unveil additional aspects of anomalies in smart homes.

AUTHOR CONTRIBUTIONS

Anitha Ambat and Jayakrushna Sahoo designed the framework and analyzed the data. Anitha Ambat processed the experimental data, performed the analyses, and drafted the manuscript. Both authors contributed to the final version of this manuscript. Jayakrushna Sahoo supervised the project.

ACKNOWLEDGMENTS

We thank the anonymous reviewers, editor, and editor-in-chief for providing thoughtful and constructive criticism, which greatly enhanced the quality, content, and presentation of this paper.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Jayakrushna Sahoo https://orcid.org/0000-0002-4514-3916

REFERENCES

1. A. Paul and R. Jeyaraj, *Internet of things: A primer*, Human Behavior Emerg. Technol. **1** (2019), no. 1, 37–47.
2. E. S. Lee, H. J. Lee, K. Lee, and J. H. Park, *Automating configuration system and protocol for next-generation home appliances*, ETRI J. **35** (2013), no. 6, 1094–1104.
3. R. Liu and Y. Ge, *Smart home system design based on internet of things*, (12th International Conference on Computer Science and Education, Houston, TX, USA), 2017, pp. 444–448.
4. H. Yar, A. S. Imran, Z. A. Khan, M. Sajjad, and Z. Kastrati, *Towards smart home automation using IoT-enabled edge-computing paradigm*, Sens. **21** (2021), no. 14, 4932.
5. J. Dogani, M. Farahmand, and H. Daryanavard, *A new method to detect attacks on the internet of things (IoT) using adaptive learning based on cellular learning automata*, ETRI J. **44** (2022), no. 1, 155–167.
6. P. Khanpara, K. Lavingia, R. Trivedi, S. Tanwar, A. Verma, and R. Sharma, *A context-aware internet of things-driven security scheme for smart homes*, Sec. Privacy **6** (2023), no. 1, e269.
7. H. Rashid and P. Singh, *Monitor: an abnormality detection approach in buildings energy consumption*, (IEEE 4th international conference on collaboration and internet computing, Philadelphia, PA, USA), 2018, pp. 16–25.
8. B. K. Sethi, D. Mukherjee, D. Singh, R. K. Misra, and S. R. Mohanty, *Smart home energy management system under false data injection attack*, Int. Trans. Electr. Energy Syst. **30** (2020), no. 7, e12411.
9. Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, and A. Amira, *Artificial intelligence based anomaly detection of energy consumption in buildings: a review, current trends and new perspectives*, Appl. Energy **287** (2021), 116601.
10. H. Wang, P. Xu, X. Lu, and D. Yuan, *Methodology of comprehensive building energy performance diagnosis for large commercial buildings at multiple levels*, Appl. Energy **169** (2016), 14–27.
11. M. Han, F. Johari, P. Huang, and X. Zhang, *Generating hourly electricity demand data for large-scale single-family buildings by a decomposition-recombination method*, Energy Built Environ. **4** (2023), no. 4, 418–431.
12. A. Malki, E. S. Atlam, and I. Gad, *Machine learning approach of detecting anomalies and forecasting time-series of IoT devices*, Alexandria Eng. J. **61** (2022), no. 11, 8973–8986.
13. 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 Build. **231** (2021), 110601.
14. S. L. Qaddoori and Q. I. Ali, *An embedded and intelligent anomaly power consumption detection system based on smart metering*, IET Wireless Sens. Syst. **13** (2023), no. 2, 75–90.
15. X. Wang and S. H. Ahn, *Real-time prediction and anomaly detection of electrical load in a residential community*, Appl. Energy **259** (2020), 114145.
16. C. Xu and H. Chen, *A hybrid data mining approach for anomaly detection and evaluation in residential buildings energy data*, Energy Build. **215** (2020), 109864.
17. R. A. Sater and A. B. Hamza, *A federated learning approach to anomaly detection in smart buildings*, ACM Trans. Internet Things **2** (2021), no. 4, 1–23.
18. S. V. Oprea, A. Bâra, F. C. Puican, and I. C. Radu, *Anomaly detection with machine learning algorithms and big data in electricity consumption*, Sustainability **13** (2021), no. 19, 10963.
19. A. Zainab, S. S. Refaat, and O. Bouhali, *Ensemble-based spam detection in smart home IoT devices time series data using machine learning techniques*, Informa. **11** (2020), no. 7, 344.
20. Y. Bouabdallaoui, Z. Lafhaj, P. Yim, L. Ducoulombier, and B. Bennadji, *Predictive maintenance in building facilities: a machine learning-based approach*, Sens. **21** (2021), no. 4, 1044.
21. M. Gaur, S. Makonin, I. V. Bajić, and A. Majumdar, *Performance evaluation of techniques for identifying abnormal energy consumption in buildings*, IEEE Access **7** (2019), 62721–62733.
22. V. Barnett and T. Lewis, *Outliers in statistical data*, 3rd ed., Wiley, New York, 1994.
23. 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 **109** (2021), no. 5, 756–795.
24. M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, *LOF: identifying density-based local outliers*, (Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Dalles, TX, USA), 2000, pp. 93–104.
25. J. Tang, Z. Chen, A. W. C. Fu, and D. W. Cheung, *Enhancing effectiveness of outlier detections for low density patterns*, (Advances in Knowledge Discovery and Data Mining: 6th Pacific-Asia Conference Taipei Taiwan), 2002, pp. 535–548.
26. Z. He, X. Xu, and S. Deng, *Discovering cluster-based local outliers*, Pattern Recogn. Lett. **24** (2003), no. 9-10, 1641–1650.
27. S. H. Brown, *Multiple linear regression analysis: a matrix approach with Matlab*, Alabama J. Math. **34** (2009), 1–3.
28. V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, *Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines*, Ore Geol. Rev. **71** (2015), 804–818.
29. D. W. Kim, S. J. Seo, C. W. De Silva, and G. T. Park, *Use of support vector regression in stable trajectory generation for walking humanoid robots*, ETRI J. **31** (2009), no. 5, 565–575.
30. C. Hu, G. Jain, P. Zhang, C. Schmidt, P. Gomadam, and T. Gorka, *Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery*, Appl. Energy **129** (2014), 49–55.

31. J. H. Friedman, *Greedy function approximation: a gradient boosting machine*, Ann. Stat. (2001), 1189–1232.

AUTHOR BIOGRAPHIES

Anitha Ambat received her BTech degree in Information Technology from Calicut University, Kerala, India, in 2009, and MEng in Computer Science and Engineering from Anna University, Tamil Nadu, India, in 2012. She is currently a PhD student in Computer Science at the International Institute of Information Technology Kottayam, Kerela, India. Her research interests include smart home security, digital forensics, Internet of Things, and machine learning.

Jayakrushna Sahoo received his PhD degree in Data Mining from the Indian Institute of Technology, Kharagpur, India, and MTech degree in Computer Science and Engineering from the International Institute of Information Technology (IIIT), Bhubaneswar, India. He worked with the Department of Computer Science and Engineering, BML Munjal University, Gurgaon, India, as an assistant professor. He is also an ad hoc faculty member with the Department of Computer Applications, National Institute of Technology, Jamshedpur, India. Since July 2019, he has been with the Department of Computer Science and Engineering at IIIT, Kottayam. His current research interests include data mining, machine learning, federated learning, digital watermarking, and deep learning applications in bioinformatics.

How to cite this article: A. Ambat and J. Sahoo, *Anomaly detection and prediction of energy consumption for smart homes using machine learning*, ETRI Journal (2024), 1–12, DOI 10.4218/etrij.2023-0155