Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 5, No. 6, December, 2021

Unsupervised Anomaly Detection for Energy Consumption in Time Series using Clustering Approach

Jesmeen M. Z. H. [1*], J. Hossen [1*], Azlan Bin Abd. Aziz [1]

[1] *Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia.*

Abstract

Recent years have seen significant growth in the adoption of smart home devices. It involves a Smart Home System for better visualisation and analysis with time series. However, there are a few challenges faced by the system developers, such as data quality or data anomaly issues. These anomalies can be due to technical or non-technical faults. It is essential to detect the non-technical fault as it might incur economic cost. In this study, the main objective is to overcome the challenge of training learning models in the case of an unlabelled dataset. Another important consideration is to train the model to be able to discriminate abnormal consumption from seasonal-based consumption. This paper proposes a system using unsupervised learning for Time-Series data in the smart home environment. Initially, the model collected data from the real-time scenario. Following seasonal-based features are generated from the time-domain, followed by feature reduction technique PCA to 2-dimension data. This data then passed through four known unsupervised learning models and was evaluated using the Excess Mass and Mass-Volume method. The results concluded that LOF tends to outperform in the case of detecting anomalies in electricity consumption. The proposed model was further evaluated by benchmark anomaly dataset, and it was also proved that the system could work with the different fields containing time-series data. The model will cluster data into anomalies and not. The developed anomaly detector will detect all anomalies as soon as possible, triggering real alarms in real-time for time-series data's energy consumption. It has the capability to adapt to changing values automatically.

Keywords:

Anomaly Detection;
Energy Consumption;
Unsupervised Learning;
Time-series Data.

Article History:

Received:	09	June	2021
Revised:	20	September	2021
Accepted:	27	October	2021
Published:	01	December	2021

1- Introduction

Time series and streaming data are generated every minute and second due to the available sensors and software. These data are obtained mainly by increasing the Internet of Things (IoT) and connecting real-time data sources. The rapid rise in real-time data sources' availability made it difficult to detect anomalies by human in streaming data. Hence, automatic anomaly detection is required by the users to spot anomalies without delay in real-time. For example, detecting an abnormality in energy consumption in smart homes will help prevent fraud, i.e., energy theft. However, anomalies can also produce different problems, such as data collection errors and natural variations [1]. Detecting anomalies is essential in this case; however, detecting unlabeled anomalies in current or streaming data has challenges. Detecting anomalies can help to take action in critical scenarios; however, the system should be reliable and able to discriminate abnormal and seasonal based consumption.

In current years, machine learning (ML) techniques have been broadly implemented to detect and predict abnormal patterns in different applications. ML methods are categorized into three parts which are supervised, semi-supervised,

* **CONTACT**: jesmeen.online@gmail.com; jakir.hossen@mmu.edu.my

DOI: http://dx.doi.org/10.28991/esj-2021-01314

and unsupervised [2]. It contains no labeled anomaly for the obtained time series data, where unsupervised learning tends to be the best. The primary purpose of the study presented in this paper is to detect anomalies for Time-series analysis using unsupervised detection techniques stated by Himeur et al. (2021) [3].

The main contributions of this research will be:

- The practical model for anomaly detection in time-series by the unsupervised learning.
- The system will be able to work with different datasets containing time-series data.

To obtain the main objectives, the significant difficulty was to train the model without labeled data and evaluate the models. The second challenge was to distinguish between normal data and anomaly data. Finally, to overcome the issue of assessing the system Excess Mass score was obtained and additional few sets of datasets with labeled data were used to evaluate the model. Furthermore, this paper proposes a system that relies on two main modules for seasonal anomaly detection. The first one is feature extraction and reduction from time-series-based data, and the second module is anomaly detection using an unsupervised clustering technique.

The structure of the paper is as follows. Section 2 is a literature review of learning models currently used for anomaly detection related to electricity usage data. Section 3 presents the overall framework of the proposed system, including data collection, data preprocessing, feature extraction and selection process, and learning models used. Section 4 contains results of the accuracy for each model for the different model hyperparameters, with some discussion on the results. Finally, in Section 5, the conclusion and future work are stated.

2- Literature Review

Considering the most non-technical losses in electricity is electricity theft, which causes the most severe problem. It results in an increase in economic cost such as excess of electricity bills, reduction of supply quality, the requirement of more significant generation load. Overall it affects the financial system.

It was reported by Holman (2019) [4], Malaysia Tenaga Nasional Bhd (TNB) searched many locations that were assumed of intentionally snooping the electricity supply panel, which had resulted in an economical cost of $25 million to the utility supplier company. This issue is not limited to Malaysia only; it is also faced by many other countries, such as Iran, Turkey, Brazil, Argentina, Venezuela, and a few European countries, including the eastern powerhouse Russia. For example, in Iran, massive tariffs of electricity usage were imposed for mining and had manipulated the miners to reduce the electricity bills.

Therefore, it is essential to find energy theft for the electricity provider by examining smart meters at the residence to discover the conceded reason it will ultimately cost labor. A practical and automatic electricity theft detection system is required to support the electricity providers to overcome the issue. The model will be able to process real energy consumption data in actual applications.

Anomaly detection can impact a different aspect; hence, current studies by the researchers of artificial intelligence and data science. In the case of anomaly detection in time series data, there are works with both supervised (e.g., k-Nearest Neighbor (KNN) [5], Gradient boosting[6] or combination of Naive Bayes and KNN [7]) and unsupervised (e.g. Artificial Neural Networks (ANN) [8]); still the majority work is done in outlier detection in batch data and not suitable for real-time data.

Puig and Carmona (2019) [9] provided a system for an energy company from both technical and business points of view, which benefits from implemented NTL detection system. They stated that the system helped the company to be more sustainable by reducing the gap between the Energy Consumption and Energy Distribution. They had used a supervised ML, which means they had labeled datasets. In this case, it is easier to classify whether the new stream data is an anomaly or not. However, it becomes challenging to train the ML model if the dataset does not contain labelled anomaly. In this case, unsupervised learning might help find unlabeled data by categorizing the data.

Anomaly detection can not only detect abnormal usage, but it can also be used to clean dirty data in time-series which might affect the performance of data analysis outcome, whether for pattern mining [10] or classification [11]. Anomaly detection is commonly also used to maintain data quality [12]. They had calculated the predicted value and confidence coefficient, and by using the threshold value, they had detected whether the value is abnormal or not. They had considered hydrologic time-series data. However, they concluded this model works well for the water level and daily flow monitoring.

Zhang et al. (2017) [13], stated that data noise could be removed by simply removing detected anomalies. However, few researchers had proved that eliminating these types of data can improve the overall performance [14, 15]. Handling abnormal data will overcome one of the data quality dimensions, i.e., inaccurate data. According to Ahmed et al. (2016) [16] study, unsupervised learning can be used to detect an anomaly, especially for the unlabeled dataset. The work-focused in this research is on using unsupervised anomaly detection to make sure the system works commonly for most

time-series data without a pre-labeled dataset. In this case, there are few standard algorithms available to detect these anomalies, such as One-Class Support Vector Machine (OCSVM) [17]. Another researcher [18], had proposed K-Means clustering, local outlier factor (LOF), and multivariate Gaussian distribution (MGD) and found K-means to work effectively for anomaly detection.

3- Research Methodology

The intelligent system for energy usage analytics platform will let the user manage energy usage online and analyse data all in one place. The research methodology to develop an intelligent model for the analytics platform is presented in Figure 1. Initially, in the first step, according to the requirement of the system, time-series data was collected, which was used in the next step. In the next step, the main system development was processed where the feature engineering was held to extract and select the best features. Next, considering the unlabelled data, the system was trained in an unsupervised model, i.e., clustering data. The model was prepared in such a way that it would detect true anomalies and seasonal anomalies. Next, the abnormalities were visualized in a time-series plot. Finally, the system was evaluated and confirmed that the developed model with selected hyper-parameter would effectively detect anomalies.

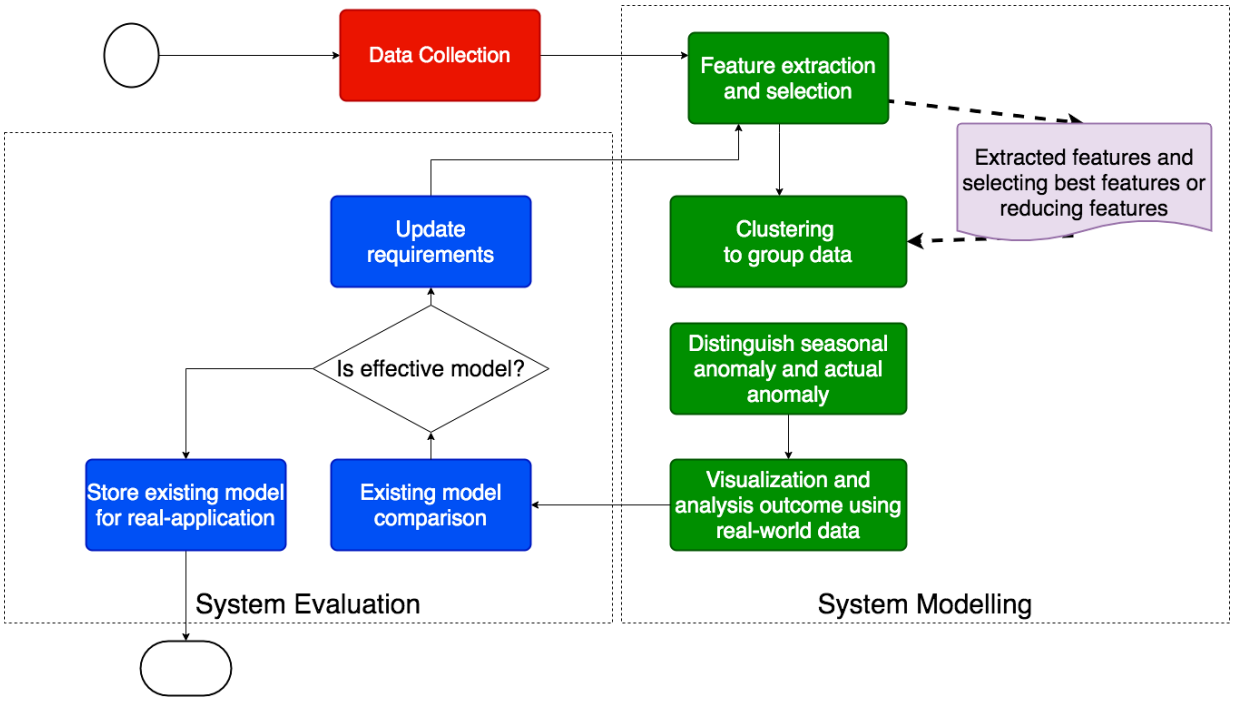


Figure 1. The overall process of the system developed.

In the system evaluation process, the developed model was compared with the existing model; if the model is not effective, the requirements and hyper-parameters will be tuned to get the best model, following with system modeling. Once the best model was obtained for the real application (i.e., analytical platform), it was further used for structuring the proposed framework. The following section presents the details of the final proposed framework used to identify abnormal energy consumption.

4- The Proposed Framework to Identify and Visualize Unusual Energy Consumption

The research aims to identify unusual usage by an unsupervised learning approach in time-series data. The detection will be different for different homes. Hence, the model train is more reliable and not dependent on the specific household. However, the detection is not statistically based, and each model must perform independently according to household electricity consumption.

Figure 2 presents an overall process of the developed system. In phase one, Data was collected from real work smart meters and stored in time-series data for further feature engineering for phase two. Next, features were extracted according to the time domain. To present these features in two dimensions, the PCA algorithm was implemented with 2-components. Finally, after further testing and validating the dataset, LOF outperformed the anomaly detection (comparison presented in Result and discussion section); hence LOF was set in the final framework. Moreover, LOF [19] is a density-based method to detect local outliers and was helpful in several fields, such as fraud detection, intrusion detection, and more [20].

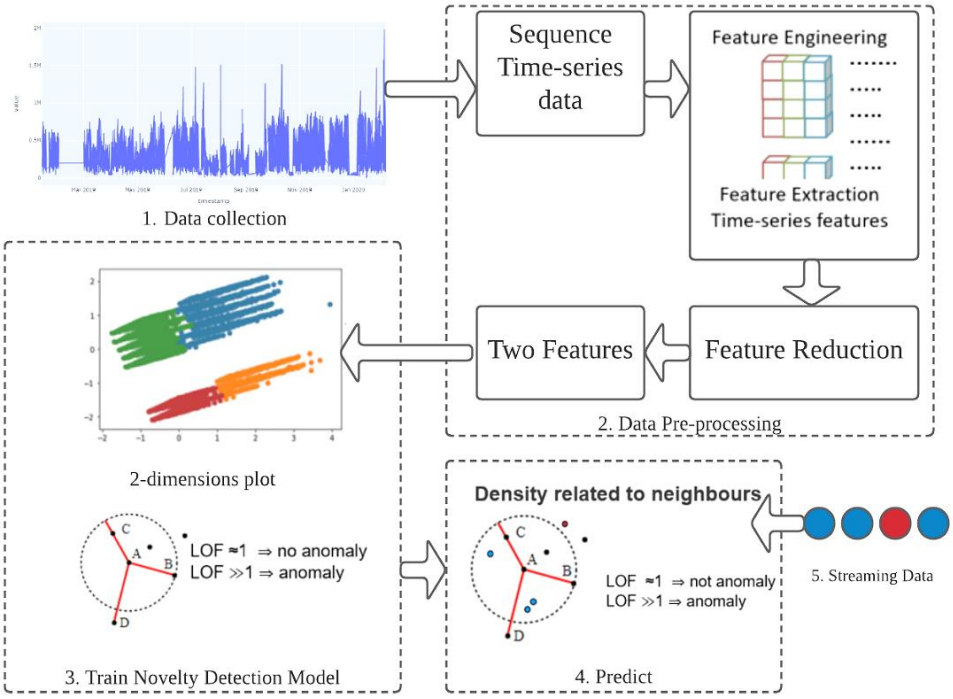


Figure 2. The overall process of the system developed.

The step-by-step procedure is described in this section.

4-1- Data Collection

In the experimental analysis, this paper used a dataset provided by a company from Malaysia for testing and performance comparison of different algorithms:

(1) Dataset 1: Energy Consumption: Real-time dataset provided by Telecom Malaysia is based on the SEMS in a real scenario. The dataset does not contain labeled anomalies. Hence, it is best to test the proposed system. A visualization diagram of time-series data plotted is presented in Figure 3. Here value in the y-axis is energy consumption.

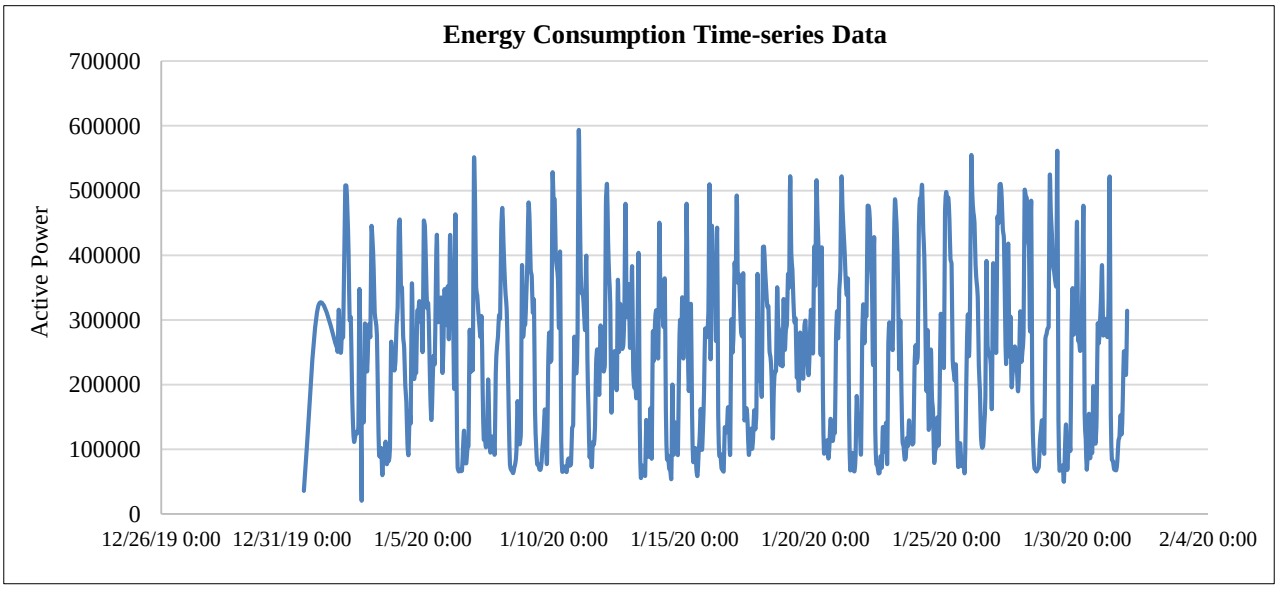


Figure 3. One Month Energy Consumption Time-series Data Plotted.

Consider, D value as $\{d(n)\}N$ n=1 denotes a dataset composed of N independent sequences of observations(here N= 9000). Each sequenced(n) has T timesteps, i.e. d(n) = (d(n) 1 , d(n) 2 , ..., d(n) T), and for every consumption at timestep t (i.e., x(n) t) has dx-dimensional vector. As the selected dataset is the univariate dataset, dx is 1. Therefore, the training dataset $\mathcal{X}$ has dimensions (N,…,1). After feature engineering and feature reduction, selecting two top features, the $\mathcal{X}$ training set with ((N,..., 2) dimensions.

Furthermore, to evaluate the proposed system of the energy consumption data, a subset (DF) of the dataset $\mathcal{X}$ of univariate time series ($dx = 1$) was obtained from smart meters from residential installation. The train dataset x of 752 daily sequences of January 2020 were selected from the dataset's complete set. Samples contained 30 days of daily observation of around 24 instances of sequence data, as shown in Figure 4.

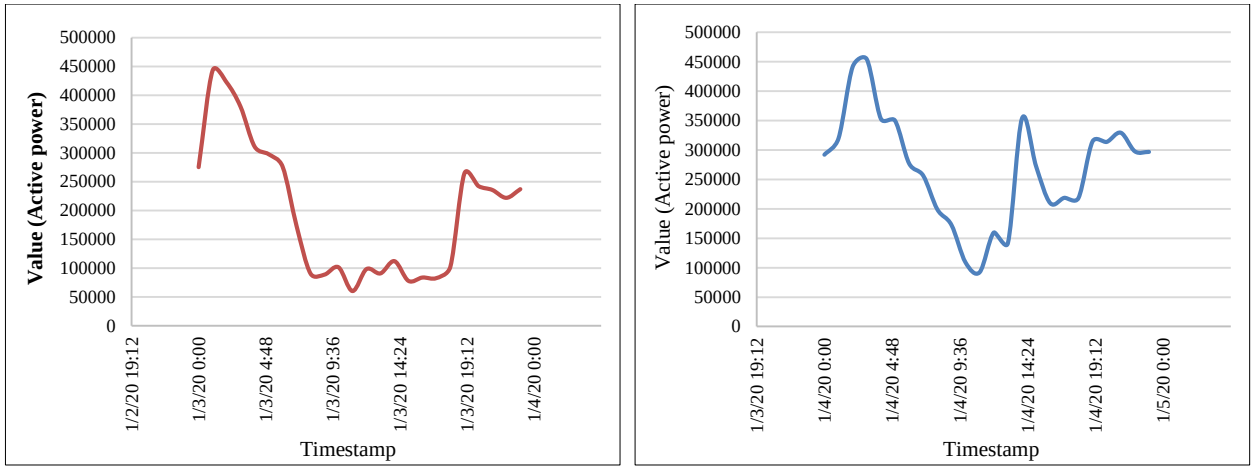


Figure 4. 24-hour Energy Consumption Time-series Data Plotted.

4-2- Data Preprocessing

4-2-1- Feature Engineering

The dataset contains univariate data. For better detection, few valuable features were extracted from the data as following, where enumerating the properties of a timestamp that was helpful for the anomaly detection implementation for residential buildings:

- Hour of day.
- Night or day
- Weekend or not (the day of the week (Monday=0, Sunday=6))

Finally, the data from the meter was classified into one of the four temporal contexts as following; The extracted features were plotted in a histogram for better visualization. The time was converted to 'int' value for better plotting, as shown in Figure 5:

- Weekend at daytime: in the daytime when the possibility to have more significant usage.
- Weekend at nighttime: in time low light, possibility to have common usage.
- Weekday at daytime: in the daytime when the possibility to have more significant usage but lower or higher than weekend daytime, depending on the active power value
- Weekday at nighttime: in the nighttime when the possibility to have low usage.

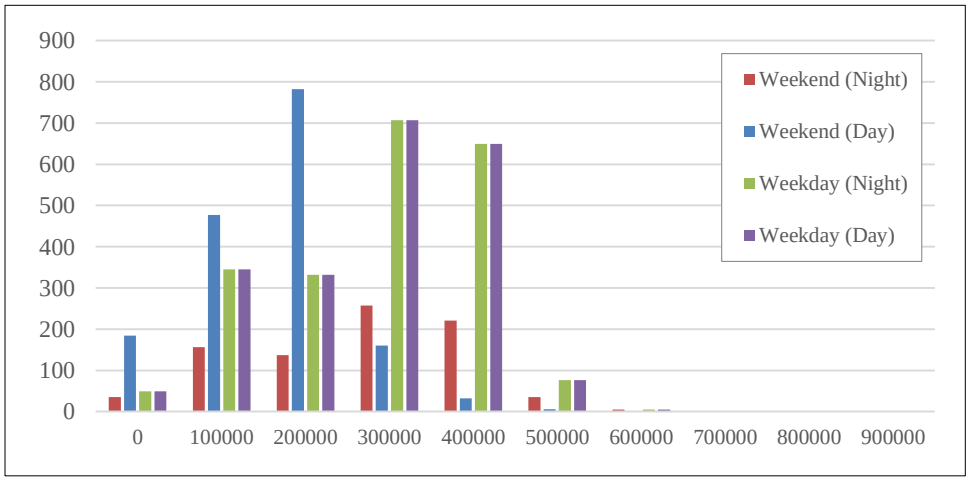


Figure 5. Plotting of weekend/weekday and night/day.

4-2-2- Feature Reduction

A total of five features (i.e., value, hours, daylight, Day of The Week, Week Day) were extracted from the previous step. Hence, it had many dimensions and must be reduced into lower dimensions with two features containing such vital information as possible—the best way to do it was by using a correlation algorithm for feature dimension reduction. Principal component analysis (PCA), the most common unsupervised technique, was used to shrink two-dimension features. It is able to retain complete information by getting the most out of the variance between data [21]. There is another well-known unsupervised feature reduction algorithm Singular Value Decomposition (SVD). Algorithm steps are different in PCA and SVD. However, they generate the exact value of eigenvectors and eigenvalues [22]. Here both algorithms were tried and the same features were obtained.

Another standard feature reduction method IsoMap was also evaluated and it is found that PCA is outperforming in performance. PCA is very well known for linear structure datasets. However, in our case, the non-linear data tends to provide good results due to feature engineering.

4-3- Train Novelty Detection

The implemented approaches in this paper are designed in such a way that they can handle low-dimensional system characterization and energy consumption analysis. Therefore, it will mainly help detect anomalies in a smart energy monitoring system (SEMS) containing streaming or time-series data.

4-3-1- Unsupervised Learning Model

Comparison of the local outlier factor (LOF) algorithm presented in Figure 1 with other standard unsupervised learning models, one class SVM [23] (OCSVM), Elliptic Envelope(EE) [24], isolation forest algorithm [25] (iForest), and Gaussian Mixture Modelling (GMM).

- OneClassSVM (OCSVM): (supervised or semi-supervised) is considered for the dataset, which can fit in a tight decision boundary around a set of regular points. This model will not perform well when the training dataset is mixed already with the dataset containing outliers.
- Elliptic Envelope (EE): (unsupervised or supervised) similarly to OCSVM, it fits well with the tightest Gaussian (smallest volume ellipsoid). It basically distinguishes the expected data from anomaly by creating an oval shape boundary.
- IsolationForest (iForest): (unsupervised) follows the decision tree technique but trains the model without a labeled dataset. As a result, it performs well in the case of higher-dimensional sets.
- Gaussian Mixture Modelling (GMM): works by using a parametric probability density function; it represents a weighted sum of Gaussian component densities [26]. A GMM is implemented as a probabilistic model for clustering active power data. It considers all usage points derived from a finite Gaussian distribution mixture with unknown parameters. GMM model is used as unlabeled cluster data. It does account for variance type of data.

The LOF algorithm presented in [19] is a very efficient way of performing anomaly detection, for the ability to define the detection is for outlier or novelty. Indicating abnormal data was calculated by a score reflection known as the local outlier factor. First, as shown in Figure 1, Step 3, the model was trained by calculating the local deviation of a provided data point (A) concerning its neighbor data points. Next, in step 4, once the streaming data was entered into the model, it would be able to plot whether the information is normal or not.

It is processed by comparing one point's local density close to others and identifying that the new point has a substantially lower density score than their neighbors. Here, the score is calculated from its density and its k-nearest neighbors' value. For example, let the Euclidean distance between two data points A and B be ED(A,B). And KD(A) be the distance of point A and its k^{th}-nearest neighbor.

Reachability distance (RD) among the two data points, 'a' and 'p' is calculated using RD(a, p) by Equation 1.

$$\mathrm{RD}(\mathrm{a},\mathrm{p}) = \max\{\mathrm{K}-\mathrm{distance}(\mathrm{a}),\mathrm{d}(\mathrm{p},\mathrm{a})\}, \quad (1)$$

Next Local reachability density by integrating Equation 1 in Equation 2 to obtain data point 'a' density.

$$LRD_k(a) = \left(\frac{1}{k}\sum_{i=1}^{k} RD_k(a^i, a)\right)^{-1}, \quad (2)$$

Finally, as shown in Figure 1, to determine if usage data is standard or not, the Local outlier factor for point 'a', LOF(a), is obtained using Equation 3.

$$LOF_k(a) = \frac{\frac{1}{k}\sum_{i-1}^{k} LRD_k(a^i)}{LRD_k(a)} \quad (3)$$

After obtaining the value of LOF, the decision of the data anomaly is processed as follows:

- If LOF(k) ~ 1, then the data (k) is similar to density as its neighbors; hence it is normal data;
- If LOF(k) < 1, then the data (k) is greater density than neighbors; hence it is inlier (Anomaly);
- If LOF(k) > 1, then the data (k) is Lower density than neighbors; hence it is an outlier (Anomaly).

4-3-2- System Evaluation

One known technique to evaluate the anomaly detection technique is Excess Mass(EM) [27]. As for the labeled dataset, ROC or Precision-Recall (PR) method effectively measures the performance of a model. However, as here we have an unlabeled dataset, it was found that EM is good to find the score to evaluate the learning models. According to [27], they had proved that EM could work alternatively to ROC and PR. Basically, ROC and PR are based on EM and Mass-Volume (MV) curves too. Even by using the available labelled anomaly data of time-series, the system was evaluated further by calculating precision using equation 4. TP (true positive) indicates the number of anomaly windows correctly detected, and FP (false positive) indicates the number of anomaly windows detected incorrectly, i.e., the normal data presenting as anomalies.

$$precision = \frac{TP}{TP+FP} \quad (4)$$

5- Result and Discussion

The system evaluated detailed results are presented with and without PCA in Tables 1 and 2, respectively. For both cases, the anomaly fraction is 0.01, as detected abnormalities percentage is small size compared to the complete dataset. Furthermore, as the input is univariate time-series data, there is only one feature as input without feature engineering. Moreover, the PCA reduced features and calculated the top two features with feature engineering from five features. Hence, it is essential to set the selected features as '1' and '2' for processing without and with PCA, respectively, for evaluation purposes to obtain EM and MV scores.

Table 1 indicates the developed model's performance without the PCA feature reduction technique. Without PCA, OCSVM tends to perform better than other techniques. It has obtained a 5.682 e-07EM value which is more significant than other models after setting the hyperparameters to gamma value to 0.001. The gamma parameter was set to fix the value to influence the radius on the kernel. The default parameter is scale, i.e. 1 / (n_features * X.var()), where n_features is number of features used to train the model and X.var() is the variance of the train set (X). For OCSVM, different parameters were tested to evaluate the difference in results for different parameters. However, OCSVM did not tend to provide good output and was finally ignored in the main framework.

It is a comparable value to iForest and EE. Therefore, GMM is also looking to perform well in this case. However, in further analysis using labeled data, it was found that GMM showed mostly False Positive output.

Table 1. Results for the novelty detection setting without PCA, where n_features=1.

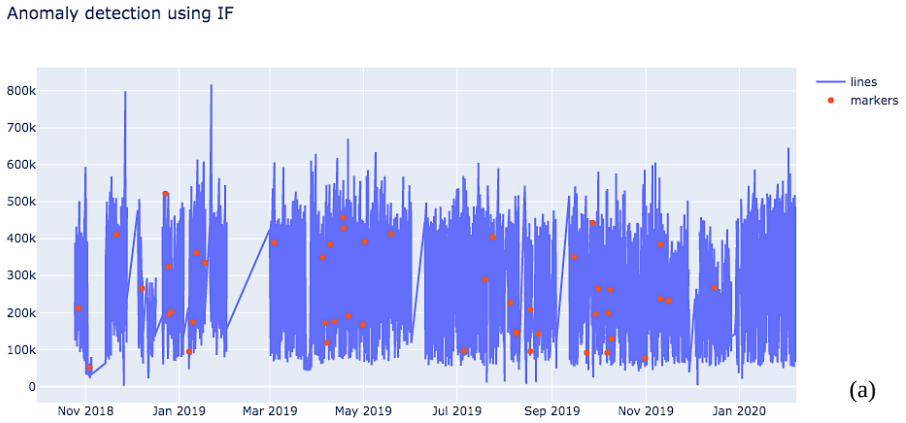
Model	Parameters changes (fraction = 0.01)	AUC EM	EM	AUC MV	MV
iForest	contamination= fraction	1.245e-07	0.0226	25131.6	256445.8
OCSVM	null	1.244e-07	0.0223	50624.6	517737.9
OCSVM	nu=0.95 * fraction	1.151e-07	0.0148	**73781.2**	**753159.3**
OCSVM	nu= fraction,kernel='rbf', gamma=.001	**5.682 e-07**	**0.0591**	15475.8	157919.6
LOF	novelty=True	1.104e-07	0.0132	25129.7	256426.0
EE	contamination=fraction, random_state=1	1.244e-07	0.0228	25129.7	256426.0
GMM	random_state=1	1.198e-07	0.0218	25129.7	256426.0

Table 2. Results for the novelty detection setting with PCA, where n_features=2.

Model	Parameters changes (fraction = 0.01)	AUC EM	EM	AUC MV	MV
iForest	contamination= fraction	0.00593	0.02857	0.30352	3.09717
OCSVM	null	0.00555	0.01551	1.16204	11.87662
OCSVM	nu=0.95 * fraction	0.00833	0.01747	0.94527	9.66020
OCSVM	nu= fraction,kernel='rbf', gamma=.001	0.00556	0.00825	**1.34642**	**13.75148**
LOF	novelty=True	**0.01263**	**0.05715**	0.30378	3.09980
EE	contamination=fraction, random_state=1	0.00554	0.00912	1.38157	14.10990
GMM	random_state=1	0.00554	0.00906	1.34570	13.74465

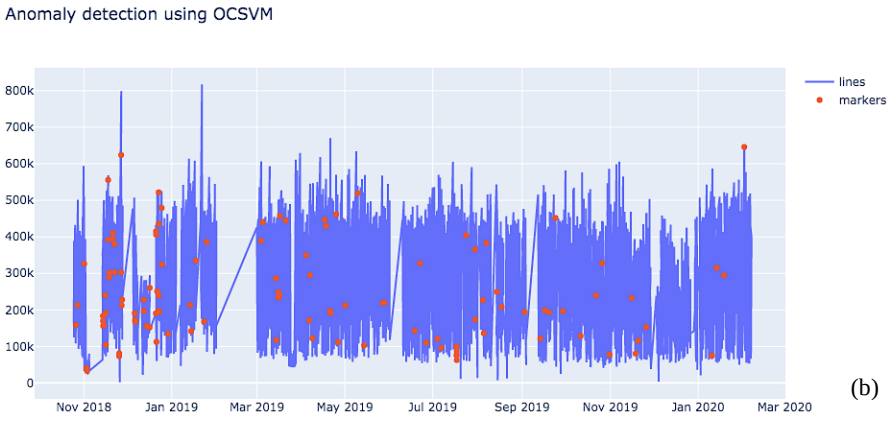
By comparing these results with Table 2 results, it was found that the system's overall performance of AUC EM score is better with PCA processing. For example, the PCA AUC EM value is 0.01263 in LOF, and without a PCA-based model, the AUC EM score is 1.104e-07. It is a significant difference. Moreover, the EM score for LOF is 0.05715 with PCA, and without PCA the score is 0.0132. It was also found that LOF tends to perform better in the model based on the feature training based on the PCA.

The next step used the PCA model with the unsupervised model to plot anomalies in the time series. According to the results in table 2, LOF fits best for detection purposes. However, the anomalies are plotted for every model to demonstrate the actual occurrence of an anomaly in the time frame (Figure 6). It shows OCSVM showing more than actual anomalies. The visualization of anomalies is presented in Figure 6. Here, N is 9000 (Recent 9000 consumption data). The actual consumption of matrices is plotted in blue lines, and anomalies are highlighted using red points. The output is shown for five trained models (a) iforest, (b) OCSVM, (c) EE, (d) LOF, and (e) GMM. The detection output is seasonally based. The dataset is divided into categories of weekend, weekday, day, and night. In the case of time-series consumption data, this is one of the important characteristics and should be considered. If the system could not distinguish between actual and seasonal anomaly, most of the anomaly would be the point to maximum usage (for example, usage beyond 600K of active power).



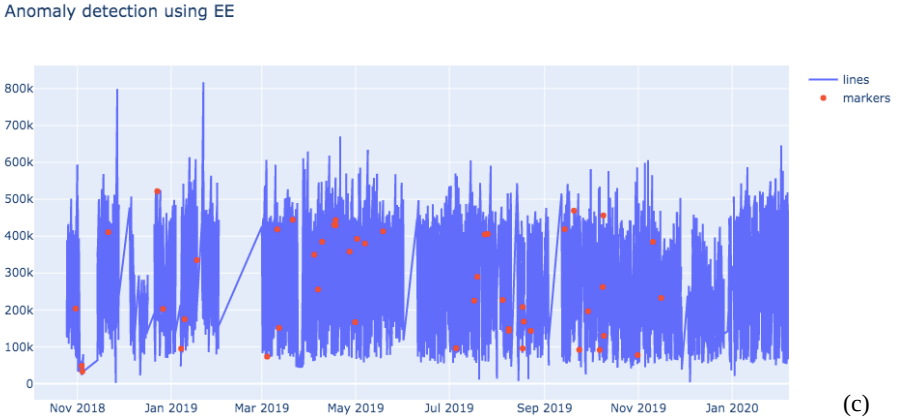
(a)

Normal usage: 8953, anomalies: 47



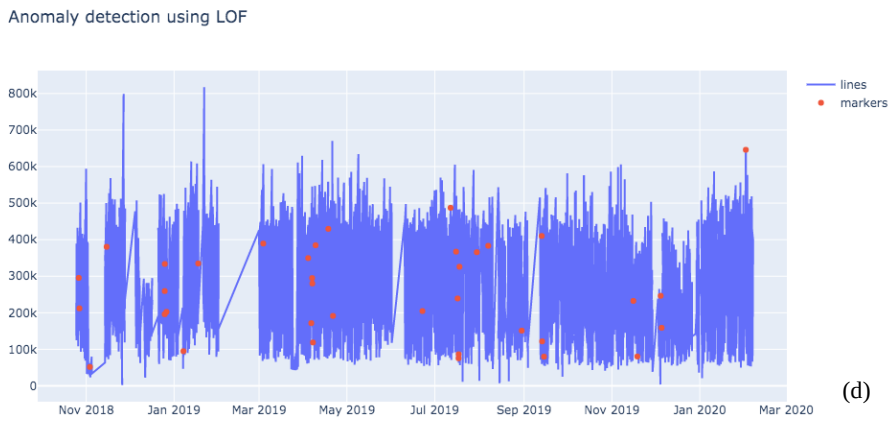
(b)

Normal usage: 8888, anomalies: 112



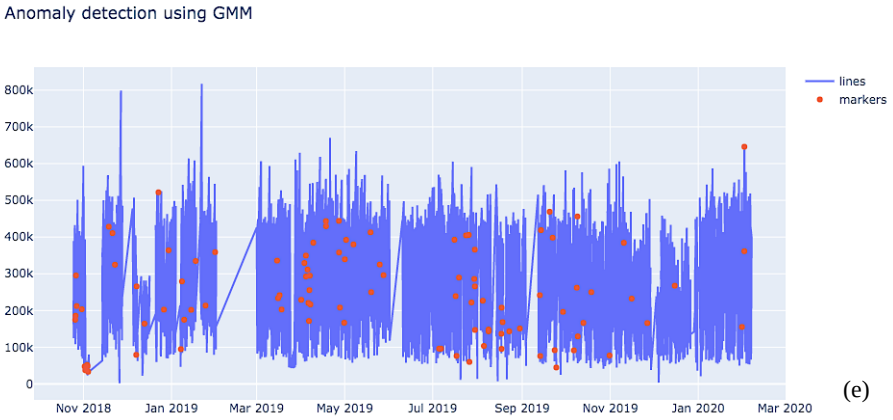
(c)

Normal usage: 8953, anomalies: 47



(d)

Normal usage: 8968, anomalies: 32

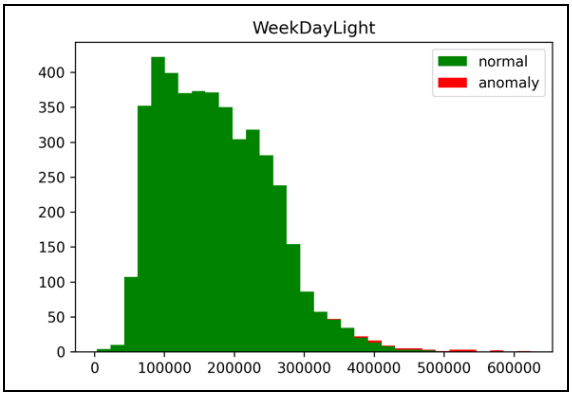


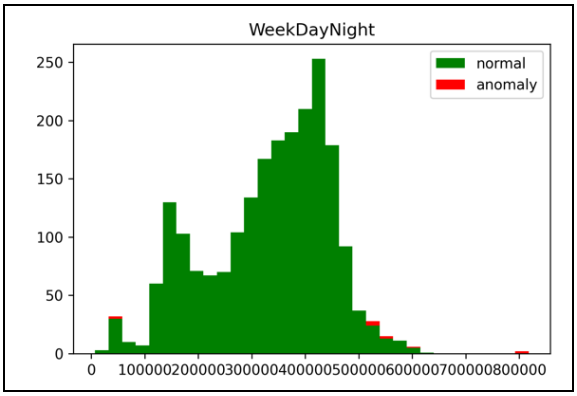
(e)

Normal usage: 8900, anomalies: 100

Figure 6. Categorical anomaly detection (weekend, weekday, day, night) for outlier percentage 0.5% (a) iForest (b) OCSVM(c) EE (d) LOF (e) GMM.

According to the plot shown in Figure 6, it can be seen that some anomalies points in the different models are similar. However, after further investigating, the detection anomalies value is plotted in the histogram, as shown in Figures 7 to 11. These figures show histograms with anomalies identified as points in red color and normal consumption in green. It demonstrates that at weekend nighttime, around 0.6m active power is considered anomaly white using iForest and GMM. Active power greater than 0.5m is considered an anomaly during the weekend daytime, while using iForest, EE and GMM. By analyzing weekday light time, more significant than ~0.45m active power is regarded as anomalies for iForest, OCSVM, EE, and GMM. The evaluation data presented in Table 2 shows that LOF provides a good detection score. Next, here LOF is compared with other models and finding the common anomalies with LOF. It can clearly understand that the number of anomalies detected by LOF is smaller than other models. However, anomalies point in LOF detected is almost same in different model's detection output. It can be visualized from Figure 10 LOF model, in the weekday nighttime, active power at 0.8M is considered anomaly as by other models (iForest, OCSVM, EE, and GMM). Moreover, this anomalies detection is not based on active power value, because the system was developed in such a way it will detect seasonal anomalies, rather than anomaly beyond a value of the active power value.





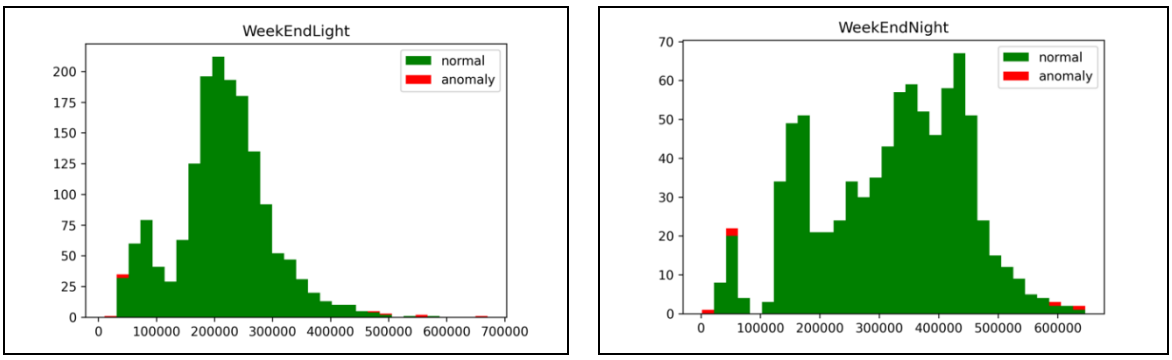


Figure 7. Plot frequency of normal and anomaly data detected by IsolationForest.

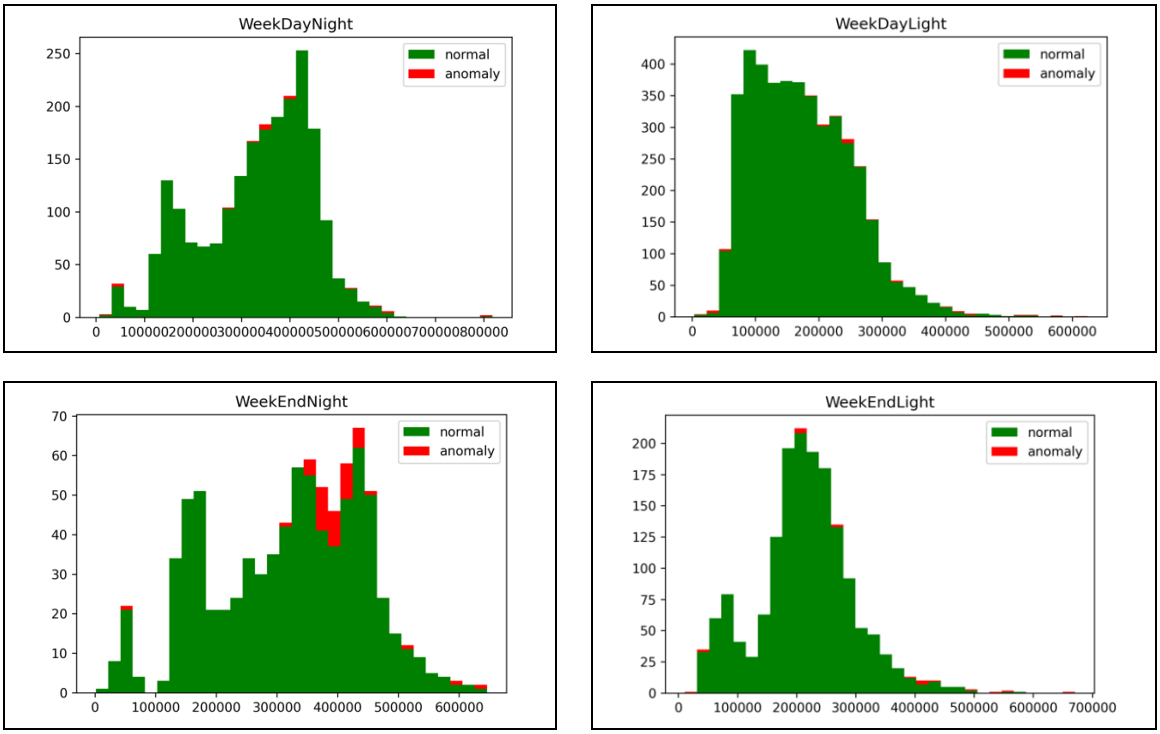


Figure 8. Plot frequency of normal and anomaly data detected by OCSVM.

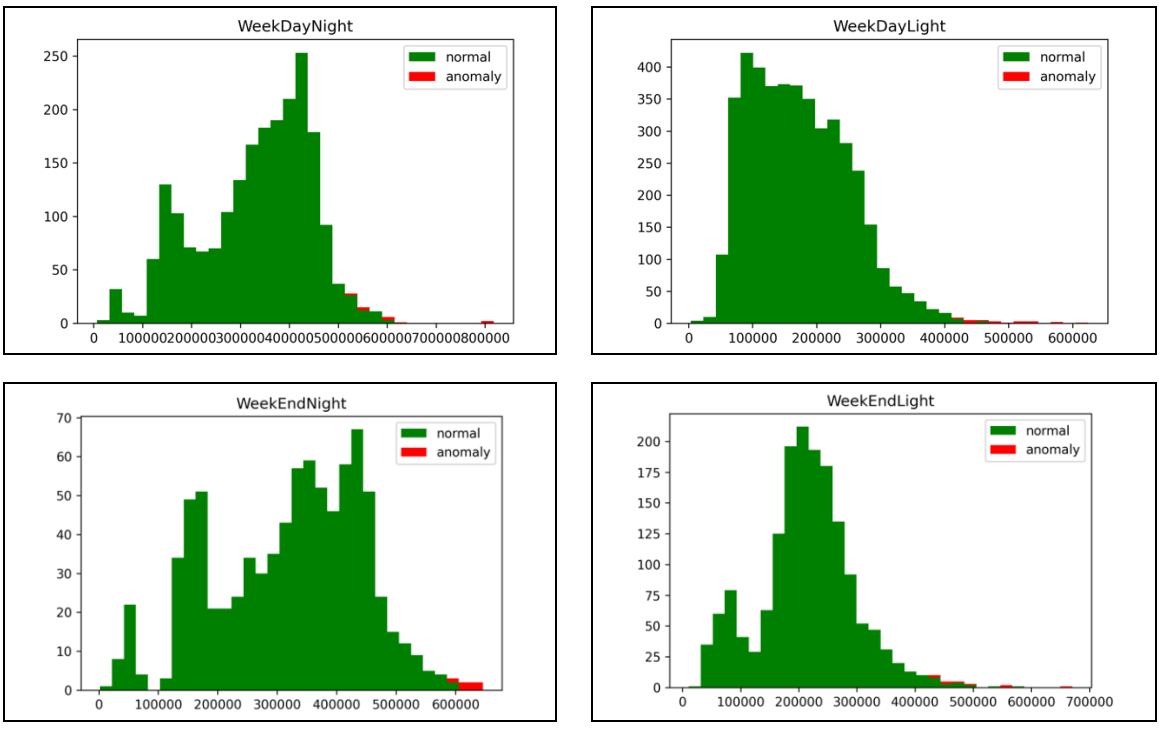


Figure 9. Plot frequency of normal and anomaly data detected by EE.

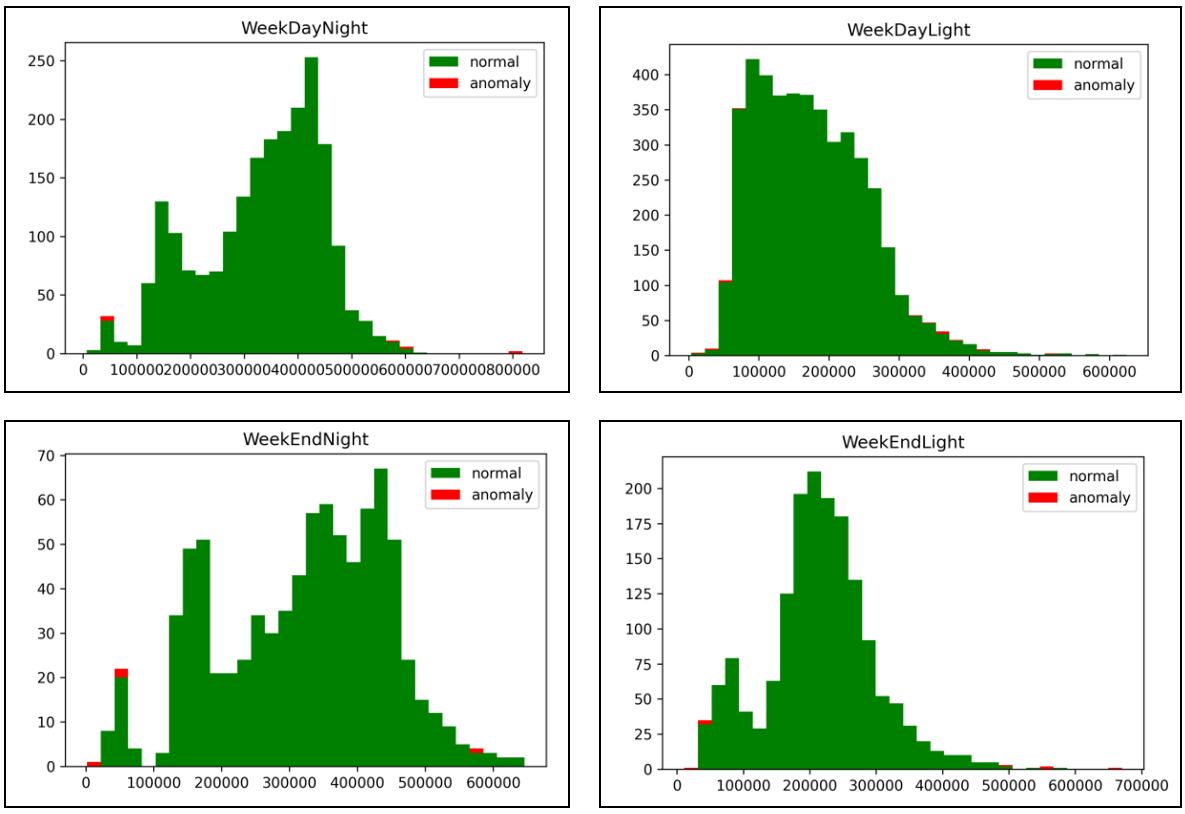


Figure 10. Plot frequency of normal and anomaly data detected by LOF.

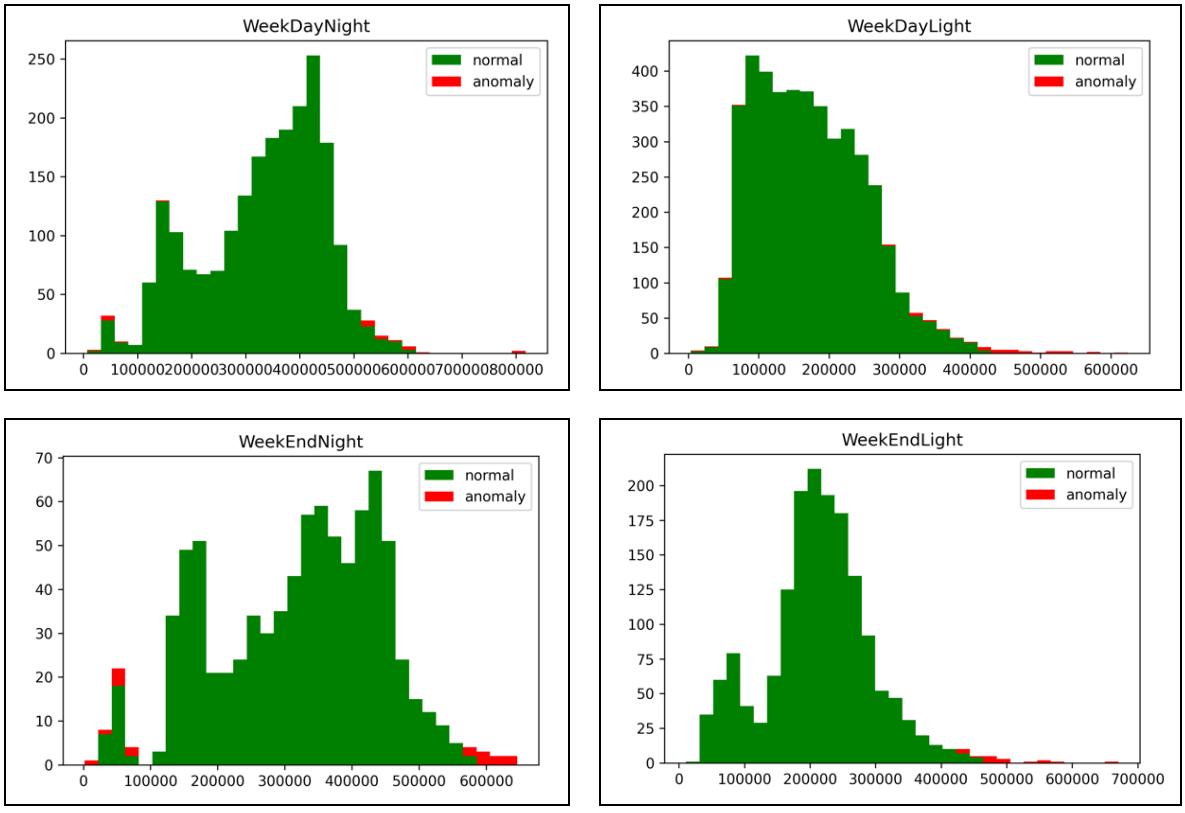


Figure 11. Plot frequency of normal and anomaly data detected by GMM.

For further validating the developed system, the unsupervised learning was trained using a benchmark dataset [28], known as NAB. Here, the dataset was used from the NAB. The three datasets are (a)ec2_request_latency_system_failure, (b) machine_temperature_ system_failure, and (c) art_daily. These three datasets are time-series data, where art_daily data contain data sets without anomalies and anomalies. This set of data will help to confirm that the system also works for novelty anomaly detection.

Table 3. Results for the novelty detection setting with PCA for 'ec2_request_latency_system_failure', where n_features=2.

Model	Parameters changes (fraction = 0.001)	AUC EM	EM	AUC MV	MV
iForest	contamination= fraction	0.00786	0.02905	0.10402	1.06143
OCSVM	null	0.00720	0.01477	1.04453	10.66418
OCSVM	nu=fraction	0.00936	0.01714	0.79956	8.16405
OCSVM	nu= fraction,kernel='rbf', gamma=.001	0.00574	0.00939	1.43030	14.60228
LOF	novelty=True	**0.01794**	**0.04024**	0.10402	1.06144
EE	contamination=fraction, random_state=1	0.00431	0.01302	**1.86950**	**19.0755**
GMM	random_state=1	0.00575	0.01008	1.30745	13.34970

As presented in Table 3, the output score for novelty detection using a different dataset (ec2_request_latency_system_failure), here too seasonal anomaly detection output is presented. Where LOF tend to work better than other models. Where GMM is mostly providing false positive anomaly detection, and OCSVM was detecting more than expected anomalies. The visualization output is presented in Figure 12.

Figure 12. Visualization of detected anomaly for ec2_request_latency_system_failure from NAB dataset.

Taking another benchmark dataset from NAB, the proposed model detected the anomaly and compared it with other available models. The overall performance of LOF tends to be better than other models. The results obtained are presented in Table 4. The detected anomaly is visualized in Figure 13. It was also evaluated without PCA. The model provided a lower score than the system with feature extraction and PCA.

Table 4. Results for the novelty detection setting with PCA for 'machine_temperature_system_failure', where n_features=2.

Model	Parameters changes (fraction = 0.003)	AUC EM	EM	AUC MV	MV
iForest	contamination= fraction	0.01036	0.01538	0.22281	2.27359
OCSVM	nu= fraction,kernel='rbf', gamma=.001	0.01034	0.00509	0.94044	9.59662
LOF	novelty=True	**0.02536**	**0.03963**	0.22350	2.28067
EE	contamination=fraction, random_state=1	0.00945	0.00917	0.96243	9.82039
GMM	random_state=1	0.00945	0.00522	0.96031	9.79879

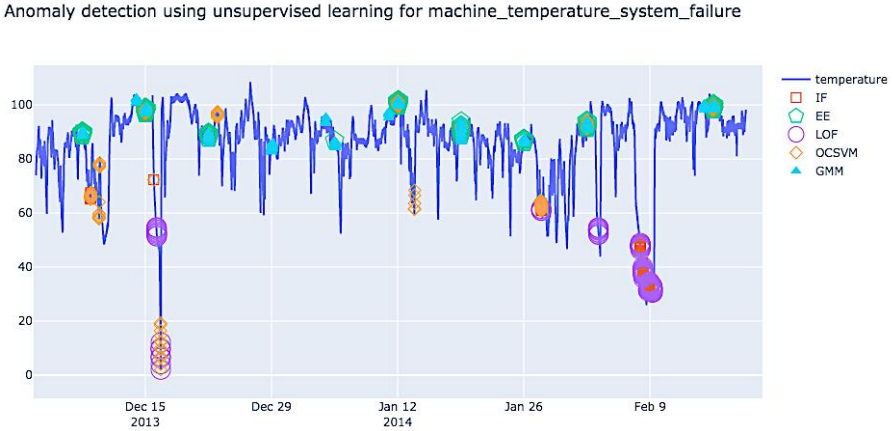


Figure 13. Visualization of detected anomalies for machine temperature system failure from NAB dataset.

As indicated in Figure 13, it shows that LOF detected anomaly in 4 positions where 3 of them are true positive detection as stated by the dataset provider in [28]. Details number of anomaly detection positions with precision is calculated and presented in Table 5. Here, precision is computed using equation 4. The precision of the proposed model with LOF showing to provide 0.75, which is the best performance. Were EE and GMM calculated precision is 0, this means detection using EE and GMM provides all false positive detection.

Table 5. Percentage error obtained for machine temperature system failure from NAB dataset.

Model	TP	FP	Precision
iForest	1	2	0.33333
OCSVM	2	5	0.28571
LOF	3	1	**0.75000**
EE	0	5	0.00000
GMM	0	11	0.00000

Following the concept of novelty detection, the model was trained using data without containing anomaly and then predicted the anomaly using a dataset containing anomalies. It can be clearly visualized that EE had shown wrong detected anomaly compared to other models where they indicated that LOF, OCSVM and GMM exhibited almost similar anomaly. Moreover, IF also detected the jumps up data, but it was not able to detect all the anomalies. According to the EM score, LOF tends to perform the best for anomaly detection.

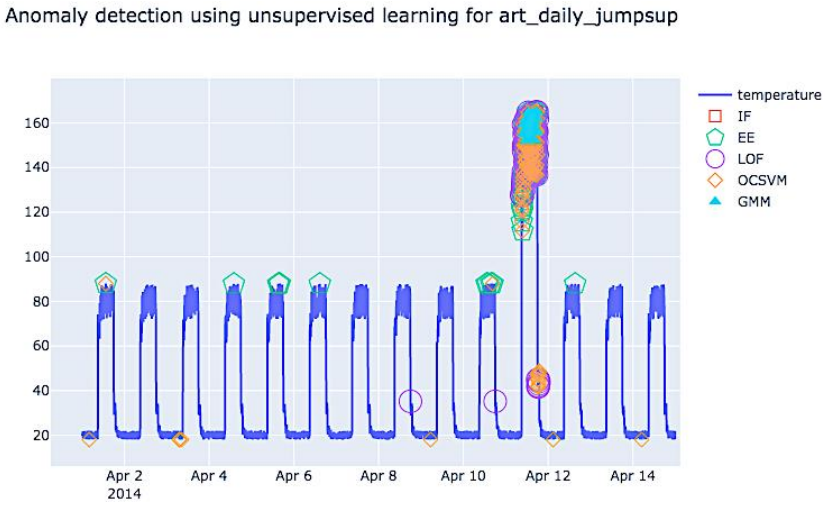


Figure 14. Visualization of novelty anomaly detection using trained unsupervised learning model for jumps up in art daily dataset from NAB.

From the indication of GMM shown in Figure 14 it shows that GMM indicates correct anomalies. However, once the jump is down than the normal data, it shows that the normal data is an anomaly (as shown in figure 15). So here LOF also is revealing the correct anomalies.

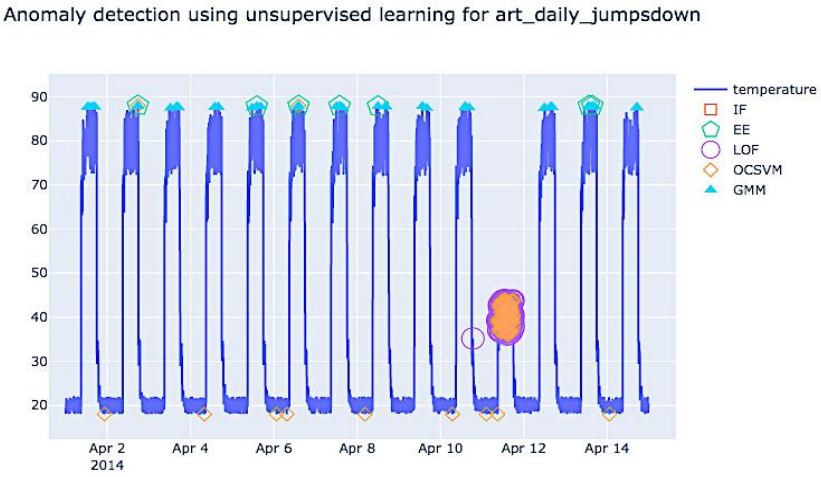


Figure 15. Visualization of novelty anomaly detection using trained unsupervised learning model for jumps down in art daily dataset from NAB.

Here, LOF tends to outperform in this research as it took both local and global properties of the training data into consideration. It can achieve the wanted outcome even when the dataset contains abnormal data samples with different densities values. It is able to find the isolated usage value depending on the value of its surrounding.

6- Conclusion

In this paper, the research is a process to solve the issue of abnormal data without any given labeled truth dataset. This model is capable of solving inaccurate data obtained in time-series data to improve the overall performance of the analytical process. The study presents anomaly detection in the real dataset of electricity usage and obtains LOF to work better than other available models. It has also proved that the feature engineering and reduction process tend to enhance the framework of the detection process. A system is proposed and developed using PCA and LOF to extract time-domain features and automatically detect anomalies. The model is currently tested with two types of time-series datasets. The model was first trained in residential electricity usage data, and it illustrated a significant improvement in anomaly detection using the proposed framework. For further confirmation of the working of the system, the model was trained using the NAB dataset and also found that PCA-LOF tends to work for different datasets. On the NAB dataset, the precision of the proposed model is best with a 0.75 score, compared to OCSVM, EE, iForest and GMM. Moreover, due to features extraction, the model will be able to detect the abnormal data in seasonal based.

In future, unsupervised known feature reduction algorithms (such as PUFS and KPCA) can be implemented to obtain more accurate experimental results. Currently, the features extraction and reduction process are conducted using univariate time-series data. In future, the system can consider using multivariate time-series data, which might enhance the overall performance.

7- Declarations

7-1- Author Contributions

J.M.Z.H., J.H., and A.B.A.A. contributed to all the research methodology process and implementation of the proposed research, to the analysis of the results and the writing of the manuscript. All authors have read and agreed to the published version of the manuscript.

7-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7-3- Funding and Acknowledgements

The authors would like to thank Multimedia University Malaysia (MMU) for supporting this research under Fisabilillah R&D Grant Scheme (FRDGS). This research was conducted in Centre for Engineering Computational Intelligence, Faculty of Engineering & Technology, Multimedia University, Malaysia.

7-4- Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

8- References

[1] Madhuri, G. Sandhya, and M. Usha Rani. "Anomaly Detection Techniques." SSRN Electronic Journal 7 (2018): 449–453. doi:10.2139/ssrn.3167172.

[2] Oladipupo, Taiwo. "Types of Machine Learning Algorithms." In New Advances in Machine Learning, 2010. doi:10.5772/9385.

[3] Himeur, Yassine, Khalida Ghanem, Abdullah Alsalemi, Faycal Bensaali, and Abbes Amira. "Artificial Intelligence Based Anomaly Detection of Energy Consumption in Buildings: A Review, Current Trends and New Perspectives." Applied Energy 287 (2021): 116601. doi:10.1016/j.apenergy.2021.116601.

[4] Holman, Trevor. "Electricity Theft for Bitcoin Mining Imposes Loss of $25 Million in Malaysia." Cryptonewsz, 2019. Available online: https://www.cryptonewsz.com/electricity-theft-for-bitcoin-mining-imposes-loss-of-25-million-in-malaysia/37197/ (accessed on February 2020).

[5] Wang, Zhe, Thomas Parkinson, Peixian Li, Borong Lin, and Tianzhen Hong. "The Squeaky Wheel: Machine Learning for Anomaly Detection in Subjective Thermal Comfort Votes." Building and Environment 151, no. January (2019): 219–27. doi:10.1016/j.buildenv.2019.01.050.

[6] Frery, Jordan, Amaury Habrard, Marc Sebban, Olivier Caelen, and Liyun He-Guelton. "Efficient Top Rank Optimization with Gradient Boosting for Supervised Anomaly Detection." Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017. doi:10.1007/978-3-319-71249-9_2.

[7] Pajouh, Hamed Haddad, Gholam Hossein Dastghaibyfard, and Sattar Hashemi. "Two-Tier Network Anomaly Detection Model: A Machine Learning Approach." Journal of Intelligent Information Systems 48, no. 1 (2017): 61–74. doi:10.1007/s10844-015-0388-x.

[8] Cauteruccio, Francesco, Giancarlo Fortino, Antonio Guerrieri, Antonio Liotta, Decebal Constantin Mocanu, Cristian Perra, Giorgio Terracina, and Maria Torres Vega. "Short-Long Term Anomaly Detection in Wireless Sensor Networks Based on Machine Learning and Multi-Parameterized Edit Distance." Information Fusion 52 (2019): 13–30. doi:10.1016/j.inffus.2018.11.010.

[9] Puig, Bernat Coma, and Josep Carmona. "Bridging the Gap between Energy Consumption and Distribution through Non-Technical Loss Detection." Energies 12, no. 9 (2019). doi:10.3390/en12091748.

[10] Moerchen, Fabian. "Algorithms for Time Series Knowledge Mining." In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006:668–73, 2006. doi:10.1145/1150402.1150485.

[11] Saad, Akram, and N. Sisworahardjo. "Data Analytics-Based Anomaly Detection in Smart Distribution Network." In International Conference on High Voltage Engineering and Power Systems, ICHVEPS 2017 - Proceeding, 2017-January:1–5, 2017. doi:10.1109/ICHVEPS.2017.8225855.

[12] Yu, Yufeng, Yuelong Zhu, Shijin Li, and Dingsheng Wan. "Time Series Outlier Detection Based on Sliding Window Prediction." Mathematical Problems in Engineering 2014 (2014). doi:10.1155/2014/879736.

[13] Zhang, Aoqian, Shaoxu Song, Jianmin Wang, and Philip S. Yu. "Time Series Data Cleaning: From Anomaly Detection to Anomaly Repairing." Proceedings of the VLDB Endowment 10, no. 10 (2017): 1046–57. doi:10.14778/3115404.3115410.

[14] Song, Shaoxu, Chunping Li, and Xiaoquan Zhang. "Turn Waste into Wealth: On Simultaneous Clustering and Cleaning over Dirty Data." In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015-August:1115–24. New York, NY, USA: Association for Computing Machinery, 2015. doi:10.1145/2783258.2783317.

[15] He, Guoliang, Yong Duan, Rong Peng, Xiaoyuan Jing, Tieyun Qian, and Lingling Wang. "Early Classification on Multivariate Time Series." Neurocomputing 149, no. PB (2015): 777–87. doi:10.1016/j.neucom.2014.07.056.

[16] Ahmed, Mohiuddin, Abdun Naser Mahmood, and Jiankun Hu. "A Survey of Network Anomaly Detection Techniques." Journal of Network and Computer Applications 60 (2016): 19–31. doi:10.1016/j.jnca.2015.11.016.

[17] Wang, Xinlin, and Sung Hoon Ahn. "Real-Time Prediction and Anomaly Detection of Electrical Load in a Residential Community." Applied Energy 259, no. 114145 (2020). doi:10.1016/j.apenergy.2019.114145.

[18] Zhao, Shizhen, Wenfeng Li, and Jingjing Cao. "A User-Adaptive Algorithm for Activity Recognition Based on K-Means Clustering, Local Outlier Factor, and Multivariate Gaussian Distribution." Sensors (Switzerland) 18, no. 6 (2018). doi:10.3390/s18061850.

[19] Breuniq, Markus M., Hans Peter Kriegel, Raymond T. Ng, and Jörg Sander. "LOF: Identifying Density-Based Local Outliers." In SIGMOD Record (ACM Special Interest Group on Management of Data), 29:93–104, 2000. doi:10.1145/335191.335388.

[20] Rai, Arun Kumar, and Rajendra Kumar Dwivedi. "Fraud Detection in Credit Card Data Using Unsupervised Machine Learning Based Scheme." In Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020, 421–26, 2020. doi:10.1109/ICESC48915.2020.9155615.

[21] Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal Component Analysis." Chemometrics and Intelligent Laboratory Systems 2, no. 1–3 (August 1987): 37–52. doi:10.1016/0169-7439(87)80084-9.

[22] Oliveira, Jadson Jose Monteiro, and Robson Leonardo Ferreira Cordeiro. "Unsupervised Dimensionality Reduction for Very Large Datasets: Are We Going to the Right Direction?" Knowledge-Based Systems 196 (2020): 105777. doi:10.1016/j.knosys.2020.105777.

[23] Schölkopf, Bernhard, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. "Estimating the Support of a High-Dimensional Distribution." Neural Computation 13, no. 7 (2001): 1443–71. doi:10.1162/089976601750264965.

[24] McKinnon, Conor, James Carroll, Alasdair McDonald, Sofia Koukoura, David Infield, and Conaill Soraghan. "Comparison of New Anomaly Detection Technique for Wind Turbine Condition Monitoring Using Gearbox SCADA Data." Energies 13, no. 19 (2020). doi:10.3390/en13195152.

[25] Liu, Fei Tony, Kai Ming Ting, and Zhi Hua Zhou. "Isolation Forest." In Proceedings - IEEE International Conference on Data Mining, ICDM, 413–22. Data Mining, ICDM, (2008). doi:10.1109/ICDM.2008.17.

[26] Reynolds, Douglas. Gaussian Mixture Models BT - Encyclopedia of Biometrics. Edited by S Z Li and A Jain. Boston, MA: Springer US, 2009. https://doi.org/10.1007/978-0-387-73003-5_196.

[27] Goix, Nicolas. "How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?," (2016).

[28] Lavin, Alexander, and Subutai Ahmad. "Evaluating Real-Time Anomaly Detection Algorithms - The Numenta Anomaly Benchmark." Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015, 2016, 38–44. doi:10.1109/ICMLA.2015.141.