Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Detecting abnormal electricity usage using unsupervised learning model in unlabeled data

M. Z. H. Jesmeen [1], G. Ramana Murthy [2, *], J. Hossen [1], Jaya Ganesan [3], A. Abd Aziz [1], K. Tawsif [1]

[1] *Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia*
[2] *Department of Electronics and Communication Engineering, Vignan's Foundation for Science Technology and Research, Vadlamudi, India*
[3] *Faculty of Business, Multimedia University, Melaka, Malaysia*

ARTICLE INFO

Article history:
Received 26 February 2021
Received in revised form
23 May 2021
Accepted 24 June 2021

Keywords:
Electricity usage
Unsupervised learning
Excess mass
Machine learning
iForest
Gaussian mixture model

ABSTRACT

Smart-home systems achieved great popularity in the last decade as they increase the comfort and quality of life. Reduction of energy consumption became a very important desiderate in the context of the explosive technological development of modern society with a major impact on the future development of mankind. Moreover, due to the large amount of data available from smart meters installed in households. It makes leverage to able to find data abnormalities for better monitoring and forecasting. Detecting data anomalies helps in making a better decision for reducing energy usage wasted. In recent years, machine learning models are widely used for developing intelligent systems. Currently, researchers' main focus is on developing supervised learning models for predicting anomalies. However, there are challenges to train models with unlabeled data indicating data anomaly or not. In this paper, abnormalities are detected in electricity usage using unsupervised learning and evaluated using Excess Mass. The unsupervised anomaly detection model is based on Gaussian Mixture Model (GMM) and Isolation Forest (iForest). The models are compared with Local Outlier Factor (LOF) and One-class support vector machine (OCSVM). The proposed framework is tested with actual electricity usage and temperature data obtained from Numenta Anomaly Benchmark (NAB), which contains normal and anomaly data in time series. Finally, it has been observed that the iForest out-performed as the detection model for the selected use case. The outcome showed that the iForest can quickly detect anomalies in electricity usage data with only a sequence of data without feature extraction. The proposed model is suitable for the Smart Home Energy Management System's practical requirement and can be implemented in various houses independently. The proposed system can also be extended with the various use cases having similar data types.

1. Introduction

With a combination of The Internet of Things (IoT) and Smart Home Energy Management Systems (SHEMS), it becomes a scorching topic nowadays for a better process of extracting valuable knowledge. It helps for better management and visualization of electricity usage. In India, it was stated by Sial et al. (2019) that over 65% of electricity is generated from non-renewable assets, i.e., thermal power plant fuel, such as coal. In 2018, 26.9% of greenhouse gas emissions were raised due to electricity, and the residential and commercial sector distributes 32% of this emission (EPA, 2018). These gases have far-ranging environmental and health effects; if individuals get alert to abnormal electricity usage, they can make a few small changes to save electricity. As MEC (2014) said, it does not require significant physical changes to have electricity and contribute to a greener environment. Moreover, it was reported in Malaysia and in different places in the world; during the Movement Control Order (MCO) and Conditional Movement Control Order (CMCO) period in COVID-19, electricity consumption in the household increased. If consumers can track this consumption, they will be able to take action

* Corresponding Author.
Email Address: drgrm_ece@vignan.ac.in (G. R. Murthy)
https://doi.org/10.21833/ijaas.2021.09.014
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-3556-9294
2313-626X/© 2021 The Authors. Published by IASE.

accordingly. The government (MEC, 2019) indicates the importance of saving electricity and reduce wastage by the monitoring system.

The anomaly detection approaches are grouped into three methods, i.e., supervised, semi-supervised, and unsupervised (Ayodele, 2010). Where supervised and semi-supervised methods require labeled datasets indicating anomaly or not. Simultaneously, the unsupervised method will work for the unlabelled dataset. Here, the main objective is to detect abnormal usage, which is not labeled and different for different types of consumers. Hence, the task is to train the unsupervised machine learning (ML) model, i.e., Gaussian Mixture Model (GMM) and Isolation Forest (iForest), for anomaly detection in sequence batch Active power data. The algorithm is designed for batch learning without labeled data. For the result analysis in this paper, the data was obtained from Malaysia's telecom company. By pointing out the active power value as two categories: normal and abnormal usage. The results are finally compared with other known algorithms to get the best model in this case.

The rest of the paper is organized as follows: Section II contains a few recent literature reviews related to anomaly detection, following by the methodology described in section III; section IV contains results and discussion. Finally, section V presents the conclusion and future work related to the research presented in this paper.

2. Literature review

As there is a demand for electricity in sociality, an energy supply crisis is also a significant bottleneck to an economy. Wastage of energy usage needs to be reduced. An individual has to wait for monthly billing to understand electricity consumption change. It became easy to visualize with the current resources of smart meters and SHEMS. However, it is still difficult for a normal user to point out abnormal usage. Hence, researchers have done few works related to this anomaly detection. In this section, a few very recent works are discussed

Wang and Ahn (2020) had worked with anomaly detection for real-time by classification-model using labeled data. They had integrated the support vector machine (SVM) algorithm, the k-nearest neighbors (kNN) model with the cross-entropy loss function for developing an anomaly detection process for finding the data correctness in the electrical load dataset.

Yip et al. (2018) used anomaly detection technology to evaluate energy usage behavior for identifying the outliers caused due to electricity frauds and faulty meters. They had used Linear Programming and trained using labeled datasets too. Similarly, another network fault prediction architecture was developed, but it also contains labeled data indicating which session fault (Emerson et al., 2020).

Moreover, Sial et al. (2019) used four different heuristic methods to indicate the anomalies obtained data from the smart meters installed in the IIIT-Delhi campus hostels. They had presented empirical evaluation to demonstrate the effectiveness of their system.

Another work (Saad and Sisworahardjo, 2017) had presented a contextual anomaly detection model for detecting irregular power usage using an unsupervised ML algorithm with temporal context obtained from meter data.

Hu et al. (2018) had proposed a system involving one feature selection for a multivariate dataset for time-series data and the next trained model using the OCSVM model. However, the author's main issue is the number of identified discords is usually more than the detected outliers in real-world situations.

3. Methodology

This section presents the presented technique for computing anomaly detection for electricity usage sequence data in this paper. An overall all system flow is illustrated in Fig. 1. Time-series data is preprocessed by reshaping to one column to train the models. The Excess Mass (EM) method is used for evaluating the outcome, which indicates the trained model's performance. The best model is stored for the selected house. The algorithm evaluations conclude the specific technique will give better outcomes and use as abnormal electricity usage detection. Finally, the best model is stored for future detection. The process is briefly explained in this section.

3.1. Data collection

Initially, data was collected from Malaysia's Telecom Company from the SHEMS using API for this research work. The data is the Active Power value obtained from smart meters installed in a house's residential area. The dataset contains hourly data from 2019-01-10 to 2020-02-06. The Active power value is plotted in Time-series in Fig. 2a. Here, assume that a univariate time series data, $D_{\text{Train}}=\{x_1, x_2, x_3,.., x_{L_{train}}\}$ and L_{train} is the length of sequence data D_{Train}. Further, to test the model, this system is tested using benchmark data used to evaluate an anomaly detection system (Lavin and Ahmad, 2015), known as Numenta Anomaly Benchmark (NAB). This dataset can confirm that the system works if the data value is evaluating, as it contains real streaming evaluated data from different domains and applications. The temperature value obtained from NAB data is plotted in Time-series in Fig. 2b.

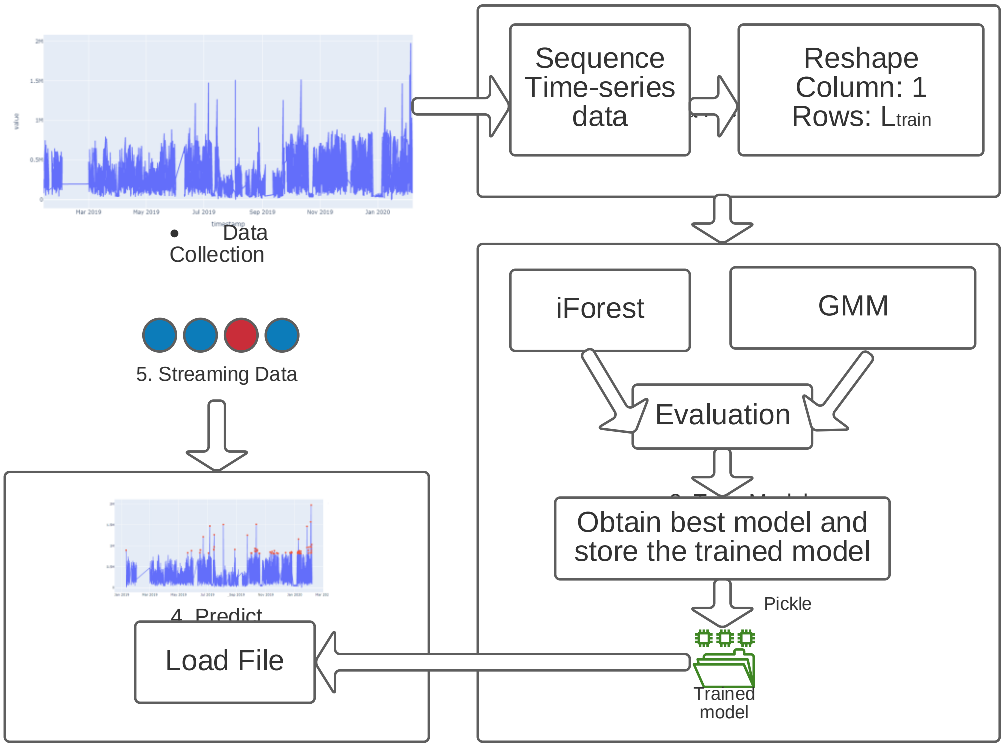
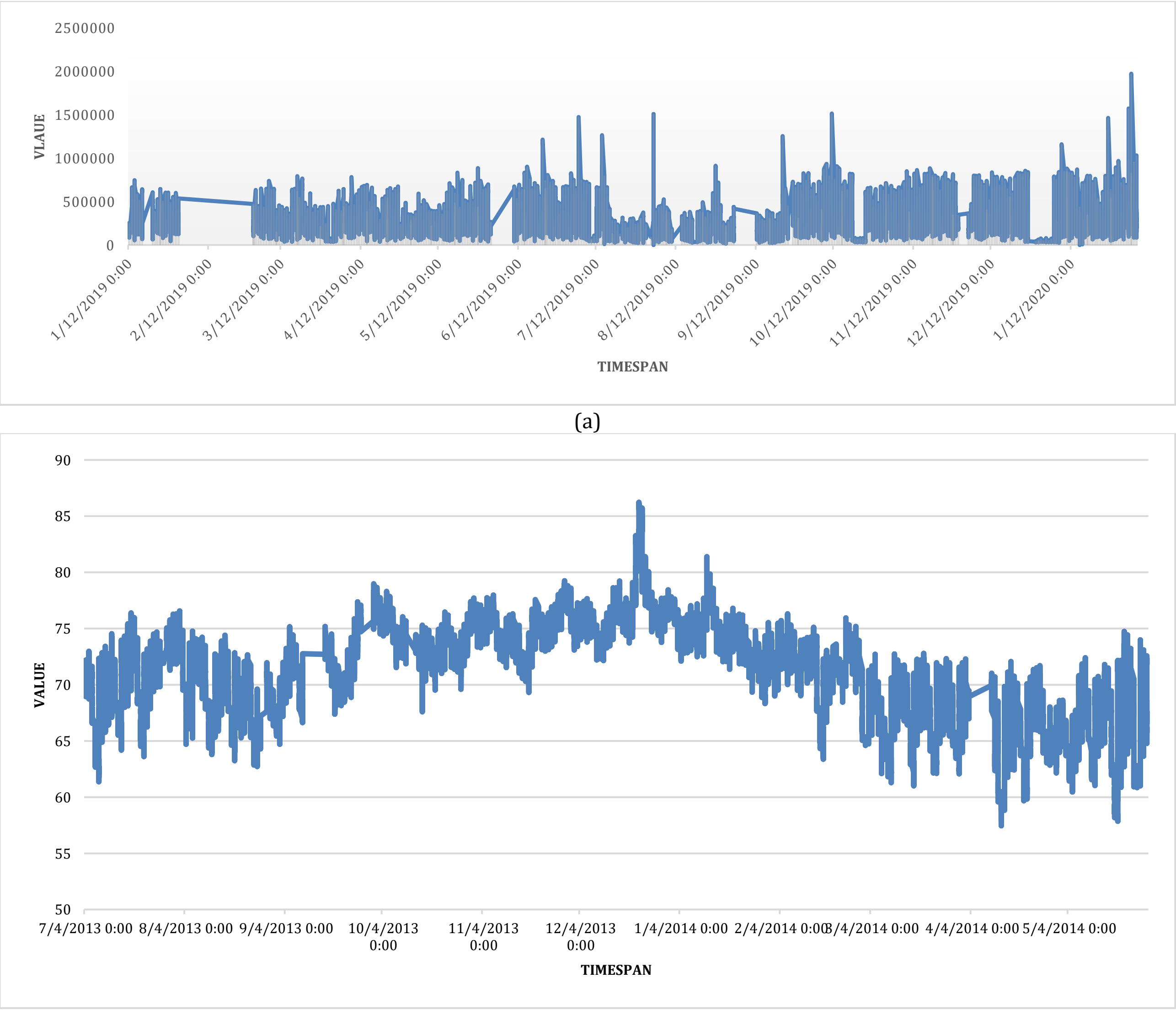

Data Collection
Sequence Time-series data
Reshape Column: 1 Rows: Ltrain
iForest
GMM
Evaluation
Obtain best model and store the trained model
Pickle
Trained model
5. Streaming Data
4. Predict
Load File

Fig. 1: The overall framework of time-series anomaly detection



(a)

(b)

Fig. 2: Visualization in time-series (a) Active power (b) Temperature

From the graph presented in Fig. 2. It can be understood that both data set is in a different form. But both are evaluated data. In Fig. 2a, there is a slight increase in electricity usage from March 2019 to January 2020. In Fig. 2b, data value increases from September 2013 to January 2014 and decreases from January 2014 to May 2014. For better visualization of the dataset D_{Train} selected, the total data value is plotted in the histogram, as shown in Fig. 3. The plot indicates the active power value and its frequency of use in the selected dataset.

3.2. Unsupervised machine learning model

The provided dataset from the SHEMS API does not contain any labeled data indicating anomaly or not. Hence the model is trained so it will group data according to the data. A probabilistic model Gaussian Mixture Modelling (GMM) and a tree-based model Isolation Forest (iForest) (Garcia-Font et al., 2018) are developed. These models are also compared with other available models (i.e., Local Outlier Factor (LOF) and One-class support vector machine (OCSVM)). For training, these models total L_{train} = 7,739 and L_{train} = 7267 of D_{Train} was selected finally from electricity usage and temperature data, respectively.

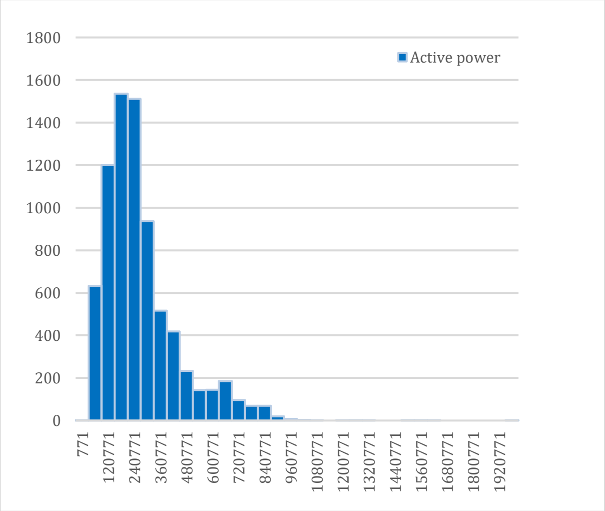


Fig. 3: Histogram plotted for active power

3.2.1. Gaussian mixture modeling

GMM model is used for obtaining a trained unsupervised ML model. A GMM works by using a parametric probability density function; it represents a weighted sum of Gaussian component densities (Reynolds, 2009). 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 unlabelled cluster data. It does account for variance type of data. Mathematically, Univariate Case (One-dimensional) Gaussian model's probabilities can be calculated using Eq. 1, followed by N value calculation using Eq. 2.

$$p(x) = \sum_{n=1}^{M} \emptyset_n N(x|\mu_n, \sigma_n) \tag{1}$$

$$N(x|\mu_n, \sigma_n) = \frac{1}{\sigma_i\sqrt{2\pi}}\, exp\left(-\frac{(x-\mu_n)^2}{2\sigma_n^2}\right) \tag{2}$$

where, μ_m indicates means value and σ_m is indicates calculation of variance for mth component. Moreover, $\emptyset_n$ is the value of weight for clustering 'm'.

For summarizing, in this approach, 'm' Gaussians to the data is fitted. Then finds the Gaussian distribution parameters μ_m and σ_m for each cluster and the weight of a group. Finally, for each data plot, probabilities are calculated, which belong to each collection.

3.2.2. Isolation forest

Isolation Forest (Liu et al., 2008) algorithm can explicitly point out abnormal data instead of first grouping the normal data. The concept is the same as the fundamental processes of decision tree algorithms. The tree partitions are generated randomly using the selected feature and then split randomly by the selected feature's minimum and maximum value. Next, an anomaly score is calculated to make decisions by using Eq. 3.

$$s(x,n) = 2^{\frac{-E(h(x))}{c(n)}} \tag{3}$$

where h(x) is the path length of usage value x, and n is the size of the selected set. Moreover, c(n) is stated by using Eq. 4.

$$c(n) = \begin{cases} 2H(n-1) - \frac{2(n-1)}{L_{train}} & for\; n > 2 \\ 1 & for\;\; n = 2 \\ 0 & otherwise \end{cases} \tag{4}$$

here, L_{train} is the size of the training dataset. H is the harmonic value. This can be calculated using Eq. 5, where γ is 0.5772156649 (known as Euler-Mascheroni constant).

$$H(k) = \ln(k) + \gamma \tag{5}$$

3.3. Excess mass (EM)

EM method is used here to evaluate the trained model. EM is known well to evaluate unsupervised anomaly detection algorithms (Goix, 2016). EM is based on the notion of density contour clusters. In section IV, the EM score is presented for evaluation. The higher the score indicates a well-trained model.

3.4. Tools

For developing the proposed system, Python Language is used. Hence, Python 3.0 environment was installed. It also used a few python libraries for different other tasks. Such as 'Numpy' and 'Pandas' libraries for data structures and data processing tools. 'Matplotlib' library was used for data visualization. Finally, the 'sklearn' library was integrated for the Gaussian Mixture model implementation.

4. Result and discussion

4.1. Evaluation using energy consumption data

The reason is to select GMM for clustering because GMM can handle the variate type dataset, as shown in Fig. 4, and the calculated probability is plotted.

For evaluating and cross-checking the EM score, the model is trained using different subsets and tabulated in Table 1 for GMM, here L_{train} is 2,361. It can be clear that the EM score is excellent and close to each other. It has also been clear while detecting the first half and second half data; it plots 12 anomalies. While detecting abnormalities in the complete dataset, it detects 24 anomalies. Hence, it shows the anomaly detection is possibly correct.

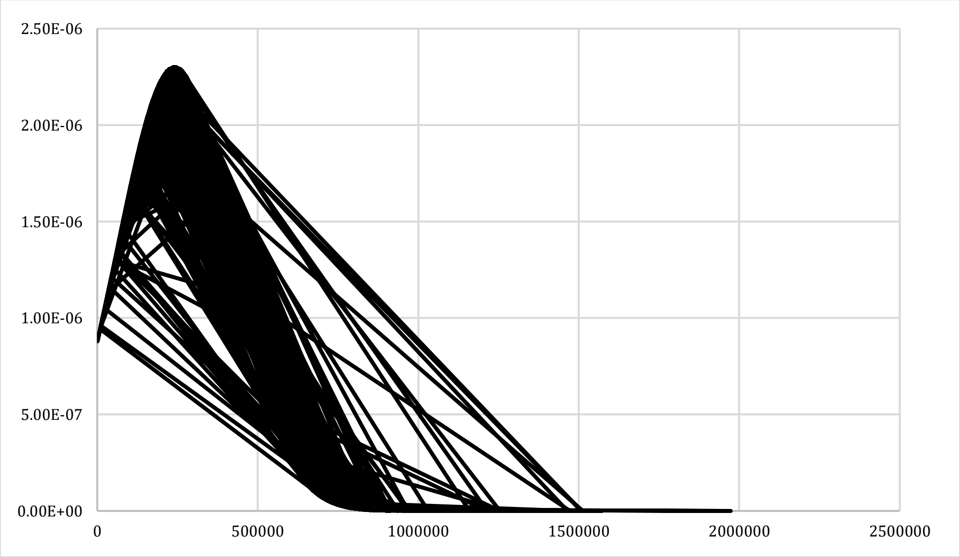


Fig. 4: Visualization of 1D GMM of Active power 'x' vs. probability 'p(x)'

Table 1: EM score and number of anomalies detected

Data	Details	Outcome		
		EM score	Anomaly	Not Anomaly
678	Nov 2019	0.008265	669	9
596	Dec 2029	0.009061	590	6
752	Jan 2020	0.006497	744	8
1,168	First half	0.007865	1156	12
1,192	Last half	0.008469	1180	12
2,361	All Data	0.008362	2337	24

For further evaluation, each model's EM score is calculated after training using the decision function to predict confidence scores, presented in Table 2. The iForest and GMM tend to out-perform the detection performance.

Table 2: EM score of trained anomaly detected

#	Model Name	EM Value
1	One Class SVM	0.012944
2	Isolation Forest	**0.017387**
3	LOF	0.011177
4	GMM	**0.015516**

According to evaluation using EM score, all data is clustered using OCSVM, iForest, LOF, and GMM, and plotted anomalies, as shown in Fig. 5. Here, it is clear iForest and GMM detection of abnormalities plotted is almost the same. Other than the excellent EM score of iForest and GMM, they both tend to detect similar anomalies.

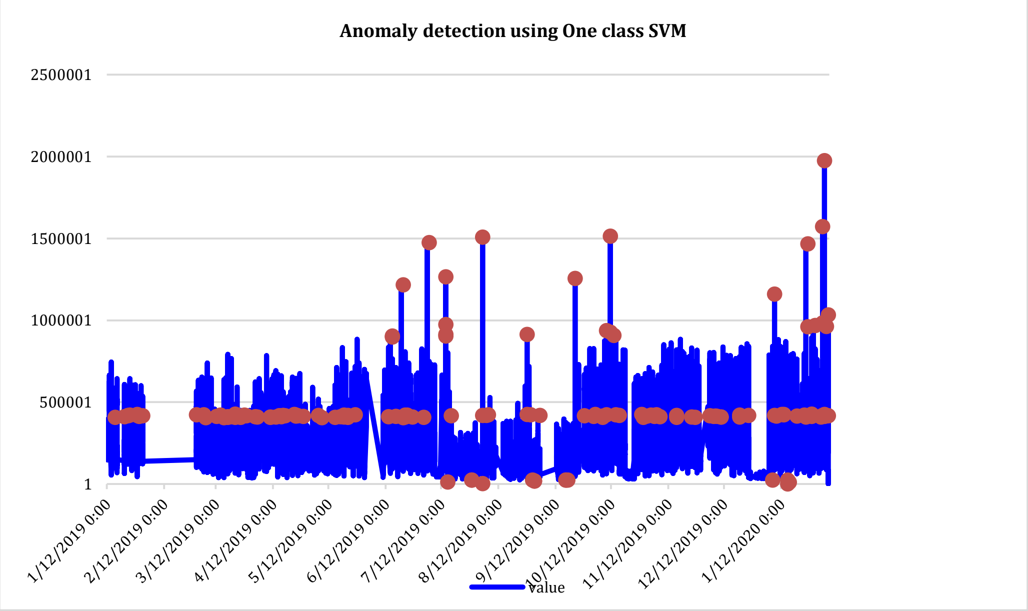
Moreover, in Fig. 6, the histogram indicates normal usage and abnormal usage. These plots will present the user with a clear visualization of abnormal energy usage on the specific active power. It demonstrates that 0.75M to 1.0M active power is an anomaly usage of the selected house. This detection will be different for different homes. Hence, the model train is more reliable and not dependent on the specific household. The detection is not statistically based, and each model must perform independently according to household usage.

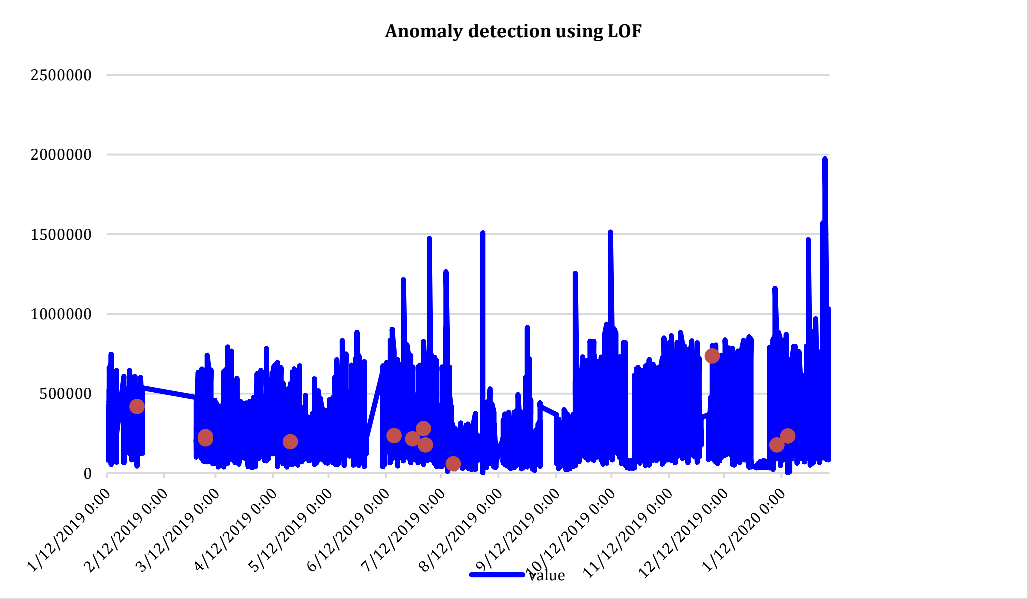
The number of anomalies detected is not similar in the four methods. However, detecting abnormal usage for iForest and GMM is almost the same; even the detection number is also the same. The number of anomalies detected is presented in Table 3 for each trained model. Here, as stated by Hu et al. (2018), it is also proven that the number of anomalies detected is more than usual.

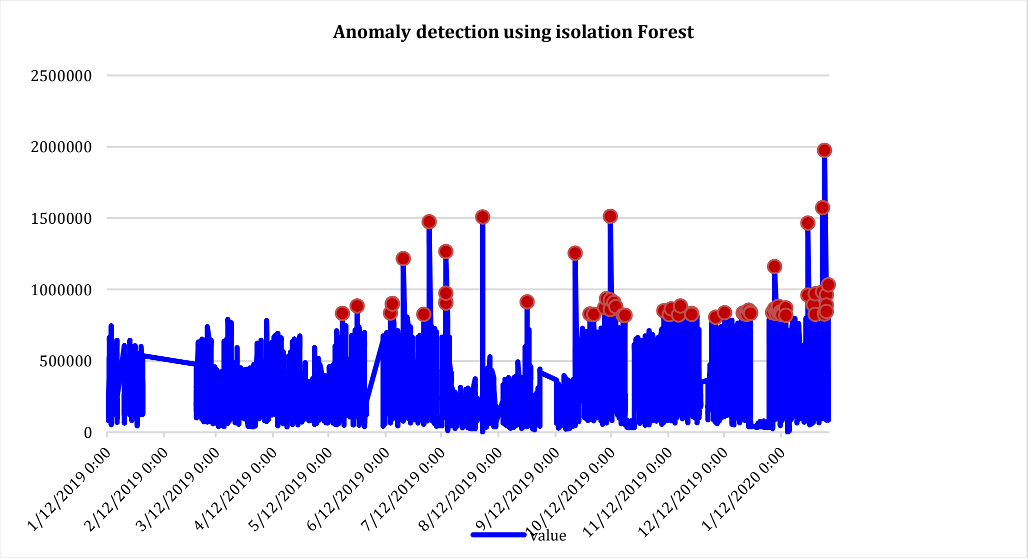
Table 3: EM score and number of anomalies detected

	Anomaly Detected	Normal usage
OCSVM	90	7649
iForest	77	7662
LOF	11	7728
GMM	78	7661

Finally, can conclude that iForest and GMM can be used for abnormal unlabeled electricity usage detection process. Hybrid of iForest and GMM can help to find the exact anomaly point. According to Fig. 5, iForest and GMM output clearly indicate anomalies in high active power.


Anomaly detection using One class SVM
2500001
2000001
1500001
1000001
500001
1
1/12/2019 0:00
2/12/2019 0:00
3/12/2019 0:00
4/12/2019 0:00
5/12/2019 0:00
6/12/2019 0:00
7/12/2019 0:00
8/12/2019 0:00
9/12/2019 0:00
10/12/2019 0:00
11/12/2019 0:00
12/12/2019 0:00
1/12/2020 0:00
value


Anomaly detection using LOF
2500000
2000000
1500000
1000000
500000
0
1/12/2019 0:00
2/12/2019 0:00
3/12/2019 0:00
4/12/2019 0:00
5/12/2019 0:00
6/12/2019 0:00
7/12/2019 0:00
8/12/2019 0:00
9/12/2019 0:00
10/12/2019 0:00
11/12/2019 0:00
12/12/2019 0:00
1/12/2020 0:00
value


Anomaly detection using isolation Forest
2500000
2000000
1500000
1000000
500000
0
1/12/2019 0:00
2/12/2019 0:00
3/12/2019 0:00
4/12/2019 0:00
5/12/2019 0:00
6/12/2019 0:00
7/12/2019 0:00
8/12/2019 0:00
9/12/2019 0:00
10/12/2019 0:00
11/12/2019 0:00
12/12/2019 0:00
1/12/2020 0:00
value

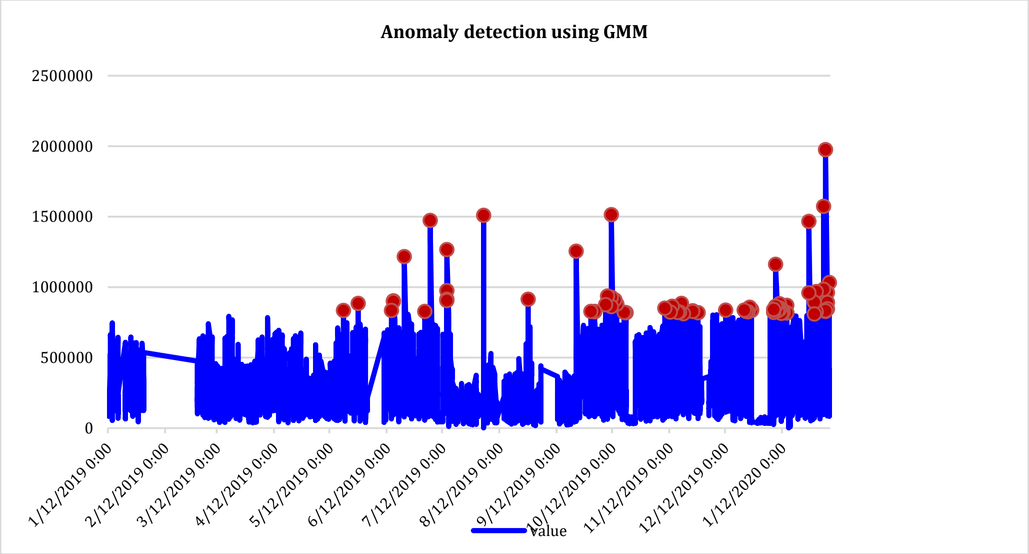
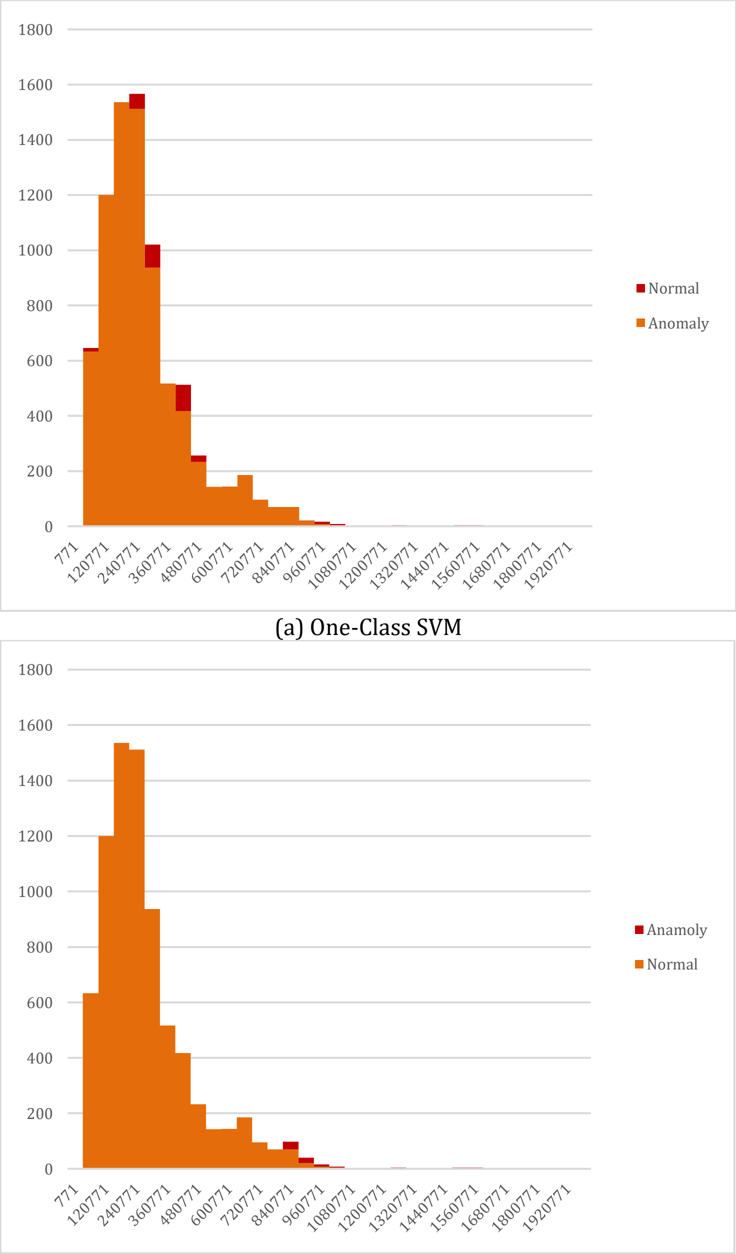
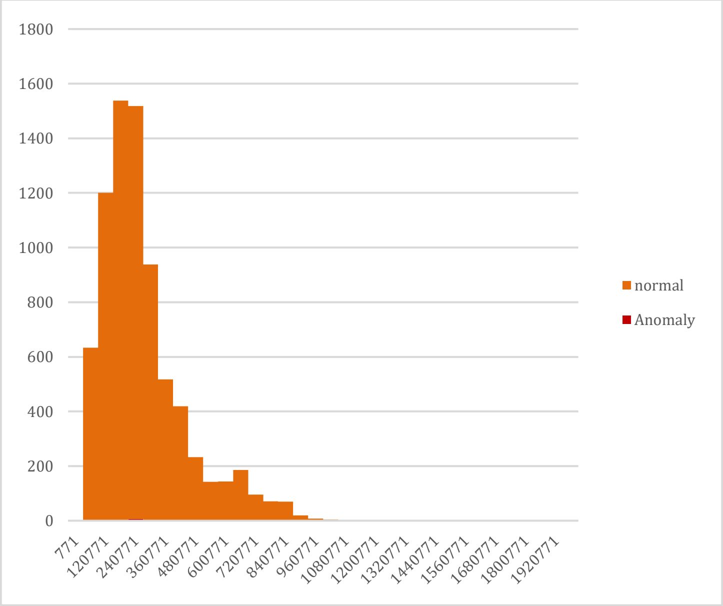


Fig. 5: Time-series of active power with anomaly markers

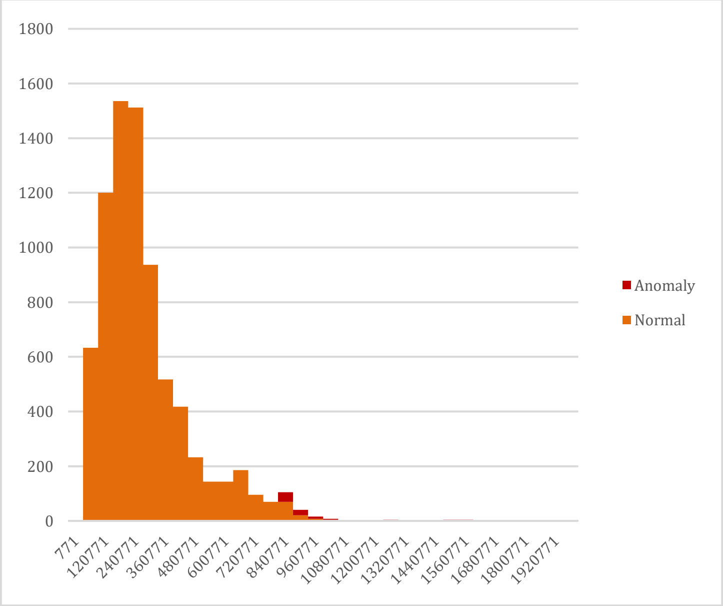


(a) One-Class SVM

(b) Isolation forest



(c) LOF



(d) GMM

Fig. 6: Histogram plotted for active power indicating normal and abnormal data

4.2. Evaluation using benchmark data

A known dataset NAB is used to train each model (i.e., One-Class SVM, iForest, LOF, and GMM) and then obtained EM score by using decision function by predicting confidence scores. Here, the outcome obtained from temperature sensors data is used, and the EM score is presented in tablature form in Table 4. Here also like energy consumption data, iForest and GMM had out-perform the detection of anomaly process. The NAB dataset can confirm that the selected algorithm can detect anomaly data, as it contains clean data and anomaly data.

Table 4: EM score of trained anomaly detected for nab dataset

#	Model Name	EM Value
1	One Class SVM	0.005843
2	Isolation Forest	**0.010054**
3	LOF	0.008150
4	GMM	**0.009573**

Like detecting anomalies in electricity usage, the second database also detected similar anomalies using iForest and GMM algorithms. The time series plot with anomaly detection for the database is presented in Fig. 7.

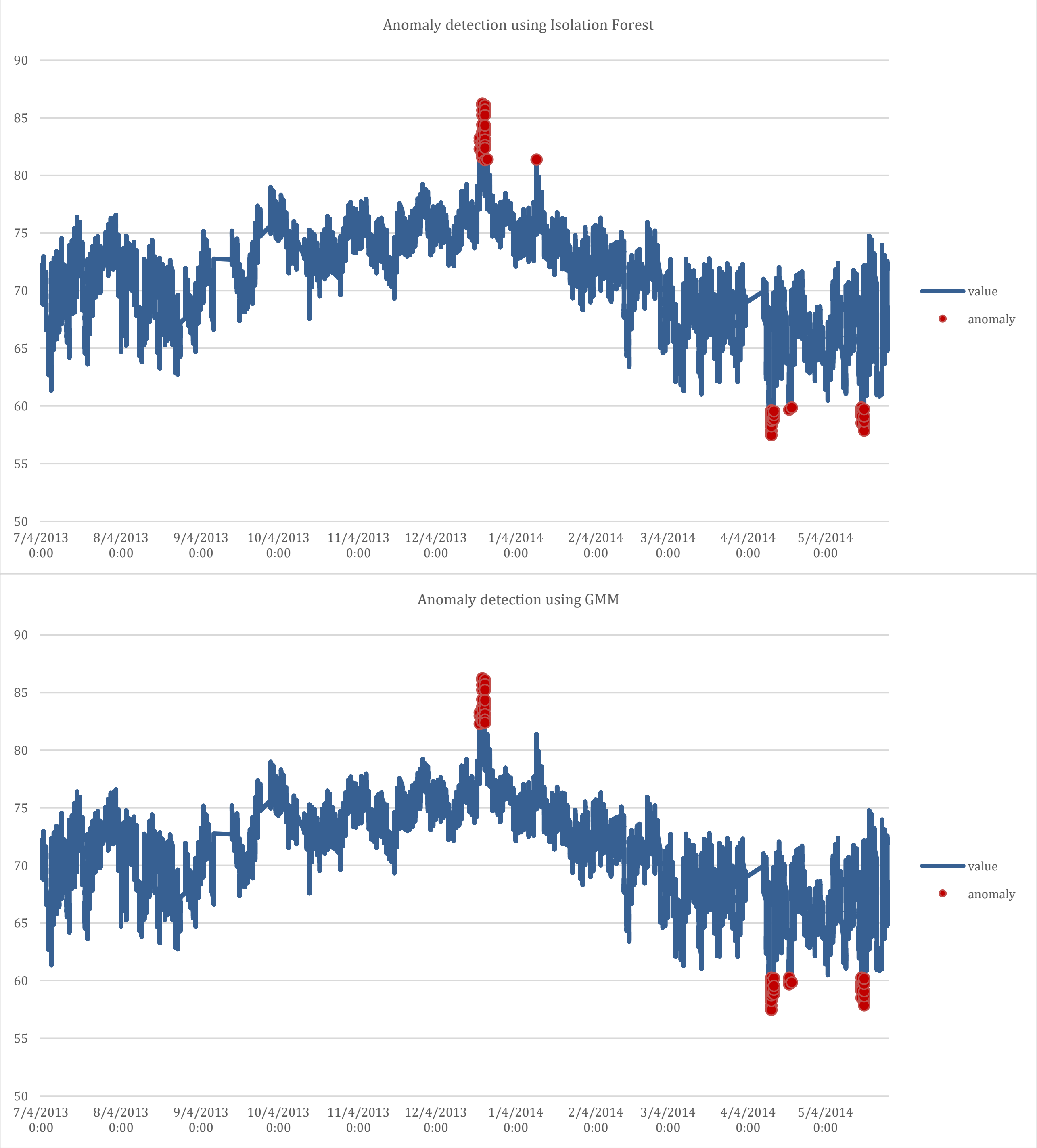


Fig. 7: Time-series of temperature with anomaly markers for iForest and GMM

5. Conclusion

In this paper, a system is presented and evaluated to study abnormal data in sequence active power data. It is the best way to detect unusual electricity consumption. It was obtained the first iForest and next GMM trained with good EM score for detecting anomalies. To conclude, this strategy of detecting abnormal usage could eventually benefit any individual using a smart home energy monitoring system. The beneficiation is from saving electricity by reducing energy wastage and cost. Correctly implementing the system's approach will also play a significant role in the smart home market. The study had proved that the unsupervised technique allowed the consumers to show that these models will bring a great deal for smart homes. It can also be considered for unusual energy usage or energy theft detection or for other similar evaluating datasets.

In the future, more investigation will be processed by using other unsupervised ML models (such as DBSCAN (Sheridan et al., 2020)). The evaluation was currently processed considering data outlier; next can consider novelty detection. As mentioned by Carreño et al. (2019), there is a difference between outlier and novelty. Next, on data visualization, it can be seen there are few missing data in the time series. It is also essential to handle this missing value (Jesmeen et al., 2018; 2019). Moreover, handling these missing data might enhance the EM score. In the case of energy consumption anomaly detection, it is also required to detect anomalies in seasonal-based and discriminate actual anomaly and seasonal anomaly.

Funding

This work was supported in part by Telekom Malaysia under Grant TMRND [MMUE/190007].

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Ayodele T (2010). Types of machine learning algorithms. In: Zhang Y (Ed.), New advances in machine learning: 19-48. BoD–Books on Demand, Norderstedt, Germany.

Carreño A, Inza I, and Lozano JA (2019). Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. Artificial Intelligence Review, 53: 3575–3594. https://doi.org/10.1007/s10462-019-09771-y

Emerson RJ, Hossen J, Ervina E, Tawsif K, and Jesmeen M (2020). Broadband network fault prediction using complex event processing and predictive analytics techniques. Journal of Engineering Science and Technology, 15(4): 2289-2300.

EPA (2018). Sources of greenhouse gas emissions. The United States Environmental Protection Agency, Washington, USA.

Garcia-Font V, Garrigues C, and Rifà-Pous H (2018). Difficulties and challenges of anomaly detection in smart cities: A laboratory analysis. Sensors, 18(10): 3198. https://doi.org/10.3390/s18103198

Goix N (2016). How to evaluate the quality of unsupervised anomaly detection algorithms? Available online at: https://arxiv.org/abs/1607.01152

Hu M, Ji Z, Yan K, Guo Y, Feng X, Gong J, and Dong L (2018). Detecting anomalies in time series data via a meta-feature based approach. IEEE Access, 6: 27760-27776. https://doi.org/10.1109/ACCESS.2018.2840086

Jesmeen MZH, Hossen A, Hossen J, Raja JE, Thangavel B, and Sayeed S (2019). AUTO-CDD: Automatic cleaning dirty data using machine learning techniques. Telkomnika, 17(4): 2076-2086. https://doi.org/10.12928/telkomnika.v17i4.12780

Jesmeen MZH, Hossen J, Sayeed S, Ho CK, Tawsif K, Rahman A, and Arif EMH (2018). A survey on cleaning dirty data using machine learning paradigm for big data analytics. Indonesian Journal of Electrical Engineering and Computer Science, 10(3): 1234-1243. https://doi.org/10.11591/ijeecs.v10.i3.pp1234-1243

Lavin A and Ahmad S (2015). Evaluating real-time anomaly detection algorithms--The Numenta anomaly benchmark. In the IEEE 14th International Conference on Machine Learning and Applications, IEEE, Miami, USA: 38-44. https://doi.org/10.1109/ICMLA.2015.141

Liu FT, Ting KM, and Zhou ZH (2008). Isolation forest. In the 8th IEEE International Conference on Data Mining, IEEE, Pisa, Italy: 413-422. https://doi.org/10.1109/ICDM.2008.17

MEC (2014). Towards a world-class energy sector: Energy commission. Volume 3, Malaysia Energy Commission, Putrajaya, Malaysia.

MEC (2019). Shaping the future of Malaysia's energy sector. Volume 18, Malaysia Energy Commission, Putrajaya, Malaysia.

Reynolds D (2009). Gaussian mixture models. In: Li SZ and Jain A (Eds.), Encyclopedia of biometrics: 659–663. Springer, Boston, USA. https://doi.org/10.1007/978-0-387-73003-5_196

Saad A and Sisworahardjo N (2017). Data analytics-based anomaly detection in smart distribution network. In the International Conference on High Voltage Engineering and Power Systems, IEEE, Denpasar, Indonesia: 1-5. https://doi.org/10.1109/ICHVEPS.2017.8225855

Sheridan K, Puranik TG, Mangortey E, Pinon-Fischer OJ, Kirby M, and Mavris DN (2020). An application of dbscan clustering for flight anomaly detection during the approach phase. In the AIAA SciTech 2020 Forum, Orlando, USA: 1851. https://doi.org/10.2514/6.2020-1851

Sial A, Singh A, and Mahanti A (2019). Detecting anomalous energy consumption using contextual analysis of smart meter data. Wireless Networks: 1-18. https://doi.org/10.1007/s11276-019-02074-8

Wang X and Ahn SH (2020). Real-time prediction and anomaly detection of electrical load in a residential community. Applied Energy, 259: 114145. https://doi.org/10.1016/j.apenergy.2019.114145

Yip SC, Tan WN, Tan C, Gan MT, and Wong K (2018). An anomaly detection framework for identifying energy theft and defective meters in smart grids. International Journal of Electrical Power and Energy Systems, 101: 189-203. https://doi.org/10.1016/j.ijepes.2018.03.025