Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1650-1661
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

A Tunable Electricity Theft Detection Method Based on Usage Habits of Customers

Chunjiang Yan[1], Feng Ma[1], Weigang Nie[1], Xiaokun Han[1], Yuejie Xu[2], Yanlin Peng[2*]
[1]State Grid Beijing Electric Power Company, Xicheng District, Beijing, China
[2]School of Electrical and Electronic Engineering, North China Electric Power University, Changping District, Beijing, China
*Corresponding Author.

Abstract:

With the wide application of the Advanced Measurement Infrastructure in power grid, electricity theft detection methods based on data analysis become a main stream for diagnosis of customers with electricity theft behavior. However, the existing anomaly submergence problem may affect the accuracy of electricity theft detection. In addition, the threshold selection of the existing unsupervised learning algorithm is relatively fixed and cannot adapt to changing detection scenarios. To solve this problem, this paper proposes an electricity theft detection method with a variable threshold. Based on the user's load shape dictionary obtained by weighted clustering, the cosine distance between the user's load and the load shape dictionary is used as the standard of the user's consumption anomaly degrees, and the effectiveness and applicability of the proposed method are verified by numerical experiments.

Keywords*: Electricity theft detection, K-means, Load shape dictionary, Data mining.*

I. INTRODUCTION

Transmission loss in power grid contains technical loss (TL) and non-technical loss (NTL) [1]. The TL is the normal losses in the process of power transmission, such as the copper and core losses of transformers. The NTL is the remaining losses that cannot be explained by theory, such as electricity theft. Besides, electricity theft can seriously damage the economic benefits of power utilities and lay down potential safety hazards such as power outages, equipment damage, and casualties. With the application of smart meters and the establishment of advanced metering infrastructure (AMI), a large amount of electricity consumption data makes data mining technology more suitable for electricity theft detection. However, the software and communication technology used in AMI make it possible to tamper with smart meters and intrude into the information flow of the power grid through cyberattacks [2].

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1650-1661
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

Current data-driven electricity theft detection methods (ETDMs) can be divided into three categories according to the type of data they use [3]. Methods in the first category assume that granular power consumption data is available and consumption patterns of fraudulent users differ from those of benign users, and this type of ETDM utilizes logistic regression [4] or artificial intelligence such as classification [5-7] and clustering [8] to analyze the load profiles of customers for electricity theft detection. Specifically, supervised methods like classification usually involve vast labeled historical electricity usage data to train the detection models. Examples including support vector machines (SVM) [5], convolutional neural networks (CNNs) [6] and other artificial neural networks [7] have been tested in literature. In contrast, unsupervised methods like clustering, focus on the information without labels. They usually extract the load shape dictionary (LSD) from the load profiles of users and calculate the anomaly degrees by quantifying the difference between the load profiles and LSD.

The existing data-driven ETDMs have some limitations. First, the supervised methods need vast reliable theft samples to train the detection models. But the small proportion of theft users and the data poisoning (the false labeled samples) [9] limited their accuracies. Worse yet, they might not distinguish between electricity theft and non-malicious activities like meter reinstallation [5]. Second, the unsupervised methods cannot assure that the power patterns of fraudulent users are always deviates from the normal LSD considering the fact of anomaly submergence [10] which means that theft pattern of one user might be similar with normal pattern of another user since the huge differences of usage habits among different users. Finally, the thresholds of unsupervised methods are usually fixed thus are not flexible enough to handle the detection for different scenes and users. The main contributions of this study are as follows.

1) We propose a weighted LSD extraction method for single individual based on *K*-means to ease the negative impact of anomaly submergence.

2) By calculating the cosine distance between load curve and LSD as the anomaly degree, we observe the distribution differences of anomaly degrees of normal curves and abnormal ones.

3) Based on above distribution differences, we develop a threshold tunable electricity theft detection method and corresponding strategy for threshold adjustment. The effectiveness and applicability of the proposed method are verified by numerical experiments.

This paper is organized as follows. In Section I, we review existing data-driven ETDMs in literature. In Section II, a method of extracting user's LSD is proposed. Section III presents a

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1650-1661
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

tunable electricity theft detection method based on user's LSD. Numerical experiments are conducted and the results are shown in Section IV. Finally, we conclude this paper in Section V.

II. EXTRACTION OF TYPICAL DAY LOAD CURVE

2.1 Common Monthly LSD Extraction Methods for Single User

From the perspective of simplicity and practicability, it can choose the load profile of a certain day in one month as the monthly LSD for a single user. For example, the load profile of a typical working day or the load profile of monthly maximum load day. However, this method is too simple to contain the information about the usage habits in other time periods. Another method to extract the monthly LSD in practice is to get the average load profile in one month. This method considers the usage habits in different time periods, but the procedure to get the average load profile may distort the load shape thus need further improvement.

2.2 LSD Extraction Method based on Weighted *K*-means

This paper presents a monthly LSD extraction method based on weighted *K*-means, the methodology is as follows.

Let us denote the set of load profiles of a user in one year as $\boldsymbol{X}$, i.e.

$$\mathbf{X}=\left\{\mathbf{x}_1,\mathbf{x}_2,\cdots\mathbf{x}_i,\cdots,\mathbf{x}_{365}\right\} \tag{1}$$

Where, $\mathbf{x}_i$ is the load profile of *i*-th day. Then, let us reconstruct $\boldsymbol{X}$ according to the month and date of the load profile as (2)

$$\mathbf{X}=\left\{\mathbf{x}_{m,d}\right\}\Big|_{d\in\{1,2,3,\cdots,31\}}^{m\in\{1,2,3,\cdots,12\}} \tag{2}$$

Where $\mathbf{x}_{m,d}$ is the load profile of the *d*-th day in month *m*. For example, $\mathbf{x}_{1,4}$ is the load profile of Jan. 4th. Then, the *K*-means is utilized to analyze $\boldsymbol{X}$ and to get the set of cluster centers $\boldsymbol{Y}$, i.e.:

$$\mathbf{Y}=\left\{\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_k,\cdots,\mathbf{y}_K\right\} \tag{3}$$

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1650-1661
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

Where the center of *k*-th cluster. After that, we can count number of the load profiles in month *m* belonging to *k*-th cluster. And let us denote this number as $D_{m,k}$. Then, we can calculate the weight factor that cluster center $\mathbf{y}_k$ account for the load profiles in month *m* via dividing $D_{m,k}$ by the number of total days in month *m*, i.e.

$$\omega_{m,k} = \frac{D_{m,k}}{D_m} \tag{4}$$

Where D_m is the number of days in month *m*. Finally, the LSD of month *m* can be got by (5):

$$\hat{\mathbf{x}}_m = \sum_{k=1}^{K} \mathbf{y}_k \omega_{m,k} \tag{5}$$

Where $\hat{\mathbf{x}}_m$ is the LSD of month *m*. The LSD of this year is composed of $\hat{\mathbf{x}}_m$ of 12 months, i.e.

$$\hat{\mathbf{X}} = \left\{ \hat{\mathbf{x}}_m \right\} \Big|_{m \in \{1,2,\cdots,12\}} \tag{6}$$

It is apparently that the LSD of month *m* is the linear combination of cluster centers and the combination coefficient is the weight factor $\omega_{m,k}$.

III. DETECTION FOR ELECTRICITY THIEVES

3.1 Principles of Electricity Thieft Detection

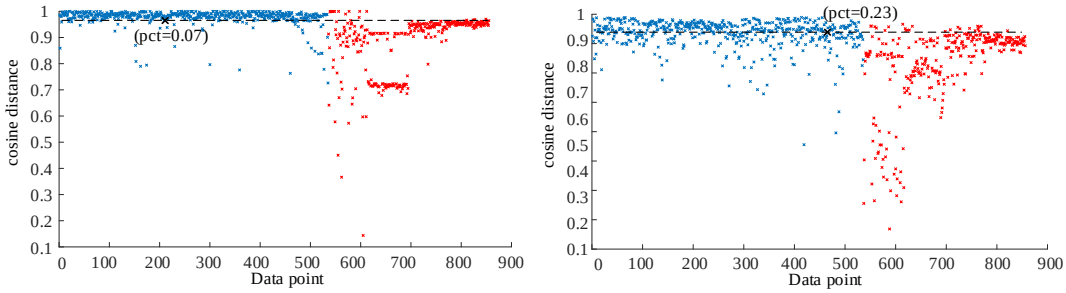
This paper calculates the cosine distance (CD) between the load profile and the LSD as the its deviation degree from the LSD by (7)

$$R(\boldsymbol{x}_{m,d}, \ \hat{\boldsymbol{x}}_m) = \frac{\sum_{t=1}^{T} (x_{m,d,t} - \bar{\boldsymbol{x}}_{m,d})(\hat{x}_{m,,t} - \bar{\hat{\boldsymbol{x}}}_m)}{\sqrt{\sum_{t=1}^{T} (x_{m,d,t} - \bar{\boldsymbol{x}}_{m,d})^2} \sqrt{\sum_{t=1}^{T} (\hat{x}_{m,,t} - \bar{\hat{\boldsymbol{x}}}_m)^2}} \tag{7}$$

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1650-1661
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

Where $R(\boldsymbol{x}_{m,d},\ \hat{\boldsymbol{x}}_m)$ is the CD between $\boldsymbol{x}_{m,d}$ and $\hat{\boldsymbol{x}}_m$, and $R(\boldsymbol{x}_{m,d},\ \hat{\boldsymbol{x}}_m) \in [-1,1]$. A lower $R(\boldsymbol{x}_{m,d},\ \hat{\boldsymbol{x}}_m)$ indicates that the load vector $\boldsymbol{x}_{m,d}$ is more far away from the LSD $\hat{\boldsymbol{x}}_m$.

It is believed that the usage patterns of electricity thieves will deviate from those of normal users. However, this conclusion is rather invalid and need to be further studied. To validate this point, we calculate the CDs of the normal load vector and fraudulent load vector. Figure 1 show the scatter plot of the CDs of two customer. User A is a customer with rather fixed usage habit, while user B has more random usage habit. And the blue point is the normal sample and the red point is the fraudulent sample.



(a) User A (b) User B

Fig 1: Comparison of CD scatter plots of users' normal consumption and electricity theft.

It can be seen from Fig 1(a) that, the CDs of normal sample of user A distribute in a very narrow range near 1, which means that the normal usage habit of user A is fixed and the cluster number of user A's load profiles is low. While the CDs of fraudulent samples of user A distribute in a wide range and some of them are far from 1. Under this circumstance, the distribution of the CDs of normal samples and fraudulent ones has significant difference, which means that the fraudulent samples deviate from the normal ones. Thus, we can set the threshold to a higher value. On the other hand, Fig 1(b) shows that, if a user has random usage habit, the CDs of his normal samples distribute dispersedly. It means that certain number of normal samples are also deviated from the LSD, the threshold needs to be a lower value.

The examples of user A and user B reveal two significant clues:

1) We can distinguish fraudulent samples from normal ones via their values of CDs;

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1650-1661
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

2) To ensure detection effect, the threshold θ to judge the fraudulent samples is related to the usage habit of the user.

This paper uses percentage rate (PCT) to measure the threshold, which is calculated by (8)

$$\mathrm{PCT} = \frac{C_{outlier}}{C_{normal}} \tag{8}$$

Where the $C_{outlier}$ is the number of normal samples whose CDs is lower than the threshold; C_{normal} is the total number of normal samples. To measure the degree of randomness of a customer's usage habit, the standard deviation σ of the CDs of his normal samples is calculated. The threshold, PCT and the standard deviation σ these three factors are related to each other. Since complexity of modeling, this relation is unlikely to be formulated through theoretical analysis. We only get their relation in statistics by conducting numerous experiments.

3.2 The Relation between PCT and σ in Statistics

Suppose that there are *n* users. To analyze the relation between PCT and σ, we must get their PCT and σ. Take user *i* for example, the procedure to get PCT_i and σ_i is as follows:

1) For user *i*, its scatter plot needs to be drawn like user A and user B in Section 3.1;

2) Calculate the standard deviation σ_i of user *i*;

3) The *K*-means (*K*=2) is adopted to classy normal samples of user *i* into two cluster (the one is outlier, the other is non-outlier) according to their CDs;

4) The boundary of these two clusters is set as the threshold θ_i of user *i*;

5) According to θ_i, the PCT_i of user *i* can be calculated based on (8).

For each user, we can obtain a pair of PCT and σ. Finally a scatter plot of PCT-σ can be

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1650-1661
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

drawn. It can be concluded that, with ascending of σ, the value of PCT also increases. But their relation is neither linear nor monofonic. If we know the σ of a user, his PCT can be easily obtained with the help of the scatter plot of PCT-, and its threshold can be derived by (8.

3.3 The Detection Procedure

The detection procedure in practice can be divided into two steps: the one is the training process; the other is detection process. The training process need to know the basic information of the normal and fraudulent samples of users. And its goal is to get the scatter plot of PCT-σ. The detection process is to obtain the threshold θ of the user to be detected, which is as follows:

1) For the user *j* to be detected, its scatter plot needs to be drawn like Figure 1 in Section 3.1;

2) Calculate the standard deviation σ_j of user *j*, and the PCT_j can be obtained with the scatter plot of PCT-σ;

3) Ranking the load profiles of user *j* in the ascending order based on their CDs. The threshold θ_j is the CD of the load profile whose ranks at PCT_j.

After the θ_j is obtained, the fraudulent samples can be judged by comparing its CD with θ_j: if $\theta_j > R(\boldsymbol{x}_{m,d}, \hat{\boldsymbol{x}}_m)$, $\boldsymbol{x}_{m,d}$ is a fraudulent sample; otherwise, $\boldsymbol{x}_{m,d}$ is a normal sample.

IV. NUMERICAL EXPERIMENTS

4.1 Evaluation Index

The detection of electricity theft is a two-class model. The positive category is the user's electricity theft data, and the negative category is the user's normal electricity consumption data. The effect of the model is evaluated by a confusion matrix, as is shown in Table I.

TABLE I. Confusion matrix

	Actual Positive	Actual Negative

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1650-1661
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

In this paper, we use the following performance metrics to evaluate the results of detection: Accuracy (ACC), Precision (Pre), Recall (Rec) and False positive rate (FPR). Among them, ACC and FPR are main evaluation indexes. In electricity theft detection, when an electricity theft user is detected, the power supply company needs to assign employees to verify in field, and if FPR is too high, it can waste manpower and material resources of the power supply company. Therefore, the false positive rate is an important indicator in evaluating the detection model. The metrics are calculated as in (9)-(12).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{12}$$

4.2 Experiments No.1

The data source is the Irish Smart Energy Trial [11], which was released by Electric Ireland and Sustainable Energy Authority of Ireland (SEAI) in 2012. Since each user participated in the experiment voluntarily, each participator is considered to be normal consumption user. Some of the commercial users are selected for experiment, and the electricity theft data is modified from the normal electricity consumption data according to Table II [6]:

TABLE II. 6 typical ways of electricity theft

Type	**Theft method**
type1	$f_1(x_t) = \gamma x_t, 0.2 < \gamma < 0.8$

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1650-1661
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

type2	$f_2(x_t)=\begin{cases} x_t, & x_t \le \gamma \\ \gamma, & x_t > \gamma \end{cases}, \gamma \le \max(x)$
type3	$f_3(x_t)=\max\{x_t-\gamma,0\}, \gamma < \max(x)$
type4	$f_4(x_t)=\beta x_t, \beta=\begin{cases} 1, t_1 < t < t_2 \\ 0, \text{otherwise} \end{cases}$
type5	$f_5(x_t)=\alpha_t x_t, 0.2 < \alpha_t < 0.8$
type6	$f_6(x_t)=\alpha_t \bar{x}, 0.2 < \alpha_t < 0.8$

In experiment No.1, 200 groups of training users are utilized, and each group includes 535 days of normal data and 480 days of abnormal data (6 types of electricity theft for 80 days each). 80 users are utilized for electricity detection, and 300 days data of each user are assumed with normal line loss. The remaining 235 days of normal consumption and 240 days of abnormal electricity consumption (6 types of electricity theft for 40 days each) are utilized as the actual test samples, a total of 80 ×475=38000 data. The results of detection are shown in Table III.

TABLE III. Detection result of various types of electricity theft

	Accuracy	**Precision**	**Recall**	**FPR**
Type 1	0.501	0.260	0.039	0.075
Type 2	0.618	0.439	0.276	0.060
Type 3	0.845	0.917	0.728	0.066
Type 4	0.926	0.935	0.89	0.059
Type 5	0.853	0.929	0.746	0.053
Type 6	0.837	0.605	0.573	0.051

As is shown in Table III, we can conclude that the performance of the proposed method is different for the detection of 6 types of electricity theft. In the detection of type 1 and type 2, the ACC of the method proposed in this paper is 50.1% and 61.8%, and the false detection rate is 7.5% and 6%, respectively. The detection effect is poor and the accuracy rate is low. For theft type 3, the accuracy rate can reach 84.5%, and the false positive rate is 6.6%. Type 4 has the best effect, with an accuracy rate of 92.6% and a false positive rate of 5.9%. The ACC of Type 5 and Type 6 are 85.3% and 83.7%, and the FPR is 5.3% and 5.1%.

From the perspective of users' electricity consumption habits, the results of several typical users in the calculation example are shown in Table IV.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1650-1661
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

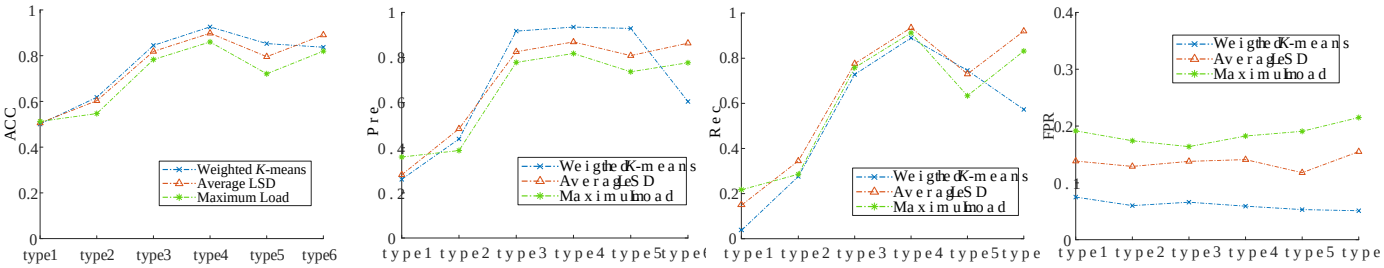
TABLE IV. Detection result of typical users

User ID	Standard deviation	Accuracy	Precision	Recall	UNITS
1	0.015	0.902	0.985	0.819	0.013
2	0.025	0.884	0.956	0.808	0.038
3	0.037	0.883	0.953	0.808	0.040
4	0.063	0.648	0.761	0.444	0.143
5	0.075	0.717	0.776	0.621	0.182
6	0.083	0.594	0.673	0.381	0.189

It can be seen from Table IV that the detection effect is also related to the user's electricity consumption habits. For users with a small standard deviation, that is, users with relatively regular electricity consumption habits, such as user 1 and user 2, the detection effect of electricity theft is better, the accuracy rate is higher, and the false detection rate is also guaranteed to be at a low value. For users with a large standard deviation, that is, users with irregular electricity consumption habits, the accuracy of the proposed method may be affected significantly.

4.3 Experiments No.2

This paper also uses the traditional typical day LSD acquisition method, namely, the average LSD, the maximum load method and LSD extracted by weighted *K*-means method mentioned in this paper to carry out the electricity theft detection effect comparison experiment. The data source of the experiment is the same as experiment No.1 and the comparison of detection results are shown in Figure 2.



(c) ACC. (d) Precision. (c) Recall. (d) FPR

Fig 2: Comparison of three LSD methods for electricity theft detection results.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1650-1661
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

It can be seen from the experimental results in Fig 2 that in terms of the main evaluation index (ACC and FPR), the LSD extracted by the maximum load is inferior to the other two methods. The reason is inferred that selecting the maximum load cannot well represent other electricity consumption behaviors in the month. Although the *K*-means method mentioned in this paper is not as accurate as the average LSD for the detection of type 6, the detection effect of the other 5 methods of electricity theft is better than the average LSD, and the FPR is lower than the average LSD as a whole. Experiment 2 generally reflects the effectiveness of the typical load extracted by weighted *K*-means proposed in this paper.

V. CONCLUSION

This paper proposes a new electricity theft detection method based on the user's electricity consumption behavior with a tunable threshold. The abnormal degree of user profiles is calculated according to CDs between the load curve and LSD proposed in this paper. To demonstrate the effectiveness of the detection and LSD method, numerical experiments and comparisons with other LSD extract method are conducted. Results show that based on the proposed LSD method, the detection technique can precisely detect electricity thefts. However, constrained by the CDs, the method does not specialize in detecting of type1 and type2. Therefore, it is worthwhile for us to investigate how to supplement the detection for these types in next step.

ACKNOWLEDGEMENTS

This research was supported by the Fundamental Research Funds for the Central Universities (Grant No. 2019MS10).

REFERENCES

[1] Viegas JL, Esteves PR, Melício R, Mendes VMF (2017) Solutions for detection of non-technical losses in the electricity grid: A review. Renewable and Sustainable Energy Reviews 80: 1256-1268.
[2] Federal Bureau of Investigation (2012) Intelligence section: smart grid electric meters altered to steal electricity. Available at: https://krebsonsecurity.com/2012/04/fbi-smart-meter-hacks-likelyto-spread/
[3] Chen Q, Zheng K, Kang C, Huangfu F (2018) Detection methods of abnormal electricity consumption behaviors: review and prospect. Automation of Electric Power Systems 42: 189-199. (In Chinese)
[4] De Nadai M, van Someren M (2015) Short-term anomaly detection in gas consumption through ARIMA and Artificial Neural Network forecast. 2015 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems (EESMS) Proceedings: 250-255.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1650-1661
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

[5] Jokar P, Arianpoo N, Leung VCM (2016) Electricity theft detection in AMI using customers' consumption patterns. IEEE Transactions on Smart Grid 7: 216-226.

[6] Zheng Z, Yang Y, Niu X, Dai (2018) Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. IEEE Transactions on Industrial Informatics 14: 1606-1615.

[7] Ismail M, Shaaban MF, Naidu M, Serpedin E (2020) Deep learning detection of electricity theft cyber-attacks in renewable distributed generation. IEEE Transactions on Smart Grid 11: 3428-3437.

[8] Zheng K, Wang Y, Chen Q, Li Y (2017) Electricity theft detecting based on density-clustering method. 2017 IEEE Innovative Smart Grid Technologies - Asia: 1-6.

[9] Takiddin A, Ismail M, Zafar U, Serpedin E (2021) Robust electricity theft detection against data poisoning attacks in smart grids. IEEE Transactions on Smart Grid 12: 2675-2684.

[10] Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y (2019) Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Transactions on Smart Grid 10: 841-851.

[11] Commission for Energy Regulation (2012) CER Project Electricity customer behaviour trial, 2009–2010. Available at: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/