A Novel Unsupervised Data-Driven Method for Electricity Theft Detection in AMI Using Observer Meters

Ruobin Qi, Jun Zheng, *Member, IEEE*, Zhirui Luo, and Qingqing Li

***Abstract*—The smart meter data of the advanced metering infrastructure (AMI) can be tampered by electricity thieves with advanced digital instruments or cyber attacks to reduce their electricity bills, which causes devastating financial losses to utilities. A novel unsupervised data-driven method for electricity theft detection in AMI is proposed in this article. The method incorporates observer meter data, wavelet-based feature extraction, and fuzzy c-means (FCM) clustering. A new anomaly score is developed based on the degree of cluster membership information produced by FCM clustering to differentiate normal and fraudulent users. We perform an ablation study to investigate the impact of key components of the proposed method on the performance using a publicly available smart meter dataset. The results show that all key components of the proposed method contribute significantly to the performance improvement. The proposed method is compared with a set of baselines including state-of-the-art methods using smart meter data of both business users and residential users. The comparison results indicate that the proposed method achieves significantly better detection performance than all baseline methods. We also show that the proposed method maintains a good performance when the detection time frame is reduced from 30 to 20 days.**

***Index Terms*—Advanced metering infrastructure (AMI), electricity theft detection, false data injection (FDI), machine learning, nontechnical loss (NTL), smart meter.**

I. Introduction

SMART grids provide more efficient and resilient power delivery and better demand-side management than traditional power grids by providing a bidirectional power and information flow through advanced communication and information technologies [1], [2]. The advanced metering infrastructure (AMI) is a key component of smart grids which uses smart meters to collect a large amount of high-frequency electricity consumption data from users. With the collected data, utilities can provide advanced data-driven services such as load forecasting [3] and demand response [4].

Manuscript received 25 February 2022; revised 20 April 2022; accepted 27 June 2022. Date of publication 11 July 2022; date of current version 15 July 2022. This work was supported in part by the National Science Foundation under EPSCoR Cooperative Agreement OIA-1757207 and in part by the Institute for Complex Additive Systems Analysis (ICASA) of the New Mexico Institute of Mining and Technology. The Associate Editor coordinating the review process was Dr. Zhenbing Zhao. *(Corresponding author: Jun Zheng.)*

The authors are with the Department of Computer Science and Engineering, New Mexico Institute of Mining and Technology, Socorro, NM 87801 USA (e-mail: ruobin.qi@student.nmt.edu; jun.zheng@nmt.edu; zhirui.luo@student.nmt.edu; qingqing.li@student.nmt.edu).

Digital Object Identifier 10.1109/TIM.2022.3189748

Although many benefits have been brought to smart grids by AMI, it also introduces a number of security vulnerabilities. Traditionally, electricity thieves manipulate meter readings by physically bypassing or destructing mechanical meters [5], [6] to reduce their electricity bills. AMI opens a new attack surface for electricity thieves where they can use advanced digital instruments or launch cyber attacks to tamper with the smart meter data. Electricity theft brings devastating financial losses every year in both developing and developed countries. A study conducted by the Northeast Group in 2017 shows that utilities lose $96 billion every year globally due to electricity theft and other nontechnical losses (NTLs) [7]. A recent news article reported that electricity theft has caused the state-owned Federal Electricity Commission (CFE) of Mexico approximately $2.5 billion loss per year [8].

Traditional methods for electricity theft detection rely on personal inspecting or video surveillance, which are time-consuming and/or labor-intensive. By taking advantage of data and information collected by AMI, data-driven methods have been developed in recent years to address the problems of traditional methods. Data-driven electricity theft detection methods can be classified into three categories according to [9]: state-based, game-theory-based, and machine-learning-based.

The state-based methods perform state estimation of the power system to find electricity usage irregularity. In [10], a method based on three-phase state estimation for distribution transformer load estimation was proposed to detect meter malfunction and electricity theft. Based on the state estimation results, an analysis of variance (ANOVA) was used to find suspicious users. A multi timescale theft estimation (MISTE) method was proposed in [11] which relies on phasor measurement unit (PMU) measurements to detect electricity theft in the distribution network. State-based methods have some limitations. The detection can only be done at the substation level instead of at the end-user level because the information of the topology and parameters of the distribution network are needed. Some methods require extra devices such as the adapted ammeter developed in [12].

The game-theory-based methods treat the tampering behaviors of electricity thieves as a game between the thieves and utilities. Amin *et al.* [13] developed a game theoretic framework to model the adversarial nature of the electricity theft problem in AMI. In [14], a distributed intelligent

TABLE I
SIX TYPES OF FDI ATTACKS

Attack Type	Modification
1	$\tilde{x_t} = \alpha x_t,\ 0.2 < \alpha < 0.8$
2	$\tilde{x_t} = \begin{cases} x_t & \text{if } x_t \le \sigma \\ \sigma & \text{if } x_t > \sigma \end{cases} \quad \sigma < max(\mathbf{x})$
3	$\tilde{x_t} = max(x_t - \gamma, 0),\ \gamma < max(\mathbf{x})$
4	$\tilde{x_t} = f(t)x_t,\ f(t) = \begin{cases} 0 & \text{if } t_1 < t < t_2 \\ 1 & \text{otherwise} \end{cases}$
5	$\tilde{x_t} = \alpha_t x_t,\ 0.2 < \alpha_t < 0.8$
6	$\tilde{x_t} = \alpha_t \bar{x},\ 0.2 < \alpha_t < 0.8$

in a time interval t, respectively. $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ is the daily electricity consumption profile where n is the number of readings in a day by the smart meter. max(x) is the maximum value of x. $\bar{x}$ is the average daily electricity consumption measured by the smart meter before tampering.

As shown in Table I, type 1 attack reduces all smart meter readings in a day with a randomly generated scaling parameter α ranging from 0.2 to 0.8. Type 2 attack clips the smart meter reading when it is larger than a randomly chosen threshold σ which is less than max(x). Type 3 attack cuts the smart meter reading by a randomly chosen value γ which is less than max(x). When type 4 attack is launched, smart meter readings in a randomly chosen time period (t_1, t_2) are replaced with zero values. The length of the time period is longer than 4 h. Type 5 attack modifies the smart meter reading in a time interval t using a randomly generated scale parameter $\alpha_t \in (0.2, 0.8)$. Finally, type 6 attack replaces the smart metering reading in a time interval t with the average consumption in the day times a randomly generated scaling parameter $\alpha_t \in (0.2, 0.8)$. An example of electricity consumption profile and the corresponding tampered profiles by using the six types of FDI attacks are shown in Fig. 2. Note that an electricity thief can tamper the smart meter data with a mixture of different types of FDI attacks.

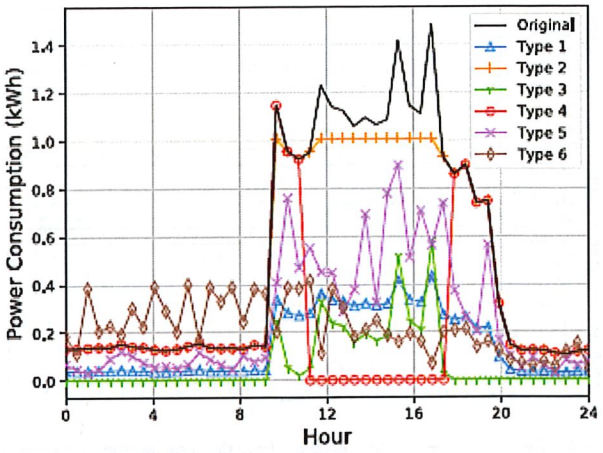

Fig. 2. Example of electricity consumption profile and its corresponding tampered profiles by applying six types of FDI attacks.

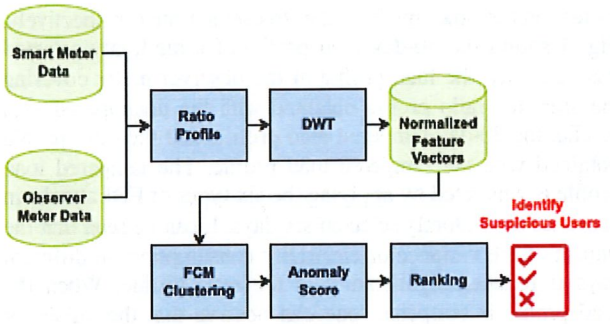

Fig. 3. Workflow of the proposed data-driven method for electricity theft detection.

III. METHODOLOGY

In this section, we introduce the proposed unsupervised data-driven method for electricity theft detection. The workflow of the method is illustrated in the flow diagram of Fig. 3 and described in detail in the following.

A. Data Preprocessing

The smart meter readings collected in AMI usually contain missing data or erroneous data (outliers), which are due to a variety of reasons such as malfunctioned smart meters, transmission errors, bad connections, and system maintenance [17]. In the data preprocessing stage, we recover missing data with the mean value method [16] as shown in the following equation:

$$x_t = \text{mean}(\mathbf{x}) \quad \text{if} \quad x_t = \text{NaN} \tag{1}$$

where x_t is the electricity consumption in a time period t, and x is the daily electricity consumption profile not including the missing data. If x_t is a missing value, it is represented as NaN (not a number).

To recover erroneous data, we apply the following equation which is based on the "three-sigma rule of thumb" [16]:

$$x_t = \begin{cases} \frac{x_{t-1}+x_{t+1}}{2}, & \text{if } x_t > 3\sigma(\mathbf{x}) \quad x_{t-1}, x_{t+1} \neq \text{NaN} \\ x_t, & \text{otherwise} \end{cases} \tag{2}$$

where $\sigma(\mathbf{x})$ is the standard deviation of the consumption profile x.

B. Ratio Profile

The electricity consumption of users at certain times could be significantly lower than other times, e.g., weekend versus weekdays for business users. When load profiles of users are directly used for detection, it may result in degraded performance due to a significant amount of false alarms. The reading of the observer meter provides a good reference to capture such effect. Thus, we derive the ratio profile of a user using the load profile and the observer meter profile which is calculated as shown in the following equation:

$$R_{i,t} = \frac{O_t}{x_{i,t}} \tag{3}$$

user's N-day load profile, the extracted N feature vectors are grouped as $M = 2$ clusters with FCM clustering. The output of FCM clustering is $\{\mu_{i,j}, i = 1, 2, \ldots, N, j = 1, 2\}$.

E. Anomaly Score and Fraudulent User Detection

The output of FCM clustering is used to derive the anomaly score which indicates the suspicious degree of a user to be fraudulent. We denote the two clusters with high and low average electricity consumption as h-cluster and l-cluster, respectively. The anomaly score for a user u, AS_u, is calculated as follows:

$$AS_u = \frac{E_h - E_l}{E_h} \tag{7}$$

where E_h and E_l are the weighted average electricity consumption of the h-cluster and l-cluster, respectively. Assuming E_i $(i = 1, 2, \ldots, N)$ is the electricity consumption of the user on the ith day, E_h and E_l are calculated as shown in the following equation:

$$E_j = \frac{\sum_{i=1}^{N} \mu_{i,j} E_i}{\sum_{i=1}^{N} \mu_{i,j}} \quad j \in \{h, l\}. \tag{8}$$

The rationale behind the proposed anomaly score is that the tampered profiles of a fraudulent user are more likely to be clustered into l-cluster which results in low average electricity consumption of l-cluster. Thus, the difference between the average electricity consumption of the two clusters of a fraudulent user will be more significant than that of a normal user. The introduction of the degree of cluster membership into the calculation of the average electricity consumption of a cluster will lead to a larger difference between the anomaly scores of normal and fraudulent users so that they are easier to be differentiated.

Finally, the users in the area are ranked in ascending order based on their anomaly scores. A user is considered as potentially fraudulent if the rank is high.

IV. Evaluation Experiments

A. Smart Meter Dataset

The smart meter dataset adopted for our performance evaluation is publicly available from the Irish CER Smart Metering Project [27] which is widely used in the research of electricity theft detection. The dataset contains more than 500 days of smart meter data collected from over 5000 residential users and small- and medium-sized business users during 2009 and 2010. The dataset is considered as not tampered and can be used as ground truth because all the users completed the pretrial or posttrial surveys. The sampling rate of the smart meter is one sample every half an hour such that a daily load profile consists of 48 samples. The load profiles of 334 business users from two consecutive 30-day time periods (July 14 to August 12 and August 13 to September 12, 2009) are used in our study. Similar to [6], we randomly and evenly divide the users into areas where each area is covered by an observer meter. In the experiments, a portion of users in an area are chosen as electricity thieves and half of their daily load profiles are tampered with certain types of FDI attacks.

B. Baseline Methods for Performance Comparison

We consider a set of baseline unsupervised electricity theft detection methods for performance comparison, which include the follows.

1) *Pearson Correlation Coefficient (PCC) and Maximum Information Coefficient (MIC)*: Two popular correlation analysis methods were used as reference methods in [6] and [18] to detect electricity theft by measuring the correlation between the NTL profile and the tampered profiles. The NTL profile is obtained by subtracting the observer meter profile by the sum of users' load profiles in the area including tampered profiles.
2) *CFSFDP Clustering*: A novel density-based clustering algorithm proposed in [28] was used in [29] to detect electricity theft using users' load profiles.
3) *Local Outlier Factor (LOF)*: A classical outlier detection method was used as a reference method in [6] and [18] to detect electricity thieves using users' load profiles.
4) *CDDA-A and CDDA-G*: A novel data-driven approach combines MIC-based correlation analysis and CFSFDP clustering for electricity theft detection using both the observer meter profile and users' load profiles [6]. The method generates a combined rank for detection using either the arithmetic mean or the geometric mean of the two ranks obtained by MIC and CFSFDP, which are denoted as CDDA-A and CDDA-G, respectively, in this article.
5) *CLOF*: A new outlier detection method proposed in [18] combines k-means clustering and LOF to detect electricity theft using users' load profiles.

C. Evaluation Metrics

Two metrics widely used for evaluating the performance of electricity theft detection methods [6], [17], [18], [29], area under the curve (AUC) and mean average precision (MAP), are adopted in our study.

AUC is a popular metric used to evaluate the classification/ranking accuracy [17]. When used for evaluating the ranking accuracy of an unsupervised electricity theft detection method, AUC is calculated as the probability that a randomly chosen electricity thief has a higher rank than a randomly chosen normal user as shown in the following equation:

$$\text{AUC} = \frac{\sum_{i \in \mathbf{T}} r_i - \frac{1}{2} N_f \times (N_f + 1)}{N_f \times N_n} \tag{9}$$

where r_i is the rank of the ith user, T is the set of electricity thieves, and N_f and N_n are the numbers of electricity thieves and normal users, respectively. Note that the users in a neighboring area are ranked in ascending order based on their anomaly scores to calculate the AUC.

MAP is a metric widely used in the information retrieval area to evaluate the retrieval quality. Based on the list of users ranked in descending order by their anomaly scores, X_k is defined as the number of electricity thieves ranked in top-k of the list. We define the precision at k $P@k$ as

$$P@k = \frac{X_k}{k}. \tag{10}$$

TABLE IV
PERFORMANCE COMPARISON RESULTS IN TERMS OF AUC (%)

Attack Type	Proposed	PCC	MIC	CFSFDP	LOF	CDDA-A	CDDA-G	CLOF
1	**94.4 (1.5)**	80.8 (3.3)	83.4 (2.5)	50.2 (4.8)	41.8 (3.9)	72.5 (3.8)	76.9 (3.2)	41.8 (3.5)
2	**90.3 (3.6)**	59.9 (3.7)	59.4 (4.0)	58.4 (4.4)	45.4 (4.1)	65.7 (4.4)	64.3 (4.6)	41.3 (4.5)
3	**99.6 (0.3)**	49.8 (4.0)	63.7 (4.0)	72.3 (3.0)	43.4 (4.2)	78.9 (3.2)	79.7 (2.9)	44.7 (4.4)
4	**98.1 (0.6)**	28.6 (3.7)	74.4 (3.2)	92.0 (1.3)	59.7 (4.1)	93.5 (1.5)	94.0 (1.4)	61.9 (4.6)
5	**91.6 (2.0)**	40.6 (3.4)	61.3 (4.1)	56.8 (3.4)	46.9 (4.3)	67.2 (3.6)	65.5 (3.8)	48.9 (4.0)
6	**90.4 (2.9)**	36.0 (5.7)	41.2 (5.1)	75.3 (2.5)	42.0 (3.6)	65.2 (4.8)	65.1 (4.6)	40.9 (3.7)
MIX	**94.0 (1.9)**	47.9 (4.3)	61.7 (4.1)	70.0 (3.5)	46.8 (4.1)	73.2 (3.8)	74.7 (3.5)	47.2 (3.7)

TABLE V
PERFORMANCE COMPARISON RESULTS IN TERMS OF MAP@20 (%)

Attack Type	Proposed	PCC	MIC	CFSFDP	LOF	CDDA-A	CDDA-G	CLOF
1	**76.9 (3.6)**	68.6 (5.2)	68.9 (5.2)	27.8 (5.2)	16.9 (3.1)	52.6 (5.5)	57.4 (5.0)	17.1 (3.0)
2	**86.7 (4.3)**	34.1 (5.3)	39.7 (6.2)	26.8 (3.7)	18.6 (3.0)	42.9 (5.6)	48.2 (5.2)	16.4 (2.9)
3	**97.8 (1.3)**	24.4 (5.3)	42.8 (6.1)	36.9 (4.2)	17.3 (3.2)	54.4 (4.5)	58.9 (4.2)	20.8 (4.7)
4	**87.5 (3.8)**	8.8 (2.6)	45.5 (5.5)	65.4 (4.9)	30.2 (4.2)	78.0 (4.7)	80.7 (4.5)	37.8 (5.8)
5	**68.1 (4.2)**	14.6 (2.9)	42.0 (5.7)	25.9 (4.4)	21.6 (4.0)	38.8 (5.3)	45.8 (5.1)	22.7 (3.5)
6	**79.1 (4.2)**	14.8 (4.4)	20.7 (5.2)	33.4 (3.6)	16.6 (3.2)	36.0 (5.7)	42.1 (5.3)	16.3 (3.1)
MIX	**84.2 (3.4)**	32.6 (7.3)	44.8 (6.9)	39.6 (4.6)	20.4 (3.6)	52.2 (5.4)	56.3 (5.1)	22.2 (3.9)

than MIC for detecting type 1 attack and CFSFDP for detecting type 6 attack, they achieve the best performance under the MIX attack among all the baselines which shows that combining the correlation analysis and clustering is a better choice than applying them individually. It can be seen from the results that LOF does not have a good performance while CLOF has slightly improved performance over LOF, which shows that the k-means clustering introduced in [18] contributes to the performance improvement but the effect is minor.

The results in Tables IV and V show that the proposed method achieves significantly better detection performance in terms of both AUC and MAP@20 than baseline methods under all types of FDI attacks. Compared with the best performed baseline CDDA-G, the average AUC and MAP@20 values of the proposed method under all types of attacks are 20.1% and 26.6% higher, respectively. It can also be observed that the proposed method has significantly lower standard deviations of AUC and MAP@20 than baselines, which indicates that the proposed method can achieve more stable detection results. The results demonstrate the superiority of the proposed method over the baselines.

3) Sensitivity Analysis of the Proposed Method: In the real world, the areas covered by observer meters may contain different numbers of users. The proportion of electricity thieves in an area and the tampering behavior of a fraudulent user such as the number of tampered days in a time frame could also vary. Thus, we want to analyze the sensitivity of the proposed method to those parameters.

We first randomly and evenly divide the users into ten areas which is the same as the experiments of the ablation study. On the other hand, we change the number of electricity thieves in each area from 1 to 8 to see how it affects the performance of the proposed method. The results are shown in Fig. 6. It can be seen that the AUC and MAP@20 values for different numbers of thieves are similar showing that the proposed method can generate stable results with varying numbers of thieves.

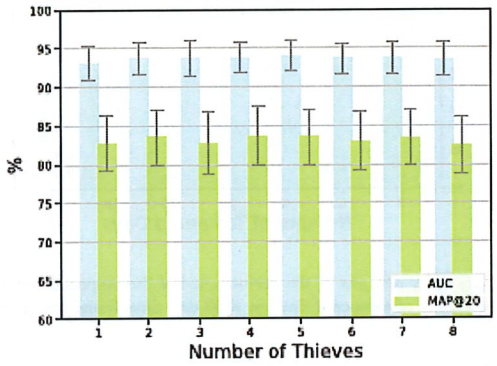

Fig. 6. Sensitivity analysis of the performance of the proposed method with different numbers of thieves per area under the MIX attack.

We then fix the number of thieves in each area as five and change the number of users in each area by dividing the users into different numbers of areas. The results are shown in Fig. 7. As can be seen from the figure, the AUC values are stable against the varying numbers of users per area. On the other hand, the MAP@20 value increases when the number of users per area decreases.

Finally, we analyze the effect of the tampering behavior of fraudulent users in terms of the number of tampered days on the performance of the proposed method. The number of users and the number of electricity thieves in each area are set the same as those used in the ablation study. On the other hand, the number of tampered days varies from 3 to 21 days in a 30-day time frame (i.e., 10%–70% of the total time), which corresponds to the degree of tampering behavior varying from light to heavy. The results are shown in Fig. 8. It can be seen that the performance of the proposed method is slightly reduced when the number of tampered days is 3 corresponding

TABLE VI
RESULTS FOR RESIDENTIAL USERS IN TERMS OF AUC (%)

Attack Type	Proposed	PCC	MIC	CFSFDP	LOF	CDDA-A	CDDA-G	CLOF
1	95.6 (1.6)	**98.4 (1.1)**	96.7 (1.7)	49.9 (4.7)	31.1 (3.5)	75.2 (3.7)	82.3 (2.6)	27.5 (3.3)
2	88.0 (3.3)	84.5 (3.2)	86.7 (3.7)	82.8 (2.5)	37.7 (3.9)	89.2 (3.1)	**90.9 (2.6)**	29.8 (4.1)
3	**99.3 (0.5)**	80.5 (2.8)	78.4 (4.0)	65.0 (3.2)	33.9 (3.1)	77.3 (3.7)	78.2 (3.6)	31.2 (3.2)
4	**98.5 (0.9)**	18.3 (3.1)	98.0 (0.8)	82.4 (2.5)	50.2 (4.1)	95.8 (1.2)	96.3 (0.9)	51.7 (4.1)
5	**94.9 (1.5)**	73.5 (4.4)	67.9 (4.0)	49.9 (4.2)	38.4 (4.4)	62.6 (4.1)	61.7 (4.3)	40.3 (4.4)
6	**95.1 (1.7)**	18.5 (3.6)	54.0 (4.0)	83.9 (1.9)	20.0 (3.0)	76.8 (3.2)	77.5 (3.0)	16.0 (2.5)
MIX	**95.1 (2.0)**	58.7 (5.6)	76.8 (3.6)	70.2 (3.5)	36.1 (4.2)	77.4 (3.5)	80.0 (3.3)	33.6 (4.5)

TABLE VII
RESULTS FOR RESIDENTIAL USERS IN TERMS OF MAP@20 (%)

Attack Type	Proposed	PCC	MIC	CFSFDP	LOF	CDDA-A	CDDA-G	CLOF
1	82.2 (5.6)	**97.5 (1.4)**	92.4 (3.0)	25.9 (4.6)	12.6 (2.9)	54.4 (6.0)	58.0 (5.2)	12.0 (3.4)
2	**83.0 (3.8)**	62.9 (5.2)	80.7 (4.6)	47.7 (4.7)	15.4 (3.7)	79.9 (5.2)	80.5 (5.2)	11.7 (3.5)
3	**95.4 (2.3)**	54.5 (5.4)	68.3 (6.2)	29.9 (3.6)	13.1 (3.0)	56.1 (5.2)	59.8 (5.0)	12.5 (3.0)
4	**90.0 (5.6)**	6.5 (3.4)	89.2 (3.3)	46.5 (4.4)	25.2 (4.7)	82.6 (3.8)	85.7 (3.7)	26.4 (5.1)
5	**74.5 (4.9)**	61.9 (6.1)	46.0 (6.3)	22.2 (4.2)	18.5 (4.6)	37.9 (5.3)	42.6 (5.0)	18.8 (3.8)
6	**80.9 (4.8)**	5.9 (3.5)	25.5 (5.0)	42.6 (3.8)	7.3 (3.2)	43.7 (4.7)	49.4 (4.6)	4.9 (2.4)
MIX	**85.5 (4.5)**	58.8 (7.7)	69.0 (5.5)	39.4 (5.2)	16.2 (4.1)	59.6 (5.5)	62.0 (5.8)	16.4 (4.8)

Section IV-D2, the correlation analysis works very well for detecting type 1 attack due to the high correlation between the NTL profile and tampered profiles. CDDA-A and CDDA-G have higher AUC values than the proposed method under type 2 attack but their MAP@20 values are lower than the proposed method. Nevertheless, the proposed method achieves close performance in terms of AUC (<3%) compared with PCC and MIC under type 1 attack, and CDDA-A and CDDA-G under type 2 attack. On the other hand, the proposed method achieves significantly higher AUC and MAP@20 values than the baseline methods under type 3–6 attacks, which results in its high performance in detecting the MIX attack. The AUC and MAP@20 values of the proposed method are 15.1% and 16.5% higher than those of the best baseline methods under the MIX attack, respectively. The results demonstrate that the proposed method is a promising electricity theft detection solution not only for business users but also for residential users.

V. CONCLUSION

Electricity theft is a serious security problem for AMI which causes a huge amount of financial losses to utilities all over the world each year. In this article, we propose a novel unsupervised data-driven method for electricity theft detection which uses observer meter data, DWT-based feature extraction, and FCM clustering. A new anomaly score is developed which is calculated using the degree of cluster membership information produced by FCM clustering for better differentiating normal and fraudulent users. Using a publicly available smart meter dataset, we perform an ablation study which shows that all the key components of the proposed method contribute significantly to the performance improvement and the most significant improvement comes from the use of ratio profile. The performance comparison with the baselines including state-of-the-art methods shows that the proposed method achieves much better performance for both business users and residential users. Although the results show that the proposed method can maintain the detection performance when being applied in a shorter time frame, it cannot detect daily tampering behaviors of a user which is the limitation of all unsupervised electricity detection methods. In the future, we will solve this problem using semisupervised learning algorithms which only use normal data to build novelty detection models so that the problems of supervised learning such as imbalanced datasets and over-fitting are avoided. There are only a few works in this area such as one-class SVM used in [9] but the detection performance is not good. The aim of our future work is to develop advanced semisupervised novelty detection algorithms that can serve as a viable solution to detect a user's daily tampering behaviors.

REFERENCES

[1] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid—The new and improved power grid: A survey," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, pp. 944–980, 4th Quart., 2011.

[2] K. Yu, M. Arifuzzaman, Z. Wen, D. Zhang, and T. Sato, "A key management scheme for secure communications of information centric advanced metering infrastructure in smart grid," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 8, pp. 2072–2085, Aug. 2015.

[3] F. L. Quilumba, W.-J. Lee, H. Huang, D. Y. Wang, and R. Szabados, "An overview of AMI data preprocessing to enhance the performance of load forecasting," in *Proc. IEEE Ind. Appl. Soc. Annu. Meeting*, Oct. 2014, pp. 1–7.

[4] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang, and Y. Zhao, "Load profiling and its application to demand response: A review," *Tsinghua Sci. Technol.*, vol. 20, no. 2, pp. 117–129, Apr. 2015.

[5] R. Jiang, R. Lu, Y. Wang, J. Luo, C. Shen, and X. Shen, "Energy-theft detection issues for advanced metering infrastructure in smart grid," *Tsinghua Sci. Technol.*, vol. 19, no. 2, pp. 105–120, Apr. 2014.

[6] K. Zheng, Q. Chen, Y. Wang, C. Kang, and Q. Xia, "A novel combined data-driven approach for electricity theft detection," *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1809–1819, Mar. 2018.

[7] A. Theron-Ord. *Electricity Theft and Non-Technical Losses Total $96BN Annually—Report.* Smart Energy International. Accessed: May 12, 2017. [Online]. Available: https://www.smart-energy.com/regional-news/africa-middle-east/electricity-theft-96bn-annually/

[8] *Theft Continues to Create Big Losses for Electricity Commission.* Mexico News Daily. Accessed: Sep. 13, 2021. [Online]. Available: https://mexiconewsdaily.com/news/theft-continues-to-create-big-losses-for-electricity-commission/