IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 10, 2021, accepted January 4, 2022, date of publication January 7, 2022, date of current version January 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3141610

A Detection Method for Group Fixed Ratio Electricity Thieves Based on Correlation Analysis of Non-Technical Loss

YINING YANG[1], RUNAN SONG[1], YANG XUE[1], PENGHE ZHANG[1], YUEJIE XU[2], JINPING KANG[2], (Member, IEEE), AND HAISEN ZHAO[2], (Senior Member, IEEE)

[1]China Electric Power Research Institute, Haidian, Beijing 100192, China
[2]State Key Laboratory of Alternative Electricity Power System With Renewable Energy Sources, North China Electric Power University, Changping, Beijing 102206, China

Corresponding author: Haisen Zhao (zhaohisen@163.com)

This work was supported by the State Grid Technology Project under Grant 5400-201925177A-0-0-00.

ABSTRACT Owing to the contagiousness of theft behaviors among customers, collaborative energy theft, such as village fraud, has become particularly common. In this study, a bunch of electricity thieves that steal energy at a constant ratio were considered. Conventional correlation-sorting-based methods may have some trouble handling these electricity thieves when they exist in the same area. To overcome such limitation, we firstly establish the mathematical model of non-technical loss (NTL) and the load data of fixed ratio electricity thieves (FRETs). Subsequently, an interesting correlation trend, which can be exploited to locate FRETs, was observed and analyzed. Based on this trend, we propose a correlation analysis-based detection method. It adopts a standardized covariance to measure the correlation between the NTL and user data. The detection of FRETs is realized by solving a combinatorial optimization problem. A corresponding framework in practice was also designed. Finally, numerical experiments based on a realistic dataset and an electricity theft dataset from an electricity theft emulator (ETE) are conducted to validate the effectiveness and superiority of the proposed method in terms of accuracy, stability, and scalability.

INDEX TERMS Data mining, electricity theft detection, fixed ratio electricity theft, covariance analysis.

NOMENCLATURE

A. INDICES

t	Index of time interval.
i, j, h	Index of user and data sample.
d	Index of day.
k	Index of optimization order

B. SETS

$\mathcal{A}$	Set of all users in an area.
$\mathcal{B}$	Set of benign users in the area.
$\mathcal{C}$	Set of fraudulent users in the area.
$\mathcal{C}_{\text{sus}}$	Set of suspicious users.
$\mathcal{C}_{\text{sus},d}$	Set of suspicious users on d-th day.
$\mathcal{P}^*$	Any user set having the increasing trend of correlation.
$\mathcal{P}_i^*$	Any user set containing user i and having the increasing trend of correlation.
$\mathcal{P}_{i\max}^*$	Best solution of the sub-problems of user i.
$\mathcal{P}_{\max}^*$	Best solution of the combinatorial optimization problem.

C. VARIABLES AND PARAMETERS

$u_{i,t}$	Ground truth load for user i at time interval t.
$\tilde{u}_{i,t}$	Recorded load for user i at time interval t.
$\tilde{\boldsymbol{u}}_i$	Recorded load vector for user i.
$\tilde{\boldsymbol{u}}_i^*$	Normalized recorded load vector for user i.
$\tilde{\boldsymbol{u}}_{i,d}^*$	Normalized recorded load vector for user i on day d.
ω_t	NTL of an area at time interval t.
$\boldsymbol{\omega}$	NTL vector of an area.
$\boldsymbol{\omega}^*$	Normalized NTL vector of an area.
$\boldsymbol{\omega}_d^*$	Normalized NTL vector of an area on d-th day.
α_i	Ratio that $\tilde{u}_{i,t}$ accounts for $u_{i,t}$

The associate editor coordinating the review of this manuscript and approving it for publication was Fabio Mottola.

β_i	Representation coefficient of the fixed ratio electricity thieves i.
$\boldsymbol{\beta}$	Coefficient vector composed of all the β_i.
$\boldsymbol{\beta}^*$	Normalized coefficient vector.
$\tilde{\boldsymbol{U}}$	Matrix composed of all the $\tilde{\boldsymbol{u}}_i$ when $i \in \mathcal{C}$.
$\tilde{\boldsymbol{U}}^*$	Matrix composed of all the $\tilde{\boldsymbol{u}}_i^*$ when $i \in \mathcal{C}$.
$\tilde{\boldsymbol{\Phi}}$	Matrix composed of the covariance between every $\tilde{\boldsymbol{u}}_i^*$ when $i \in \mathcal{C}$.
ϕ_i	Column vector of the i-th column of $\tilde{\boldsymbol{\Phi}}$.
$\tilde{\boldsymbol{u}}_{\mathcal{C}}^*$	Combined vector of all the $\tilde{\boldsymbol{u}}_i^*$ when $i \in \mathcal{C}$.
$\tilde{\boldsymbol{u}}_{\mathcal{P}^*}^*$	Combined vector of all the $\tilde{\boldsymbol{u}}_i^*$ when $i \in \mathcal{P}^*$.
$\tilde{\boldsymbol{u}}_{\mathcal{P}_i^*}^*$	Combined vector of all the $\tilde{\boldsymbol{u}}_i^*$ when $i \in \mathcal{P}_i^*$.
$\tilde{\boldsymbol{u}}_{\mathcal{P}_{\max}^*}^*$	Combined vector of all the $\tilde{\boldsymbol{u}}_i^*$ when $i \in \mathcal{P}_{\max}^*$.
m	Number of fraudulent users.
n	Number of benign users.
K	Ending order of optimization searching.
M	Total number of the days of load profiles.
M_i	Number of days of user i belongs to $\mathcal{C}_{\text{sus},d}$.
γ_i	Anomaly degree of user i.

D. FUNCTIONS

$\text{Corr}(\cdot, \cdot)$	Correlation measurement of two vectors.
$\rho(\cdot, \cdot)$	Pearson correlation coefficient of two vectors.
$\text{cov}(\cdot, \cdot)$	Covariance of two vectors.
$\max(\cdot)$	Maximum of a vector.
$\text{mean}(\cdot)$	Arithmetic average of a vector.

I. INTRODUCTION

Transmission loss in a power grid includes technical loss (TL) and non-technical loss (NTL) [1]. TL is the normal loss in the process of power transmission, such as the copper and core losses in transformers. NTL is the remaining loss that cannot be explained theoretically after the TL is removed. Abnormal power consumption behavior, such as electricity theft, is one of the main sources of NTL [1]. In addition, electricity theft can seriously harm the economic performance of power suppliers and cause safety problems, such as equipment damage, power outages, and casualties. For example, the annual bills of stolen electricity in the USA were almost \$ 10 billion in 2017 based on a report by the Northeast Group [2]. In China, the State Grid Corporation has also retrieved theft bills of nearly 13 billion RMB in the past three years.

Fixed-ratio electricity thieves (FRETs) refer to malicious users stealing energy in constant proportion. Most physically meter-tampering users, such as voltage-reducing, current-decreasing, and power factor diminishing users, are FRETs [3]. Although there have been many nonlinear false data injections (FDIs) via cyberattacks, the high thresholds of the techniques involved in cyberattacks make them less common than FRETs. This is especially true in rural areas of some developing countries [4]. In addition, owing to the contagiousness of theft behavior among customers, malicious users have recently tended to conduct electricity theft in groups. Corresponding cases include collaborative fraud of the entire village [5] and collective minor fraud [6] have been reported. For the above reasons, FRETs are the bugs in the very vitals of power utilities and need to be addressed urgently.

Advanced metering infrastructure, which consists of smart meters and sophisticated communication networks, provides massive amounts of electricity data of users. Thus, technologies such as data mining and artificial intelligence are more appropriate for electricity theft identification [7]. Current data-driven electricity theft detection methods can be classified into three groups [8]:

A. GAME THEORY-BASED

Electricity theft is supposed to be a game between power utilizes and energy thieves in the game-theory-based methods [9], [10]. By formulating a model involving all players in this game, the difference in bills between electricity thieves and normal users can be derived with the theory of Nash equilibrium. This type of method can be used to learn the interactions of different players under different situations, but it is not practically applicable because of the complexity of the modeling.

B. POWER-CONSUMPTION-PATTERN-BASED

This method assumes that the load patterns of electricity thieves deviate from those of normal users. Thus, techniques such as deep learning and machine learning can be adopted to analyze the load profiles of customers for electricity theft identification. Deep-learning-based methods require massive amounts of labeled electricity consumption data for model training. Examples include support vector machines (SVM) [11], convolutional neural networks (CNNs) [12] and other artificial neural networks, which were investigated in [13]–[15]. The accuracy of these supervised methods can be as high as 100% in experiments; however, they only work well when vast reliable samples of electricity theft can be acquired. On the other hand, machine learning-based methods focus on information without labels, that is, outliers deviating from the normal majority. In [16], a method called spatial clustering of applications with noise (DBSCAN) was used to locate abnormal users by calculating their anomaly degree. Zheng *et al.* [17] utilized clustering by fast search and find of density peaks (CFSFDP) to identify nonlinear FDIs.

C. SYSTEM-STATE-BASED

The measurements of voltage, current, and other variables should satisfy the physical equations of the power grid. Therefore, the data for these variables should be consistent. However, the meter tempering behaviors of electricity thieves may ruin this consistency and cause some abnormalities (i.e., NTL and voltage limit violation). This kind of method utilizes this fact to locate fraudulent users. For instance, Leite and Sanches Mantovani [18] exploited the concept of state estimation to monitor the bias between the estimated and measured voltages. Once the bias was detected, the sources of the NTL were located using a pathfinding procedure based on the A-Star algorithm. Another method of this type attempts to detect bypassing users using a data analytical technique derived from the three-phase state estimator [19]. Although state-based methods are capable of locating malicious users (i.e., bypassing individuals), which is beyond the capability of other types of methods, their excellent performance requires accurate information of the topology and parameters of the network. This information is not likely to be available in general situations because of the regular alternation of the network topology. The most practically feasible approach of this type is [20], in which a linear detection model requiring only active power and voltage magnitude measurements was formulated; however, it only can detect three-phase malicious users.

Recently, some hybrid methods combing the traits of power patterns and system states have been proposed [21]. And perhaps, the most directly relevant works are [22],in which a central meter recording the aggregated consumption of users in the neighborhood is deployed. Salinas *et al.* [22] realized the detection for FRETs by calculating the sparest solution of a set of underdetermined linear equations. However, obtaining a solution with low sparsity is still a challenging problem, and its sparsity may not be sufficiently low.

Based on the above brief review, current data-driven electricity detection methods may face their own problems in practice. First, game-theory-based methods are not practically applicable owing to the difficulty of formulating a model involving all players. Second, deep learning-based methods demand massive, trustworthy samples for model training. However, the unbalanced dataset and contaminated samples [25] significantly affect their accuracies. Moreover, they may mistake some normal activities such as meter installation and climate change as theft behaviors. Third, unsupervised methods do not specialize in detecting FRETs because their attacks are linear, and the tampered data are almost identical to the ground truth data after normalization (this will be discussed in Section IV-C).

Our work is an extension and perfection of the research presented in [23] and [24], where they were trying to detect FRETs by correlation-based method. In [23] and [24] the maximum information coefficient (MIC) and a combined correlation coefficient were used respectively to measure the correlation between the NTL and meter readings of users. And a stronger correlation indicated more suspicious FRETs. This correlation-sorting-based method performs well under the scenario of one or two FRETs in the same area. However, when it comes to multiple FRETs, it usually has poor true positive rate. To effectively identify multiple FRETs in the same area, we proposed a detection method based on non-technical loss covariance (NTLC). The main contributions of this study are as follows:

1) We derive the mathematical model between NTL and the load data of FRETs, and adopt standardized covariance to measure the correlation between them.

2) We observe an increasing trend of correlation between the NTL and load data of FRETs with data superposition. A theoretical validity analysis of this trend is conducted to make it exploitable for the detection of FRETs.

3) We put forward a searching approach that needs neither training samples nor accurate information about the topology and parameter of network to catch the FRETs effectively.

4) We conduct numerous extensive tests using a realistic dataset and electricity theft dataset from an electricity theft emulator (ETE). Comparisons with other state-of-the-art methods demonstrate the merits of the presented method in detecting FRETs.

II. PROBLEM STATEMENT AND MODELING

In this section, we describe the applicable scenario and the problem that this study focuses on. The mathematical model between the NTL and load of the FRETs is then established. The main purpose is to provide justifications of the search algorithm presented in the next section.

A. DESCRIPTION OF APPLICABLE SCENE

Figure 1 illustrates a typical distribution network, where a transformer supplies power to multiple areas simultaneously, and each area contains a group of adjacent users. Our method is applicable when a gateway meter is installed in each area to record the aggregated ground-truth consumption of the users monitored by it. Owing to the fine security of the gateway meter and stable topology within the area, the NTL of the area can be directly calculated by subtracting the sum of the kWh measurements of all registered users in the area from the electricity consumption W_t recorded by the gateway meter, i.e.,

$$\omega_t = W_t - \sum\nolimits_{i\in\mathcal{A}} \tilde{u}_{i,t} \tag{1}$$

where $\mathcal{A}$ is the set of all users in area A; $\tilde{u}_{i,t}$ is the recorded electricity consumption of user i during time instance t; and ω_t is the NTL of the area. For areas equipped with no gateway meters, their NTL can also be calculated indirectly using NTL evaluation methods [26], [27] with high precision. Therefore, in this case, the proposed method is feasible.

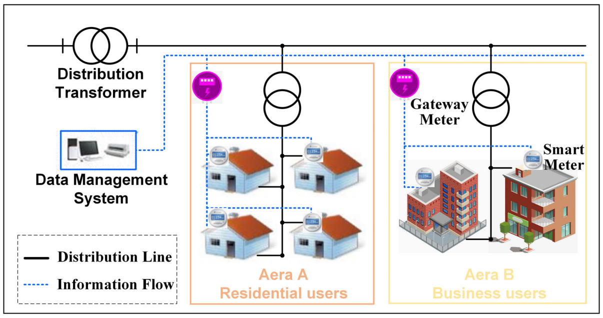


FIGURE 1. Illustration of applicable scene.

Suppose that there are several electricity thieves in area A. Assume that all smart meter function well, which means that the errors of the benign users' meters are within the normal range. Let us denote the set of fraudulent users as $\mathcal{C}$, whose size is m. The remaining benign uses are denoted as $\mathcal{B}$, whose size is n. Note that $\mathcal{A} = \mathcal{B} \cup \mathcal{C}$. The problem this study attempts to address is how to find all these electricity thieves when $\mathcal{C}$ contains only FRETs.

B. MATHEMATICAL MODEL OF NTL AND FRETS' DATA

Assume that user i and his ground-truth load data is $u_{i,t}$. If i is a FRET, the attack is linear, i.e.,

$$\tilde{u}_{i,t} = \alpha_i u_{i,t} \tag{2}$$

where $\tilde{u}_{i,t}$ is the recorded load data, $\alpha_i \in (0, 1)$ is a constant, and represents the ratio of $\tilde{u}_{i,t}$ to $u_{i,t}$. Because the gateway meter cannot be tampered with by fraudulent users, the NTL ω_t caused by them satisfies (3) when $\mathcal{C}$ only contains FRETs:

$$\omega_t = \sum\nolimits_{i\in\mathcal{C}} (u_{i,t} - \tilde{u}_{i,t}) = \sum\nolimits_{i\in\mathcal{C}} \beta_i \tilde{u}_{i,t} \tag{3}$$

where $\beta_i = 1/\alpha_i - 1 > 0$. Let us vectorize the recorded load data and NTL data for one day to obtain the load vector $\tilde{\boldsymbol{u}}_i = [\tilde{u}_{i,1}, \tilde{u}_{i,2}, \cdots, \tilde{u}_{i,T}]^{\mathrm{T}}$ and NTL vector $\boldsymbol{\omega} = [\omega_1, \omega_2, \cdots, \omega_T]^{\mathrm{T}}$. Based on (3), we can derive that $\tilde{\boldsymbol{u}}_i|_{i\in\mathcal{C}}$ and $\boldsymbol{\omega}$ satisfy (4).

$$\boldsymbol{\omega} = \tilde{\boldsymbol{U}}\boldsymbol{\beta} \tag{4}$$

where $\tilde{\boldsymbol{U}} = [\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2, \cdots \tilde{\boldsymbol{u}}_m]$ is a matrix composed of the $\tilde{\boldsymbol{u}}_i$ of m different FRETs, and $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots \beta_m]$ is a ratio vector composed of β_i. Equation (4) shows that when $\mathcal{C}$ contains only FRETs, the NTL vector $\boldsymbol{\omega}$ can be linearly represented by the recorded load vectors $\tilde{\boldsymbol{u}}_i|_{i\in\mathcal{C}}$, and the representation coefficients depend only on β_i. Thus, the correlation between $\boldsymbol{\omega}$ and $\tilde{\boldsymbol{u}}_i$ when $i \in \mathcal{C}$ should be stronger than the correlation when $i \in \mathcal{B}$, i.e.,

$$\mathrm{Corr}(\boldsymbol{\omega}, \tilde{\boldsymbol{u}}_i)|_{i\in\mathcal{C}_R} > Corr(\boldsymbol{\omega}, \tilde{\boldsymbol{u}}_i)|_{i\in\mathcal{B}} \tag{5}$$

where $\mathrm{Corr}(\cdot)$ is a correlation measurement for two vectors, such as Pearson correlation coefficient (PCC)$\rho(\cdot)$. Equation (5) is strictly true when only one or two FRETs exist within the same area. However, when it comes to the scenario of a group of FRETs, in which $\boldsymbol{\omega}$ is jiontly determined by the $\tilde{\boldsymbol{u}}_i$ of all FRETs, (5) may not be valid.

With an area containing four FRETs in Figure 2, the load curves of the four FRETs are clearly dissimilar to the NTL curve. In addition, the PCCs in Table 1 show that the linear correlation between the NTL and load data of any single FRET is not strong. In this case, traditional correlation-sorting-based methods are disadvantageous. The example in Figure 2 also reveals that if we add up the load vectors of all FRETs to obtain a combined vector $\tilde{\boldsymbol{u}}_{\mathcal{C}} = \sum_{i\in\mathcal{C}} \tilde{\boldsymbol{u}}_i$, $\boldsymbol{\omega}$, and $\tilde{\boldsymbol{u}}_{\mathcal{C}}$ may have an extremely strong correlation. Moreover, with each superposition of the FRET load vector, the combined vector and $\boldsymbol{\omega}$ become more correlated, as shown in Table 1:

$$\begin{aligned} &\rho(\boldsymbol{\omega}, \tilde{\boldsymbol{u}}_i)|_{i\in\mathcal{C}} \\ &\leq \rho(\boldsymbol{\omega}, \tilde{\boldsymbol{u}}_i + \tilde{\boldsymbol{u}}_j)\Big|_{i\neq j}^{i,j\in\mathcal{C}} \\ &\leq \rho(\boldsymbol{\omega}, \tilde{\boldsymbol{u}}_i + \tilde{\boldsymbol{u}}_j + \tilde{\boldsymbol{u}}_h)\Big|_{i\neq j\neq h}^{i,j,h\in\mathcal{C}_R} \leq \cdots \leq \rho(\boldsymbol{\omega}, \tilde{\boldsymbol{u}}_{\mathcal{C}}) \end{aligned} \tag{6}$$

TABLE 1. PCCs between NTL and load vectors of FRETs.

FRET	$\rho(\boldsymbol{\omega}, \tilde{\boldsymbol{u}}_i)$	$\rho(\boldsymbol{\omega}, \tilde{\boldsymbol{u}}_{1+2})$	$\rho(\boldsymbol{\omega}, \tilde{\boldsymbol{u}}_{1+2+3})$	$\rho(\boldsymbol{\omega}, \tilde{\boldsymbol{u}}_{1+2+3+4})$
1	0.626	0.933	0.976	0.999
2	0.508			
3	0.419	--		
4	0.619	--	--	

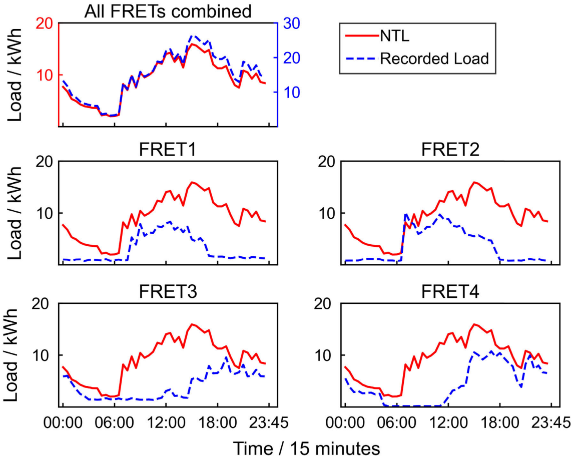


FIGURE 2. Load curves of 4 FRETs and NTL.

If this increasing trend of correlation revealed in (6) is not an individual case, it can be exploited to capture the entire group of FRETs.

C. VALIDITY ANALYSIS OF THE TREND

To exploit the above trend for locating FRETs, we must understand its prerequisite. In this section, we analyze the validity of such correlation trend and provide a theoretical foundation for our method. We first normalize all vectors by (7) to obtain $\boldsymbol{\omega}^*$ and $\tilde{\boldsymbol{u}}_i^*\big|_{i\in\mathcal{A}}$

$$\boldsymbol{x}^* = \frac{\boldsymbol{x}}{\max(\boldsymbol{x})} \tag{7}$$

Because the normalization makes all the vectors under the same dimension, the value of covariance can also measure the correlation of two vectors [28]. Thus, we merely need to explore the precondition of (8) to analyze the validity of the above trend.

$$\begin{aligned} \mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_i^*)\big|_{i\in\mathcal{C}} &\leq \mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_i^* + \tilde{\boldsymbol{u}}_j^*)\Big|_{i\neq j}^{i,j\in\mathcal{C}} \\ &\leq \mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_i^* + \tilde{\boldsymbol{u}}_j^* + \tilde{\boldsymbol{u}}_h^*)\Big|_{i\neq j\neq h}^{i,j,h\in\mathcal{C}} \\ &\leq \cdots \leq \mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_{\mathcal{C}}^*) \end{aligned} \tag{8}$$

where $\mathrm{cov}(\cdot)$ is the function of covariance.

To demonstrate (8), we must introduce the matrix of covariance between every recorded load vector $\tilde{\boldsymbol{u}}_i^*$ when $i \in \mathcal{C}$, as follows.

$$\tilde{\boldsymbol{\Phi}} = \begin{bmatrix} \mathrm{cov}(\tilde{\boldsymbol{u}}_1^*, \tilde{\boldsymbol{u}}_1^*) & \cdots & \mathrm{cov}(\tilde{\boldsymbol{u}}_1^*, \tilde{\boldsymbol{u}}_m^*) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(\tilde{\boldsymbol{u}}_m^*, \tilde{\boldsymbol{u}}_1^*) & \cdots & \mathrm{cov}(\tilde{\boldsymbol{u}}_m^*, \tilde{\boldsymbol{u}}_m^*) \end{bmatrix} = \left[\tilde{\boldsymbol{\phi}}_1, \tilde{\boldsymbol{\phi}}_2, \ldots, \tilde{\boldsymbol{\phi}}_m\right] \tag{9}$$

where $\tilde{\boldsymbol{\phi}}_i$ denotes the i-th column vector of $\tilde{\boldsymbol{\Phi}}$. After normalization, $\boldsymbol{\omega}^*$, $\tilde{\boldsymbol{u}}_i^*\big|_{i\in\mathcal{C}}$, and $\boldsymbol{\beta}^*$ also satisfy the linear equations in (10).

$$\boldsymbol{\omega}^* = \tilde{\boldsymbol{U}}^* \boldsymbol{\beta}^* \tag{10}$$

where $\tilde{\boldsymbol{U}}^* = [\tilde{\boldsymbol{u}}_1^*, \tilde{\boldsymbol{u}}_2^*, \cdots \tilde{\boldsymbol{u}}_m^*]$; $\boldsymbol{\beta}^* = [\beta_1^*, \beta_2^*, \cdots \beta_m^*]^{\mathrm{T}}$ is the normalized ratio vector and $\beta_i^* = \beta_i \max(\tilde{\boldsymbol{u}}_i)/\max(\boldsymbol{\omega})$. Therefore, the recorded load vector $\tilde{\boldsymbol{u}}_i^*$ of FRET i can be expressed as shown in (11).

$$\tilde{\boldsymbol{u}}_i^* = \tilde{\boldsymbol{U}}^* \boldsymbol{e}_i \tag{11}$$

where $\boldsymbol{e}_i$ is an m-dimensional column vector whose i-th element is 1 and the remaining elements are 0. Then, based on the property of covariance, we can obtain the covariance between $\boldsymbol{\omega}^*$ and $\tilde{\boldsymbol{u}}_i^*$ using (12).

$$\begin{aligned}\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_i^*) &= \mathrm{cov}(\tilde{\boldsymbol{U}}^* \boldsymbol{\beta}^*, \tilde{\boldsymbol{U}}^* \boldsymbol{e}_i) = \boldsymbol{\beta}_*^T \tilde{\boldsymbol{\Phi}} \boldsymbol{e}_i \\ &= \boldsymbol{\beta}_*^T \tilde{\boldsymbol{\phi}}_i \end{aligned} \tag{12}$$

Then, select a FRET j from $\mathcal{C}$ ($j \neq i$) to obtain the combined vector $\tilde{\boldsymbol{u}}_{i+j}^* = \tilde{\boldsymbol{u}}_i^* + \tilde{\boldsymbol{u}}_j^*$ given by

$$\tilde{\boldsymbol{u}}_{i+j}^* = \tilde{\boldsymbol{U}}^* \boldsymbol{e}_{i,j} \tag{13}$$

where $\boldsymbol{e}_{i,j}$ is an m-dimensional column vector whose i-th and j-th elements are 1, and the remaining elements are 0. We can then obtain the covariance between $\boldsymbol{\omega}^*$ and $\tilde{\boldsymbol{u}}_{i+j}^*$ from

$$\begin{aligned}\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_{i+j}^*) &= \mathrm{cov}(\tilde{\boldsymbol{U}}^* \boldsymbol{\beta}^*, \tilde{\boldsymbol{U}}^* \boldsymbol{e}_{i,j}) = \boldsymbol{\beta}_*^T \tilde{\boldsymbol{\Phi}} \boldsymbol{e}_{i,j} \\ &= \boldsymbol{\beta}_*^{\mathrm{T}} \tilde{\boldsymbol{\phi}}_i + \boldsymbol{\beta}_*^{\mathrm{T}} \tilde{\boldsymbol{\phi}}_j \end{aligned} \tag{14}$$

Comparing (13) and (14), the increment of covariance caused by the superposition of $\tilde{\boldsymbol{u}}_j^*$ is given by

$$\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_{i+j}^*) - \mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_i^*) = \boldsymbol{\beta}_*^{\mathrm{T}} \tilde{\boldsymbol{\phi}}_j \tag{15}$$

Based on (15), this increment depends only on the ratio vector $\boldsymbol{\beta}^*$ and the column vector $\tilde{\boldsymbol{\phi}}_j$ of the j-th column of matrix $\tilde{\boldsymbol{\Phi}}$. Because every element β_i^* of $\boldsymbol{\beta}^*$ is above zero, when the elements of $\tilde{\boldsymbol{\phi}}_j$ are non-negative, the increment $\boldsymbol{\beta}_*^{\mathrm{T}} \tilde{\boldsymbol{\phi}}_j$ must be positive and $\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_{i+j}^*)$ must be greater than $\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_i^*)$. Similarly, we can deduce that when every element of $\tilde{\boldsymbol{\Phi}}$ is non-negative, every increment caused by the superposition of FRETs is greater than 0. Under these circumstances, equation (8) is strictly true. Finally, we prove the precondition of the above trend, which is:

$$\forall i, \quad j \in \mathcal{C}, \ \mathrm{cov}(\tilde{\boldsymbol{u}}_i^*, \tilde{\boldsymbol{u}}_j^*) \geq 0 \tag{16}$$

In short, if the recorded load vectors of any two FRETs are positively correlated, the correlation between the combined load vector and $\boldsymbol{\omega}$ becomes stronger with each superposition of FRET data. Although this trend may not always be true in every situation, the precondition in (16) is not very difficult to satisfy or approximately satisfy. Therefore, we believe that this trend can be exploited to detect FRET.

III. METHODOLOGY AND DETECTION FRAMEWORK

In this section, we explain our method based on the above correlation trend. Two practical problems are also considered and analyzed. Finally, we design the framework of the method for future practical applications.

A. DETECTION FOR FRETS

Based on the validity analysis in Section II, when the condition in (16) is met, the covariance between the NTL and data of FRETs exhibits an increasing trend with data superposition. Therefore, the detection of FRETs is limited to finding the bunch of users with such a trend. However, this trend is not exclusive to FRET. To locate FRETs precisely and effectively, another significant trait is needed.

Let us denote any user set with the above increasing trend as $\mathcal{P}^*$. That is, when the load vectors of users in $\mathcal{P}^*$ are superposed one by one in a certain order, the covariance between the NTL vector and the superposed vector continues to increase. It is clear that $\mathcal{C}$ belongs to $\mathcal{P}^*$ when the condition in (16) is satisfied. According to (10), $\boldsymbol{\omega}^*$ is determined only by $\tilde{\boldsymbol{u}}_i^*\big|_{i\in\mathcal{C}}$. Thus, if we add up all vectors of $\mathcal{C}$ to obtain $\tilde{\boldsymbol{u}}_{\mathcal{C}}^* = \sum_{i\in\mathcal{C}} \tilde{\boldsymbol{u}}_i^*$, the $\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_{\mathcal{C}}^*)$ should be the maximum of the covariances between $\boldsymbol{\omega}^*$ and the combined vectors $\tilde{\boldsymbol{u}}_{\mathcal{P}^*}^* = \sum_{i\in\mathcal{P}^*} \tilde{\boldsymbol{u}}_i^*$ of all $\mathcal{P}^*$ with the above trend, that is,

$$\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_{\mathcal{C}}^*) \approx \max\nolimits_{\forall \mathcal{P}^* \subset \mathcal{A}} \left[\mathrm{cov}(\boldsymbol{\omega}^*, \sum\nolimits_{i\in\mathcal{P}^*} \tilde{\boldsymbol{u}}_i^*)\right] \tag{17}$$

Therefore, the detection of FRETs can be transformed into an optimization problem: find a user set from all $\mathcal{P}^*$ with the above increasing trend to maximize the covariance between $\boldsymbol{\omega}^*$ and its combined vector, that is,

$$\begin{aligned} &\max \left[\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_{\mathcal{P}^*}^*)\right] \\ &s.t. \begin{cases} \mathcal{P}^* \subset \mathcal{A} \& \mathcal{P}^* \neq \emptyset \\ \tilde{\boldsymbol{u}}_{\mathcal{P}^*}^* = \sum_{i\in\mathcal{P}^*} \tilde{\boldsymbol{u}}_i^* \end{cases} \end{aligned} \tag{18}$$

The best solution to (18) is denoted as $\mathcal{P}_{\max}^*$. When the condition in (16) is satisfied, $\mathcal{P}_{\max}^*$ is nearly identical to $\mathcal{C}$.

To obtain $\mathcal{P}_{\max}^*$, we first decompose (18) into $m + n$ subproblems. Each subproblem can be described as follows:

For user i, find a user set containing it and with the increasing trend to maximize the covariance between $\boldsymbol{\omega}^*$ and its combined vector, that is,

$$\begin{aligned} &\max \left[\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_{\mathcal{P}_i^*}^*)\right] \\ &s.t. \begin{cases} \mathcal{P}_i^* \subset \mathcal{A} \\ \tilde{\boldsymbol{u}}_{\mathcal{P}_i^*}^* = \sum_{j\in\mathcal{P}_i^*} \tilde{\boldsymbol{u}}_j^* \end{cases} \end{aligned} \tag{19}$$

where $\mathcal{P}_i^*$ is any one set containing user i and with the increasing trend. The best solution to sub-problem of user i is

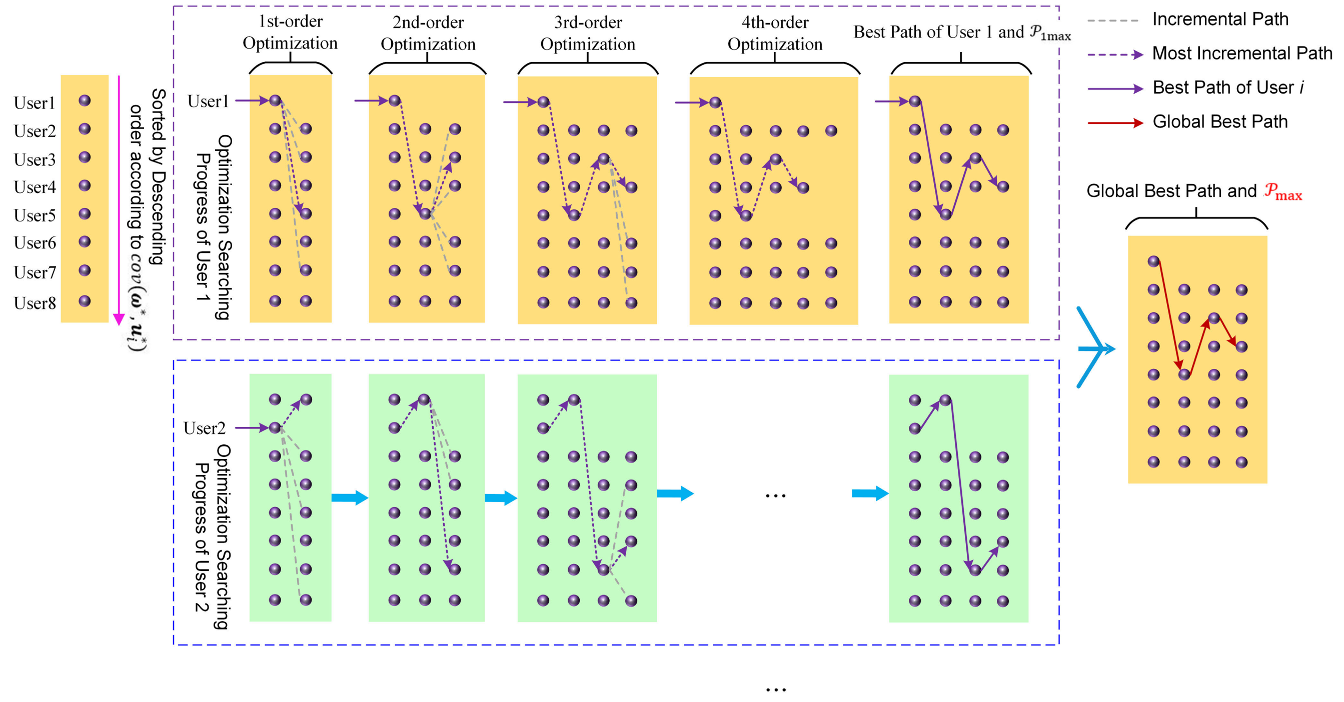

1st-order Optimization
2nd-order Optimization
3rd-order Optimization
4th-order Optimization
Best Path of User 1 and $\mathcal{P}_{1\max}$
Incremental Path
Most Incremental Path
Best Path of User i
Global Best Path
Global Best Path and $\mathcal{P}_{\max}$
User1
User2
User3
User4
User5
User6
User7
User8
Sorted by Descending order according to $cov(\boldsymbol{\omega}^*, \boldsymbol{u}_i^*)$
Optimization Searching Progress of User 1
User1
User2
Optimization Searching Progress of User 2
...
...

FIGURE 3. **Optimization searching process of NTLC method.**

denote as $\mathcal{P}^*_{i\max}$ By searching every $\mathcal{P}^*_{i\max}$, the optimal global solution $\mathcal{P}^*_{\max}$ can be easily obtained.

We propose a searching approach illustrated in Figure 3 to seek the $\mathcal{P}^*_{i\max}$ of user i. Corresponding procedures are as follows.

1) FIRST-ORDER OPTIMIZATION

If j belongs to $(\mathcal{A}-\{i\})$ and $\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_i^*) \le \mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_i^* + \tilde{\boldsymbol{u}}_j^*)$, the corresponding j is denoted as j_1, and $i \to j_1$ is the first-order incremental path (IP) of user i. The j_1 that causes the $\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_i^* + \tilde{\boldsymbol{u}}_{j_1}^*)$ to increase the most is denoted as $j_{1\max}$. The corresponding IP $i \to j_{1\max}$ refers to the first-order maximum incremental path (MIP) of user i. The goal of the first-order optimization is to find $i \to j_{1\max}$.

2) SECOND-ORDER OPTIMIZATION

Find the user $j_{2\max}$ that makes $\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_i^* + \tilde{\boldsymbol{u}}_{j_{1\max}}^* + \tilde{\boldsymbol{u}}_{j_2}^*)$ increase the most based on $\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_i^* + \tilde{\boldsymbol{u}}_{j_{1\max}}^*)$. The goal of the second-order optimization is to find $i \to j_{1\max} \to j_{2\max}$ based on $i \to j_{1\max}$.

3) k-ORDER OPTIMIZATION

On basis of $i \to j_{1\max} \to \ldots \to j_{(k-1)\max}$, find $i \to j_{1\max} \to \ldots \to j_{(k-1)\max} \to j_{k\max}$.

4) ENDING CONDITION

After k-order optimization, if $k = m+n-1$, which means that the search process has already traversed all users, or if there is no IP when conducting $(k+1)$-order optimization, which means that the covariance no longer increases, the search process is terminated.

Once the ending condition is triggered, the optimization search is terminated, and $\mathcal{P}^*_{i\max}$ is composed of the users along the MIP of every order, that is,

$$\mathcal{P}^*_{i\max} = \{i, j_{1\max}, j_{2\max}, \ldots, j_{k\max}\} \tag{20}$$

It is worthwhile to explain that the search for IP of every order ensures that the obtained $\mathcal{P}^*_{i\max}$ possesses the increasing trend, and the search for MIP of every order ensures that the covariance increases the most. Thus, the $\mathcal{P}^*_{i\max}$ sought by the procedure above is guaranteed to be the best solution for (20).

For every user i, the corresponding $\mathcal{P}^*_{i\max}$ is determined according to the search procedure above. The global solution $\mathcal{P}^*_{\max}$ is the $\mathcal{P}^*_{i\max}$ with the greatest $\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}_{\mathcal{P}^*_{i\max}})$. Once the $\mathcal{P}^*_{\max}$ is obtained, the detection for FRETs is accomplished.

B. PROBLEMS IN PRACTICE

Before introducing the detection framework of the proposed method, two practical problems must be considered. The first problem is the computational complexity of the search algorithm. From the search progress, the computational complexity depends on two factors: the number of users to be detected and the ending order when conducting the search for $\mathcal{P}^*_{i\max}$ for every user i. The number of users was unregulated. Therefore, to increase the serviceability of the data scales, it is only possible to control the ending order when conducting a search for $\mathcal{P}^*_{i\max}$. We slack the ending condition in Section III by setting an optional cut-off order K. The ending condition is adjusted as follows: After k-order optimization, if $k = K-1$, or if there is no IP when conducting $(k+1)$-order optimization, the search process is terminated. Compared with the original ending condition, it cuts off the following searching process for $\mathcal{P}^*_{i\max}$ by setting K to a suitable value, which saves a lot of time, especially when conducting a search $\mathcal{P}^*_{i\max}$ for benign users. A comparison between our search algorithm (with and without K) and other detection methods in terms of computational complexity is presented in Section IV-D.

On the other hand, fraudulent customers may steal electricity at non-fixed ratios. When an area contains a certain number of non-FRETs, the proposed method may be invalid. To address this problem, we set up a scene judgment part to block our method in cases where fraudulent users were not FRETs. Because $\tilde{\boldsymbol{u}}^*_i\big|_{i\in\mathcal{C}}$ and $\boldsymbol{\omega}^*$ become increasingly less satisfying the linear equations in (10) as the number of non-FRETs increases, the linear correlation between $\boldsymbol{\omega}^*$ and the combined load vector $\tilde{\boldsymbol{u}}^*_{\mathcal{P}^*_{\max}} = \sum_{i\in\mathcal{P}^*_{\max}} \tilde{\boldsymbol{u}}^*_i$ search by our method should be increasingly weaker. This can be verified by the value variations of $\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}^*_{\mathcal{P}^*_{\max}})$ and $\rho(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}^*_{\mathcal{P}^*_{\max}})$ in Table 2, which provides the basis for scene judgment. Because the value of PCC is fixed between -1 and 1, $\rho(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}^*_{\mathcal{P}^*_{\max}})$ is more suitable as the standard for scene judgement. After $\mathcal{P}^*_{\max}$ is found, calculate $\rho(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}^*_{\mathcal{P}^*_{\max}})$ and compare it with the threshold θ. When $\rho(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}^*_{\mathcal{P}^*_{\max}}) > \theta$, we believe that fraudulent users are mainly FRETs, and $\mathcal{P}^*_{\max}$ is given as the suspicious user set $\mathcal{C}_{\mathrm{sus}}$; when $\rho(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}^*_{\mathcal{P}^*_{\max}}) \leq \theta$, we believe that this is not the target scene, and the void set $\emptyset$ is given as $\mathcal{C}_{\mathrm{sus}}$ to block the method. After numerous experiments, it is suggested that the value of θ should be set as 0.96-0.98 to achieve a favorable result in the real unknown scene.

P.S. The total number of users in an area is 80, and the total number of electricity thieves in per area is 8.

TABLE 2. $\mathrm{cov}(\omega^*, \tilde{u}^*_{\mathcal{P}\mathrm{max}})$ **and** $\rho(\omega^*, \tilde{u}^*_{\mathcal{P}\mathrm{max}})$ **with different numbers of non-FRETs per area.**

Non-FRETs	0	1	2	3	4	5	6	7	8
$\mathrm{cov}(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}^*_{\mathrm{P_{max}}})$	0.116	0.094	0.077	0.063	0.056	0.049	0.043	0.036	0.032
$\rho(\boldsymbol{\omega}^*, \tilde{\boldsymbol{u}}^*_{\mathrm{P_{max}}})$	0.999	0.981	0.964	0.931	0.903	0.877	0.864	0.853	0.848

Algorithm of NTLC

```
Input:   User set $\mathcal{A}$, normalized load vectors $\tilde{\boldsymbol{u}}^*_i\big|_{i\in\mathcal{A}}$, NTL
         vector $\boldsymbol{\omega}$ and parameters $K$ and $\theta$.
Output:  Set of suspicious users $\mathcal{C}_{\mathrm{sus}}$.
Step1:   For every user $i$ in $\mathcal{A}$:
           Set the best solution of the sub-problem of $i$ as
           $\mathcal{P}^*_{i\max} = \{i\}$;
           Do:
             For every user $j$ in $\mathcal{A}$:
               Calculate $\mathrm{cov}(\boldsymbol{\omega}, \tilde{\boldsymbol{u}}^*_j + \sum_{h\in\mathcal{P}^*_{i\max}} \tilde{\boldsymbol{u}}^*_h)$;
               If $\mathrm{cov}(\boldsymbol{\omega}, \tilde{\boldsymbol{u}}^*_j + \sum_{h\in\mathcal{P}^*_{i\max}} \tilde{\boldsymbol{u}}^*_h)$
               $-\mathrm{cov}(\boldsymbol{\omega}, \sum_{h\in\mathcal{P}^*_{i\max}} \tilde{\boldsymbol{u}}^*_h) > 0$:
                 Add user $j$ into $\mathcal{P}^*_{i\max}$;
           While the ending condition is not triggered;
Step2:   Set the best solution of problem (18) as $\mathcal{P}^*_{\max} = \emptyset$;
         For every user $i$ in $\mathcal{A}$:
           If $\mathrm{cov}(\boldsymbol{\omega}, \sum_{h\in\mathcal{P}^*_{i\max}} \tilde{\boldsymbol{u}}^*_h) > \mathrm{cov}(\boldsymbol{\omega}, \sum_{h\in\mathcal{P}^*_{\max}} \tilde{\boldsymbol{u}}^*_h)$:
             $\mathcal{P}^*_{\max} = \mathcal{P}^*_{i\max}$;
Step3:   Calculate $\rho(\boldsymbol{\omega}, \sum_{h\in\mathcal{P}^*_{\max}} \tilde{\boldsymbol{u}}^*_h)$;
         If $\rho(\boldsymbol{\omega}, \sum_{h\in\mathcal{P}^*_{\max}} \tilde{\boldsymbol{u}}^*_h) > \theta$:
           $\mathcal{C}_{\mathrm{sus}} = \mathcal{P}^*_{\max}$;
         Else:
           $\mathcal{C}_{\mathrm{sus}} = \emptyset$;
```

C. DETECTION FRAMEWORK

Based on the above principle and methodology, the corresponding detection framework is designed with three modules: preprocessing, detection, and judgment, as shown in Figure 4.

For an area that contains $m+n$ different users with their M-day load profiles, the preprocessing module first vectorizes the load profiles each day to obtain the daily load vectors $\tilde{\boldsymbol{u}}_{i,d}$ For every $\tilde{\boldsymbol{u}}_{i,d}$, and the missing data are recovered as follows:

$$G(\tilde{u}_{i,d,t}) = \begin{cases} \mathrm{mean}(\tilde{\boldsymbol{u}}_{i,d}), & \text{when } \tilde{u}_{i,d,t} \in \mathrm{NaN} \\ \tilde{u}_{i,d,t}, & \text{otherwise} \end{cases} \tag{21}$$

where $\mathrm{mean}(\tilde{\boldsymbol{u}}_{i,d})$ is the average value of $\tilde{\boldsymbol{u}}_{i,d}$. In addition, there were some individual erroneous data under certain conditions. Therefore, the preprocessing module also recovers those data by the following equation according to "three-sigma rule of thumb".

$$G(\tilde{u}_{i,d,t}) = \begin{cases} \dfrac{\tilde{u}_{i,d,t-1} + \tilde{u}_{i,d,t+1}}{2}, & \text{if } \tilde{u}_{i,d,t} > 3\sigma(\tilde{\boldsymbol{u}}_{i,d}), \\ & \tilde{u}_{i,d,t-1}, \tilde{u}_{i,d,t+1} \neq \mathrm{NaN} \\ \tilde{u}_{i,d,t}, & \text{otherwise} \end{cases} \tag{22}$$

where $\sigma(\tilde{\boldsymbol{u}}_{i,d})$ is the standard deviation of $\tilde{\boldsymbol{u}}_{i,d}$. Next, the daily NTL vector $\boldsymbol{\omega}_d$ is calculated using (3). Subsequently, all vectors are standardized to obtain $\tilde{\boldsymbol{u}}^*_{i,d}$ and $\boldsymbol{\omega}^*_d$ according to (9).

The detection module uses $\tilde{\boldsymbol{u}}^*_{i,d}$ and $\boldsymbol{\omega}^*_d$ on the same day as the input. For every d, its suspect set $\mathcal{C}_{\mathrm{sus},d}$ is calculated according to NTLC algorithm. Then, the judgment module calculates the anomaly degree γ_i of user i by counting the number of times M_i that user i belongs to $\mathcal{C}_{\mathrm{sus},d}$ in M days, that is,

$$\gamma_i = \frac{M_i}{M} \tag{23}$$

From (23), the higher the γ_i is, the more suspicious user i is. Finally, a certain number of users with the highest γ_i are chosen for on-site inspection.

IV. VALIDATION AND EVALUATION

A. DATASET

1) BENIGN DATA

We utilize a realistic electricity usage dataset from the State Grid Corporation of China (SGCC) to conduct the experiments. Because all the users in this dataset are from areas whose transmission loss rates are less than 3% per month, their load data are measured synchronously and can be considered ground truth. Detailed information on this dataset is presented in Table 3. In particular, it consists of the load profiles of 3000 single-phase customers and 500 three-phase customers over 285 days (from April 1, 2019, to December 31, 2019). Every load profile per day is composed of kWh measurements every 15 min.

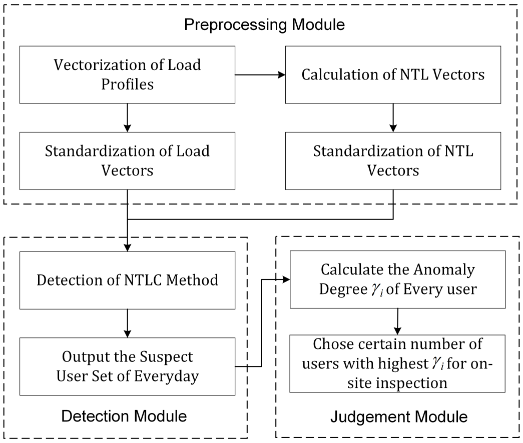

Preprocessing Module
Vectorization of Load Profiles
Calculation of NTL Vectors
Standardization of Load Vectors
Standardization of NTL Vectors
Detection of NTLC Method
Output the Suspect User Set of Everyday
Detection Module
Calculate the Anomaly Degree γ_i of Every user
Chose certain number of users with highest γ_i for on-site inspection
Judgement Module

FIGURE 4. Framework of NTLC method.

2) FRAUDULENT DATA

We rely on the electricity theft emulator (ETE) of the China Electric Power Research Institute to generate fraudulent data. This emulator, as shown in Figure 5, consists of two sectors: the load control sector and meter sector. Let us explain how fraudulent data are generated by ETE using one example. By inputting the time series of the voltage, current, and power factor according to the benign dataset, the load control sector can emulate the power usage processes of real customers. Meanwhile, smart meters in the meter sector record the data of emulated customers. If the meter is physically tampered, for example, the voltage coil of this meter is connected to a diode, the readings of this meter are the fraudulent data of such malicious user who tamper with their meters in this way. By tampering with smart meters in different ways, this emulator can precisely emulate all sorts of real physical attacks, which are far more reliable than artificially summarized FDIs. Fraudulent samples generated by this emulator are shown in Figure 6.

In this study, we mainly adopt voltage-reducing, current-decreasing and power factor diminishing these linear physical attacks to generate fraudulent data of FRETs.

B. EVALUATION METRICS AND COMPARISONS

We adopt the area under the curve (AUC) [29] and mean average precision (MAP) [30], which are two widely used classification evaluation metrics to assess the performance of the proposed method. The AUC is defined as the area under the receiver operating characteristic curve (ROC). According to [12], it can be calculated as follows:

$$\mathrm{AUC} = \frac{\sum_{i \in \mathcal{C}} \mathrm{rank}_i - 0.5m(m+1)}{m \times n} \tag{24}$$

where m is the number of fraudulent users, n is the number of benign users, and rank_i is the rank of user i in ascending order according to the anomaly degree γ_i.

MAP is typically used to estimate the quality of information retrieval. To calculate the MAP, we need to define P@S, as in (25).

$$\mathrm{P@S} = \frac{Y_s}{S} \tag{25}$$

TABLE 3. Description of the benign dataset.

Items	Description	
User types	Single-phase	Three-phase
Number of users	3000	500
Date	from April 1, 2019 to December 31, 2019	
Data type	Every load profile a day is composed of the kWh measurements of every 15 minutes.	

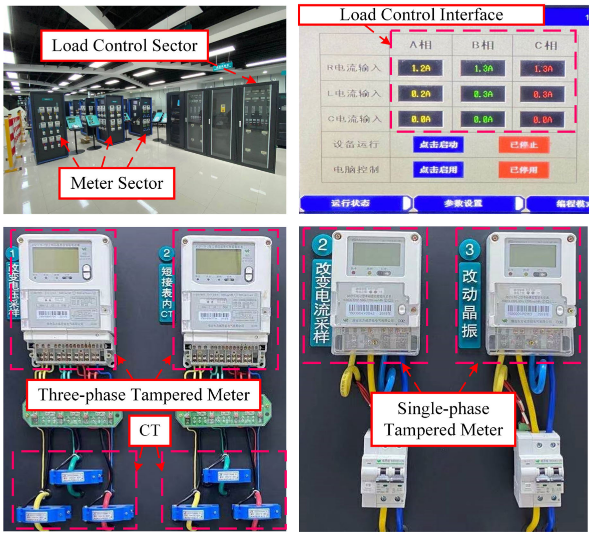


FIGURE 5. Electricity theft emulator.

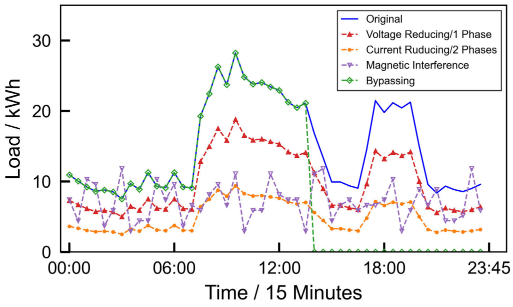


FIGURE 6. Some fraudulent samples generated by ETE.

where Y_s is the number of electricity thieves among the top S suspicious users. Given a certain number of R, MAP@R can be calculated as in (26).

$$\text{MAP@R} = \frac{\sum_{i=1}^{r} \text{P@S}_i}{r} \tag{26}$$

where r is the number of electricity thieves among the top R suspicious users, and S_i is the position of the i-th electricity thieve. It can be inferred that both AUC and MAP@R are between 0 and 1, and a higher value indicates a better performance.

To analyze the merits and demerits of the proposed method, we use some state-of-the-art methods for comparison.

1) PCC [31]: A famous bivariate correlation measurement.

2) MIC [31]: A metric that can measure strength of non-linear correlation of two vectors.

3) CFSFDP [32]: A novel clustering algorithm based on density and distance. In [17], it was adopted to calculate the anomaly degrees of users.

4) Local outlier factor (LOF) [33]: a widely used method of local-density-based outlier method.

C. PERFORMANCE OF THE DETECTION FOR FRETS

In this part, the load profiles of all 500 three-phase users from Aug.1, 2019 to Sep.31, 2019, were utilized to conduct the experiments. We randomly and evenly divided 500 three-phase users into ten areas. Six users in each area were randomly chosen as FRETs, whose load profiles were tampered with via ETE. Therefore, each area contains 50 users, and the ratio of FRETs is 12%, which is very high. The test was repeated 100 times with different combinations of areas and random selection of FRETs.

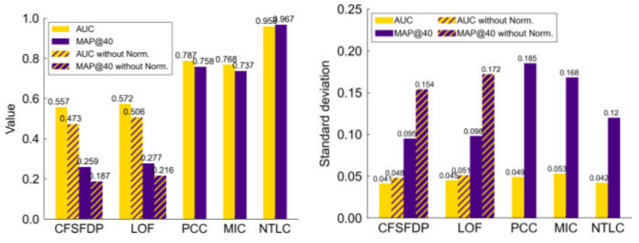
Table 4 shows the best values of AUC and MAP@40 for the five methods in the 100 experiments. Figure 7(a) shows the average scores for the five methods. To analyze the impact of normalization on the performance of power pattern-based methods in detecting FRETs, we also present the evaluation metrics of CFSFDP and LOF without normalization. It is evident that the highest MAP@40 of the power-pattern-based methods is only 0.343, whereas the lowest MAP@40 of the correlation-based methods is 0.726. At the same time, the gap in AUC between these two methods was also large. This result demonstrates that correlation-based methods are far more capable of detecting FRETs than outlier-based methods. This is because normalization makes the tampered data almost identical to the ground truth data. Therefore, the patterns of the FRETs cannot differ from the normal patterns after normalization. However, the evaluation metrics of CFSFDP and LOF dropped without normalization. It seems that non-normalization of data cannot help improve their accuracy when detecting FRETs. Furthermore, comparison among the correlation-based methods shows that both the AUC and MAP@40 of NTLC are above 0.95, which increases by nearly 25% based on PCC (0.787, 0.758) and MIC (0.768, 0.737). This indicates that the proposed method improves the performance of correlation-based methods in detecting FRETs.

Figure 7(b) shows the standard deviation σ of AUC and MAP@40 in 100 experiments for the five methods. The σ_{auc} of the 5 methods are all distributed between 0.04 and 0.06 and are not much different. On the other hand, the values of $\sigma_{\text{MAP@40}}$ are distributed between 0.09 and 0.19, in which $\sigma_{\text{MAP@40}}$ of outlier-based methods is lower than that of correlation-based methods. Nevertheless, the $\sigma_{\text{MAP@40}}$ of LLC is 0.12, which is lower than those of the other two correlation-based methods. From the experiments in this part, we can conclude that the NTLC method improves both the accuracy and stability of correlation-based methods for detecting FRETs.

D. COMPARISON OF TIME CONSUMPTION

In this part, we mainly test the time consumption of the five methods under three different data scales: 15500 load profiles, 30500 load profiles, and 76500 load profiles. To evaluate the impact of the ending order K on the time consumption and accuracy of our method in detecting FRETs, we also conducted tests on NTLC with different K values ($K = 20$, 15, 10, 5, and 3). The other experimental settings, such as the number of users in one area and the number of FRETs, are the same as in Section IV. All tests were performed on an AMD Ryzen 95900@4.7GHz desktop computer with 64GB RAM. All five methods were developed on PyCharm using Python 3.8. The average time consumption and evaluation metrics are shown in Figure 8.

Among these methods, MIC and PCC have the lowest time consumption, which is less than 25s for all data scales. The CSFDPF and NTLC without K are the most time-consuming methods. It takes NTLC 93.75s to complete the detection



(a)

(b)

FIGURE 7. Values of evaluation metrics of the five methods. (a) Average value of AUC and MAP@40. (b) Standard deviation of AUC and MAP@40.

TABLE 4. Best evaluation results of the five detection methods.

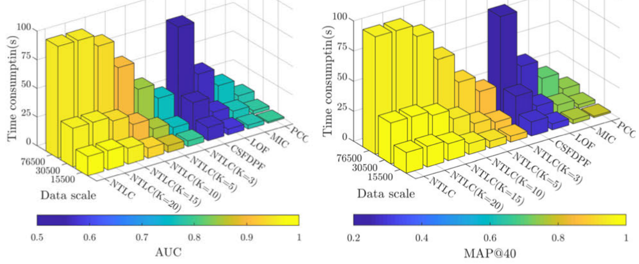
Methods	AUC	MAP@40
CFSFDP	0.593	0.301
CFSFDP without norm.	0.556	0.274
LOF	0.630	0.343
LOF without norm.	0.564	0.282
PCC	0.827	0.845
MIC	0.801	0.833
NTLC	**0.998**	**1.000**

of 76500 load profiles, which is nearly five times as long as the MIC consumption. With an increase in the data scale, the time consumption of NTLC increased geometrically. The results suggest that the proposed method indeed faces the problem of long-time consumption when handling large-scale data.

To address this problem of NTLC, we set the parameter of the ending-order K. It can be observed from Figure 8 that when K varies from 20 to 15, the time consumption of NTLC decreased not too much. This indicates that a higher K does not affect the time consumption of NTLC. However, as K decreased continuously, time consumption dropped off obviously. When $K = 5$, the time consumption of the NTLC was at the same level as that of the LOF. Meanwhile, the AUC and MAP@40 of NTLC did not decrease significantly during this process. The MAP@40 of NTLC remains above 0.85 when K was 3. The results demonstrate that we can control the time consumption of NTLC to a reasonable level (between LOF and MIC) without significant loss of accuracy when handling large-scale data by setting the value of K to a suitable level.

E. IMPACT OF NUMBER OF FRETS

To explore the impact of the number of FRETs on the above five methods, we held 50 users per area and changed the number of FRETs from 1 to 15 (step size of 2). Figure 9 shows the results of the five methods used in this study. For CFSFDP and LOF, we only present their results with normalization.

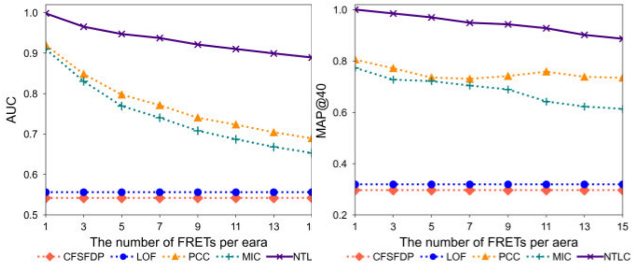


(a)

(b)

FIGURE 8. Time consumption and evaluation metrics of the five methods and NTLC with different K. (a) Time consumption and AUC with different data scales. (b) Time consumption and MAP@40 with different data scales.

The AUC and MAP@40 of the power pattern-based methods remained stable and did not change with the number of FRETs. This is the reason why the tampered load curves of FRETs are almost identical to the ground truth curves after normalization, and the outliers have already been decided from the beginning, regardless of how many users are chosen to be tampered with. In contrast, the scores of the correlation-based methods changed with the number of FRETs. To be specific, with a small number of FRETs, the three correlation-based methods have fairly closed and high AUCs. However, as the number of FRETs increased, the AUC of PCC and MIC decreased rapidly and gradually deviated from that of the LLC. Overall, the AUC damping of PCC and MIC are 0.231, from 0.92 to 0.689, and 0.251, from 0.911 to 0.654, respectively, while that of NTLC is only 0.111, from 0.998 to 0.887. The MAP@40 of these three methods exhibited the same behavior. The analysis above indicates that the number of FRETs indeed has a negative impact on the accuracy of PCC and MIC, for which these two methods are not applicable to the scenario of multiple FRETs. Although the scores of NTLC decreased slightly in the whole process, they remained at a very high level. Therefore, the proposed method performed superiorly in detecting group FRETs.



(a)

(b)

FIGURE 9. Performance of the 5 methods with different numbers of FPET users per area. (a) Values of AUC. (b) Values of MAP@40.

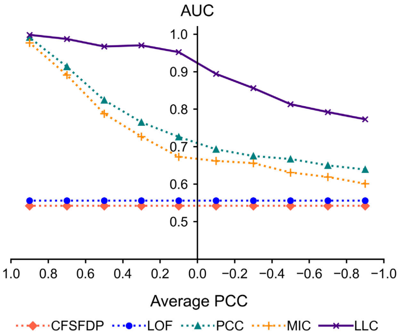


FIGURE 10. AUC values of the 5 methods with different average PCCs.

F. IMPACT OF CORRELATION OF FRETS

The analysis in Section II proves that when the load vectors of any two FRETs are positively correlated, the suspicious user set $\mathcal{C}_{\text{sus}}$ output by the LLC method is nearly identical to the set of fraudulent users $\mathcal{C}$. In this part, we explore the performance of the NTLC method when the sufficient condition is not fully met. For this purpose, we use the average PCC, which is the mean value of the PCCs between every user, to measure the correlation of a set of FRETs. Ten sets of FRETs with ten different average PCCs that were approximately evenly distributed between -1 and 1 were found. Figure 11 shows the AUCs of the five methods for detecting these 10 sets of FPET users.

The AUC of the outlier-based method behaved exactly as in previous tests for the same reason. In contrast, the curves of the correlation-based methods show that as the correlation of FPET users changes from positive to negative, the AUC of these three methods decreases differently. Further analyzing the curve of LLC, we can see that setting 0 as a boundary when the average PCC is above 0, which means that the FPET users are positively correlated, the AUC of LLC hardly decreases significantly. However, when the average PCC is below 0, which means that the FPET users are negatively correlated, the AUC reduction of LLC evidently increases. This shows that negatively correlated FPET users can affect the accuracy of the LLC method. Even so, the lowest AUC of LLC was 0.784, which significantly outperformed the other methods. Thus, the proposed method is more robust from this perspective.

V. CONCLUSION

This study proposes an NTLC-based method to detect FRETs and verifies the effectiveness of this method using a real consumption dataset provided by SGCC. The main conclusions are as follows:

1) We established a mathematical model between the NTL and load data of FRETs. We then observe that their correlation becomes increasingly strong with data superposition. The precondition for this trend is that the load data of any two FRETs are positively correlated.

2) Based on this trait, we realize the detection of FRETs and UPUs by searching a set of users who have the maximum covariance with NTL from all sets that have the above increasing trend. We also analyzed two problems of the proposed method in practice and took corresponding measures to address them.

3) We conducted a series of tests via comparisons with other state-of-the-art methods regarding accuracy, stability, and scalability. The results indicate that the proposed method has superior and stable performance in detecting FRETs, particularly when it comes to multiple FRETs in one area. In addition, the parameter K is able to reduce the time consumption of the presented method without too much loss of accuracy. In addition, the performance of the proposed method when the precondition is not fully met was also tested.

There are also some limitations to the proposed method. First, the proposed method analyzes only electricity consumption data, which may contain limited information. In addition to meter reading data, other information such as climatic factors (temperature), regional factors, and electric factors (current and voltage) are worth studying in the future. Second, our method does not specialize in detecting nonlinear FDI, which has also been adopted by cyber-attacks. Therefore, it is worthwhile to investigate ways to supplement the detection of nonlinear FDI with multiple types of information. However, it is the fact that there is no one-fit-all solution to handle all sorts of FDIs, the NTLC method is of high value to power utilities for easing the severity of collaborative fraud.

REFERENCES

[1] S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, "Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft," *Energy Policy*, vol. 39, no. 2, pp. 1007–1015, Feb. 2011.

[2] N. G. LLC. (May 2017). *Electricity Theft and Non-Technical Losses: Global Markets, Solutions, and Vendors*. [Online]. Available: http://www.northeast-group.com/reports/Brochure-Electricity%20Theft%20&%20Non-Technical%20Losses%20-%20Northeast%20Group.pdf

[3] J. Dou, X. Liu, and J. Lu, "Research on electricity anti-stealing method based on power consumption information acquisition and big data," *Electr. Meas. Instrum.*, vol. 55, no. 21, pp. 43–49, 2018.

[4] T. Tribune. (2021). *Rs, 1200 Cr Power Stolen Every Year in Punjab, Villages Account for 66% Theft*. [Online]. Available: https://www.tribuneindia.com/news/patiala/rs-1-200-cr-power-stolen-every-year-in-punjab-villages-account-for-66-theft-279647

[5] T. I. Express. (2019). *Delhi: At This Najafgarh Village, It's Hard to Pull the Plug on Electricity Theft*. [Online]. Available: https://www.indianexpress.com/article/cities/delhi/at-this-najafgarh-village-in-delhi-its-hard-to-pull-the-plug-on-electicity-theft-6054274

[6] F. Daily. (2019). *A Serious Case of Group Electricity Theft in Fuzhou! 310 Malicious Users are Captured!* [Online]. Available: https://xw.qq.com/cmsid/20180920A1D8LT/20180920A1D8LT00

[7] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3125–3148, May 2019.

[8] Q. Chen. K. Zheng, C. Kang, and F. Huangfu, "Detection methods of abnormal electricity consumption behaviours: Review and prospect," *Autom. Electr. Power Syst.*, vol. 42, no. 17, pp. 189–199, 2018.

[9] A. A. Cardenas, S. Amin, G. Schwartz, R. Dong, and S. Sastry, "A game theory model for electricity theft detection and privacy-aware control in AMI systems," in *Proc. 50th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2012, pp. 1830–1837.

[10] S. Amin, G. A. Schwartz, A. A. Cardenas, and S. S. Sastry, "Game theoretic models of electricity theft detection in smart utility networks: Providing new capabilities with advanced metering infrastructure," *IEEE Control Syst.*, vol. 35, no. 1, pp. 66–81, Feb. 2015.

[11] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216–226, Jan. 2016.

[12] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1606–1615, Apr. 2018.

[13] S. V. Oprea and A. Bara, "Machine learning classification algorithms and anomaly detection in conventional meters and Tunisian electricity consumption large datasets," *Comput. Electr. Eng.*, vol. 94, pp. 1–17, Sep. 2021.

[14] M. Ismail, M. Shahin, M. Naidu, and E. Serpedin, "Deep learning detection of electricity theft cyber-attacks in renewable distributed generation," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3428–3437, Jul. 2020.

[15] T. Hu, Q. Guo, H. Sun, T. E. Huang, and J. Lan, "Nontechnical losses detection through coordinated BiWGAN and SVDD," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 1866–1880, May 2021.

[16] L. Tian and M. Xiang, "Abnormal power consumption analysis based on density-based spatial clustering of applications with noise in power systems," *Autom. Electr. Power Syst.*, vol. 41, no. 5, pp. 64–70, Mar. 2017.

[17] K. Zheng, Y. Wang, Q. Chen, and Y. Li, "Electricity theft detecting based on density-clustering method," in *Proc. IEEE Innov. Smart Grid Technol.-Asia (ISGT-Asia)*, Dec. 2017, pp. 1–6.

[18] J. B. Leite and J. R. S. Mantovani, "Detecting and locating non-technical losses estimation in modern distribution system," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 1023–1032, Mar. 2018.

[19] L. M. R. Raggi, F. C. L. Trindade, V. C. Cunha, and W. Freitas, "Non-technical loss identification by using data analytics and customer smart meters," *IEEE Trans. Smart Grid*, vol. 35, no. 6, pp. 2700–2909, Dec. 2020.

[20] Y. Gao, B. Foggo, and N. Yu, "A physically inspired data-driven model for electricity theft detection with smart meter data," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5076–5088, Sep. 2019.

[21] G. M. Messinis, A. E. Rigas, and N. D. Hatziargyriou, "A hybrid method for non-technical loss detection in smart distribution grids," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6080–6091, Nov. 2019.

[22] S. Salinas, M. Li, and P. Li, "Privacy-preserving energy theft detection in smart grids: A P2P computing approach," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 257–267, Sep. 2013.

[23] 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. 2019.

[24] P. P. Biswas, H. Cai, B. Zhou, B. Chen, D. Mashima, and V. W. Zheng, ''Electricity theft pinpointing through correlation analysis of master and individual meter readings,'' *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3031–3042, Jul. 2020.

[25] A. Takiddin, M. Ismail, U. Zafar, and E. Serpedin, ''Robust electricity theft detection against data poisoning attacks in smart grids,'' *IEEE Trans. Smart Grid*, vol. 12, no. 3, pp. 2675–2684, May 2021.

[26] D. N. Nikovski and Z. Wang, ''Method for detecting power theft in a power distribution system,'' U.S Patent 20 140 236 506 A1, Aug. 21, 2014.

[27] P. S. N. Rao and R. Deekshit, ''Energy loss estimation in distribution feeders,'' *IEEE Trans. Power Del.*, vol. 21, no. 3, pp. 1092–1100, Jul. 2006.

[28] L. Chen, P. O. Arambel, and R. K. Mehra, ''Estimation under unknown correlation: Covariance intersection revisited,'' *IEEE Trans. Autom. Control*, vol. 47, no. 11, pp. 1879–1882, Nov. 2002.

[29] J. Davis and M. Goadrich, ''The relationship between precision-recall and ROC curves,'' in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 233–240.

[30] A. Turpin and F. Scholer, ''User performance versus precision measures for simple search tasks,'' in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2006, pp. 11–18.

[31] D. N. Reshef, Y. A. Reshef, H. K. FinucaneSharon, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, ''Detecting novel associations in large data sets,'' *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[32] A. Rodriguez and A. Laio, ''Clustering by fast search and find of density peaks,'' *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[33] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, ''LOF: Identifying density-based local outliers,'' *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, Jun. 2000.

YINING YANG received the B.E. and M.E. degrees in electrical and automation from the School of Electrical and Electronics Engineering, North China Electric Power University, Beijing, China, in 2016 and 2019, respectively.

She is currently an Assistant Engineer with China Electric Power Research Institute. Her research interests include electricity theft detection and measurement instrumentation.

RUNAN SONG received the M.E. degree in electrical and automation from the School of Electrical and Electronics Engineering, Tianjin University, Tianjin, China, in 2020.

She is currently an Assistant Engineer with China Electric Power Research Institute. Her research interests include electricity theft detection and measurement instrumentation.

YANG XUE received the M.E. degree in high voltage and insulation from the School of Electrical and Electronics Engineering, North China Electric Power University, Beijing, China, in 2011.

He is currently a Senior Engineer with China Electric Power Research Institute. His research interests include measurement instrumentation fault diagnosis and electricity theft detection.

PENGHE ZHANG received the Ph.D. degree in high voltage and insulation technology from the School of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2012.

She is currently a Professor-Level Senior Engineer with the China Electric Power Research Institute. Her research interests include measurement instrumentation fault diagnosis and electricity theft detection.

YUEJIE XU received the B.E. degree in electrical and automation from the School of Electrical and Electronics Engineering, North China Electric Power University, Beijing, China, in 2020, where he is currently pursuing the master's degree with the School of Electrical and Electronics Engineering. His research interests include electricity theft detection and electricity data mining.

JINPING KANG (Member, IEEE) received the B.E. degree in power system and automation from the Taiyuan University of Technology, Taiyuan, China, in 1996, and the M.E. and Ph.D. degrees in electric machines and apparatus from North China Electric Power University (NCEPU), Beijing, China, in 1999 and 2010, respectively.

She is currently an Associate Professor with the School of Electrical and Electronic Engineering, NCEPU. Her research interests include non-linear model and parameters of electrical machines, wireless power transfer, and data mining in power systems.

HAISEN ZHAO (Senior Member, IEEE) received the B.E. degree in agricultural electrification and automation from the Agriculture University of Hebei, Baoding, China, in 2004, and the M.E. and Ph.D. degrees in electric machines and apparatus from North China Electric Power University (NCEPU), Beijing, China, in 2007 and 2011, respectively.

He is currently an Associate Professor with the School of Electrical and Electronic Engineering, NCEPU. Moreover, he was a Visiting Scholar with the Energy Systems Research Laboratory, Department of Electrical and Computer Engineering, Florida International University (FIU), Miami, FL, USA, from December 2018 to December 2019. He has authored and coauthored more than 100 articles in peer-reviewed journals and major international conferences. Furthermore, he has 20 patents awarded in the area of electric machine design, control, and energy-saving technologies. His research interests include electrical motors design, diagnosis, energy analysis, and energy-saving technologies of electric machines and drive systems, wireless power transfer, and data mining in power systems.