Reducing Annotation Efforts in Electricity Theft Detection through Optimal Sample Selection

Liao, Wenlong; Bak-Jensen, Birgitte; Pillai, Jayakrishnan Radhakrishna; Xia, Xiaofang; Ruan, Guangchun; Yang, Zhe

Published in:
I E E E Transactions on Instrumentation and Measurement

DOI (link to publication from Publisher):
10.1109/TIM.2024.3352696

Publication date:
2024

Document Version
Accepted author manuscript, peer reviewed version

Reducing Annotation Efforts in Electricity Theft Detection through Optimal Sample Selection

Wenlong Liao, Birgitte Bak-Jensen, Jayakrishnan Radhakrishna Pillai, Xiaofang Xia, Guangchun Ruan, and Zhe Yang

***Abstract*—Supervised machine learning models are receiving increasing attention in electricity theft detection due to their high detection accuracy. However, their performance depends on a massive amount of labeled training data, which comes from time-consuming and resource-intensive annotations. To maximize model performance within a limited annotation budget, this paper aims to reduce the annotation effort in electricity theft detection through optimal sample selection. In particular, a general framework and three new strategies are proposed to select the most valuable and representative samples from different perspectives, including uncertainty, class imbalance, and diversity of samples. In-depth simulations and analyses are conducted to evaluate the effectiveness of the proposed strategies on commonly used machine learning models and a real-world dataset. Simulation results show that the proposed strategies significantly outperform baselines on datasets of different sizes and fraudulent ratios. Besides, the proposed strategies are effective in improving detection performance across a range of classifiers.**

***Index Terms*—Electricity theft, Smart grid, Machine Learning, Data annotation, Sample selection**

I. Introduction

ELECTRICITY theft detection is the process of identifying and preventing illegal consumption of electricity, which poses significant risks to power systems, including revenue losses, equipment damage, and increased operational costs. Therefore, electricity theft detection is important to ensure the reliability and sustainability of power distribution systems, and reduce financial losses [1].

Traditionally, the detection of electricity theft typically involves dispatching technical personnel or employing video surveillance methods, which are evidently time-consuming and labor-intensive [2]. In recent years, the widespread deployment of advanced metering infrastructure (AMI) in smart grids has made the analysis of historical grid meter data the mainstream approach to detect electricity theft [3]. The existing methods based on measurement data can be divided into three categories [4]: state-based methods, game theory-based methods, and machine learning-based methods.

State-based methods are types of approaches to detect electricity theft in low-voltage power distribution systems by utilizing additional measurements (e.g., current and voltage) beyond electricity consumption reading [5]. These methods leverage the inability of fraudulent users to manipulate measurements of power distribution systems, thereby creating conflicts between the system states and smart meter records, enabling the detection of electricity theft with high accuracy. However, to implement these methods, knowledge of the network topology and additional meters are required. In some cases, specific instruments or metering devices may be installed to address the issues of electricity theft. For instance, the work in [6] designs a new current ammeter to detect electricity theft in low-voltage loads, which allows the technician to make a comparison between local loggers and remote loggers. While these methods have demonstrated their effectiveness at the substation-level detection, they may not be as applicable at the end-user level due to additional costs and installation difficulties associated with some types of devices [6].

Game theory-based methods are considered as effective ways to detect electricity theft by modeling manipulation behaviors as a game between fraudulent users and electric utilities, which is a strong assumption [7]. The primary objective of these methods is to identify the Nash equilibrium for this game. For example, the work in [8] designs a game-theoretic model to discover power theft by analyzing distributional differences. Compared to state-based methods, game theory-based methods are relatively less costly. However, it is often challenging to find a suitable equation to explain the relationship between fraudulent users and electric utilities in this game.

The machine learning-based methods focus on identifying fraudulent users by analyzing their historical bills and usage patterns. Machine learning-based methods have become increasingly popular in recent years due to their low cost in acquiring historical billing or electricity consumption readings. The machine learning-based methods can be classified into unsupervised and supervised learning methods based on the availability of labeled data [4]. The unsupervised learning methods deal with discovering patterns and structures in unlabeled data, while the supervised learning methods involve learning a mapping from inputs to outputs by using labeled data.

Unsupervised electricity theft detection methods operate on the principle of identifying power profiles that deviate

This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant 62372360.

Wenlong Liao is with Wind Engineering and Renewable Energy Laboratory, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne 1015, Switzerland.

Birgitte Bak-Jensen, Jayakrishnan Radhakrishna Pillai are with the AAU Energy, Aalborg University, Aalborg 9220, Denmark.

Xiaofang Xia is with the School of Computer Science and Technology, Xidian University, Xi'an 710071, China.

Guangchun Ruan is with the Laboratory For Information & Decision Systems, Massachusetts Institute of Technology, Cambridge 02139, United States.

Zhe Yang (Corresponding author) is with Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong (email: zhe1yang@polyu.edu.hk).

significantly from the normal behavior of users, without the need for labeled training data. For instance, the authors in [9] employ the fuzzy c-means clustering model to differentiate normal and fraudulent users. The work in [10] proposes three indexes to measure the anomalies of electricity consumption readings, and then identifies fraudulent users by combining the clustering method and the sample-to-subsample decomposition method. In [11], a neural network is used to predict power consumption, and anomalies are identified based on whether the predicted values fall within a reasonable range. Although such methods are more scalable and applicable to a wider range of locations, they suffer from the limitations of high false-positive rates and difficulties in determining the root cause of the detected anomalies, as they ignore the prior knowledge (i.e., samples with labels). Therefore, the performance of these algorithms is relatively limited.

The supervised electricity theft detection methods involve learning a model from labeled training data to differentiate between benign and fraudulent behaviors. For example, the work in [12] presents an extreme gradient boosting (XGBoost) algorithm to project the relationship between consumption readings and their labels. In [13], an improved multi-layer perception (MLP) with classic tricks (e.g., regularization and skip connection) is employed to predict instances of electricity theft. To capture the latent features from electricity consumption readings, a convolutional neural network (CNN) and its variants are proposed in [14]. Although these supervised methods typically achieve higher accuracy than the unsupervised methods, they often suffer from certain limitations, which are listed below:

- They require a large amount of labeled training samples, and may not be applicable to situations where labeled data is scarce or costly to obtain.
- They may suffer from the problem of class imbalance (i.e., the number of benign samples far exceeds fraudulent samples.), which negatively affects performance.
- The performance of the supervised electricity theft detection methods heavily relies on the quality of the training data.

To address the above limitations and improve the accuracy of supervised electricity theft detection, increasing the number of labeled data, especially fraudulent samples, is an efficient way. However, collecting labeled data from electricity consumers is time-consuming and resource-intensive, which makes it difficult to collect a large number of labeled samples in practical engineering [15]. Therefore, how to select the most valuable and representative samples for annotation deserves further investigation. One traditional method to select informative samples for annotation is random sampling, where samples are selected at random [16]. This approach can lead to the selection of inefficient or unreliable samples, as it may select uninformative or redundant samples. Another traditional method is to use clustering-based sampling, which selects samples for annotation based on their cluster membership, with the assumption that samples within the same cluster have similar properties. However, clustering-based sampling is sensitive to the choice of clustering algorithm and parameters, and may not perform well in complex and high-dimensional datasets [17].

In this context, this paper aims to reduce the annotation effort in electricity theft detection through optimal sample selection. Specifically, three new strategies are proposed to select the most valuable and representative samples from different perspectives. By applying the novel sample selection strategies to classifiers (e.g., MLP, CNN, XGBoost, etc.), this study demonstrates significant improvements in electricity theft detection performance. The selected samples effectively boost the performance of classifiers within a limited annotation budget, enhancing the overall effectiveness of the detection process.

The main innovations and contributions are summarized as follows:

- **New perspective to improve model performance**: In contrast to most works that focus on improving the model structure (i.e., model-centric perspective), this paper discusses electricity theft detection from a data-centric perspective, i.e., how to select useful samples to maximize model performance within a limited annotation budget.
- **Novel Sample Selection Strategies**: This paper introduces three innovative strategies for selecting valuable and representative samples in electricity theft detection: uncertainty-based sample annotation, fraud class-based sample annotation, and distance-based sample annotation. These strategies address different challenges, including uncertainty, class imbalance, and sample diversity.
- **Real-World Experimental Validation**: The proposed strategies and methodologies are thoroughly evaluated using actual electricity theft datasets and various classifiers. The comprehensive experimental study provides valuable insights into the advantages and limitations of each strategy, establishing their practical relevance and usefulness in electricity theft detection.

The remaining sections are organized as follows. Section II briefly introduces data preprocessing followed by data analysis, from which the need to annotate representative samples within a limited budget is derived. Section III presents a general framework and three new strategies to select the most valuable and representative samples for annotation from different perspectives. Simulation results and analysis are reported in section IV. Section V presents the discussion. Finally, section VI shows the conclusions and future work.

II. Data Pre-processing and Analysis

In this section, a brief overview of data preprocessing is presented, followed by a data analysis, which shows that annotating representative samples within a limited budget is necessary. This step is essential for the rest of the study.

A. Data Pre-processing

In raw electricity consumption data, missing values are a common issue due to a variety of reasons, including but not limited to cyber-attacks, communication blockage, and sensor malfunctions. These gaps in data can cause performance degradation. To mitigate this effect, a widely used interpolation method in [14] is employed to impute missing values in raw consumption readings. This helps to ensure that the subsequent

analysis is based on complete and reliable data.

$$F(X_i)=\begin{cases}0 & X_i \in \text{NaN}, X_{i-1} \text{ or } X_{i+1} \in \text{NaN}\\ X_i & X_i \notin \text{NaN}\\ 0.5X_{i-1}+0.5X_{i+1} & X_i \in \text{NaN}, X_{i-1} \,\&\, X_{i+1} \notin \text{NaN}\end{cases} \quad (1)$$

where X_i represents the i^{th} data point; and NaN represents a missing value.

In addition to missing values, electricity consumption readings often contain erroneous values and outliers. To mitigate their potential negative effects on detection accuracy, a common approach is to apply the three-sigma rule of thumb to identify and correct the erroneous values. This involves setting a threshold of three standard deviations from the mean to define the range of normal values, with outliers falling outside this range. This approach is widely used in practice and can be easily implemented in various applications. The following formula can be used to implement this approach:

$$F(X_i)=\begin{cases}\bar{X}+3\cdot\sigma_X & \text{If } X_i > \bar{X}+3\cdot\sigma_X\\ X_i & \text{Otherwise}\end{cases} \quad (2)$$

where $\bar{X}$ represents the mean value; and σ_X represents the standard deviation.

Data normalization is a critical step in preparing data for model training. Without normalization, the input data may have widely varying scales and ranges, which can lead to slower convergence during training and make it difficult for the model to learn the underlying patterns in the data [14]. Normalization techniques, such as min-max scaling, can help standardize the data and make it more suitable for training. The following formula can be used to implement this approach:

$$F(X_i)=\frac{X_i-X_{\min}}{X_{\max}-X_{\min}} \quad (3)$$

where $X_{\max}$ represents the maximum value; and $X_{\min}$ represents the minimum value.

B. Data Analysis

Table I provides a comprehensive overview of the widely used and solely publicly available State Grid Corporation of China (SGCC) dataset in [14], which contains electricity consumption data for 42372 household users from January 1, 2014 to October 31, 2016. Of these users, 3615 are labeled as fraudulent, and 38757 are labeled as benign. The percentage of fraudulent users is 8.53%. There are approximately 25.63% missing values in the raw dataset. Data points are recorded daily, resulting in a time resolution of 1 day and a time series length of 1035 points for each sample.

TABLE I
THE DETAIL OF THE ACTUAL SGCC DATASET

Property	Description
Time duration	January 1, 2014-October 31, 2016
Total sample count	42372
Benign sample count	38757
Fraudulent sample count	3615
Fraudulence proportion	8.53%
Sample form	1×1035
Time resolution	1 day
Unit	kWh

After preprocessing the raw dataset, two benign samples and two fraudulent samples are randomly selected for visual analysis, as shown in Fig. 1. Note that the data presented has been normalized.

The plot reveals a significant fluctuation in the electricity consumption readings, making it difficult to identify key features to distinguish between benign and fraudulent cases. Moreover, both benign and fraudulent consumption readings (e.g., benign sample 1 and fraudulent sample 1) may exhibit a valley, which could be attributed to various factors, such as electricity theft, changes in consumption patterns, and missing data. These findings underscore the need for expert knowledge and manual inspection in the annotation process, which can be extremely time-consuming and labor-intensive.

Therefore, the conventional approach of annotating a massive number of users to improve model performance is a high-cost strategy that can be prohibitive for electricity theft detection tasks. Alternatively, annotating a small set of the most valuable and representative samples is a promising way to maximize model performance with a limited annotation budget [18].

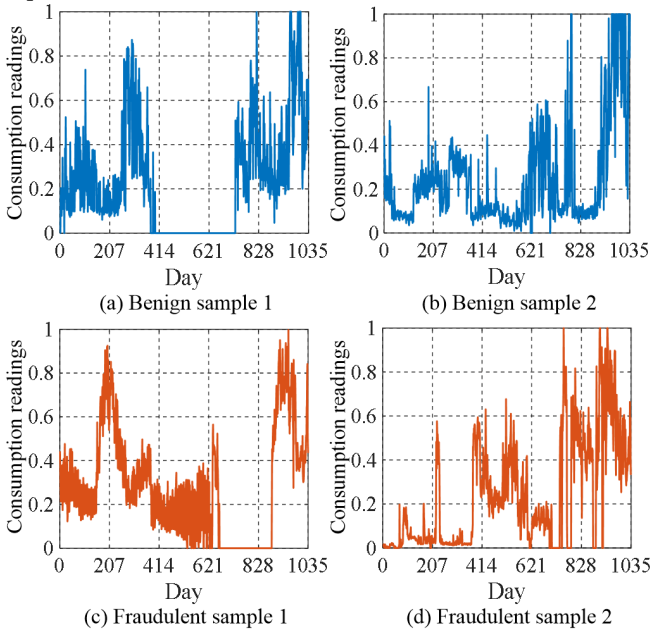

Fig. 1. The benign and fraudulent samples after data pre-processing.

III. EFFICIENT SAMPLE ANNOTATION STRATEGIES

A. A General Framework

The previous data analysis has shown that electricity theft detection usually relies on expert knowledge and manual inspection, as it is difficult to distinguish benign samples from fraudulent ones. With limited budgets, the practical application usually selects a set of the most valuable and representative samples, rather than all of them, for annotation. Fig. 2 illustrates a general sample annotation framework.

First, an efficient sample selection strategy is used to select a set of the most valuable or representative samples, which will be annotated by experts. Second, these samples are added to the annotated dataset for updating model performance. The machine learning model (e.g., MLP, CNN, XGBoost, etc.) will guide the selection of unlabeled samples. Then, the above steps repeat until the preset annotation budget or performance requirement is reached. Finally, the machine learning model will be used to detect fraudulent users. In summary, the main idea is to continuously select the most valuable samples for

annotation and update the machine learning model, so as to maximize model performance with a limited annotation budget.

In the following subsections, three new sample annotation strategies are proposed to select the most valuable and representative samples for annotation from different perspectives, including uncertainty, class imbalance, and diversity of samples.

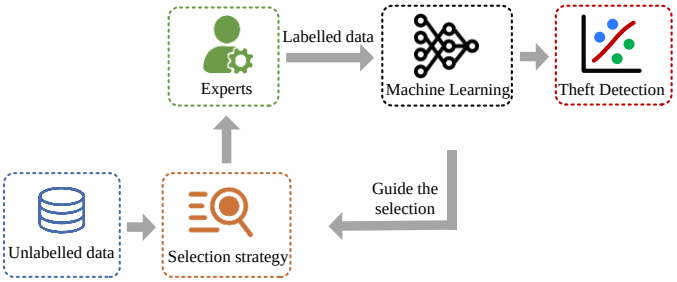

Fig. 2. A general sample annotation framework.

B. Strategy 1: Uncertainty-Based Sample Annotation

There is a highly effective and targeted learning method called error collection learning. It identifies knowledge gaps from incorrectly answered questions, and then provides targeted supplementation to improve learning performance.

Inspired by the error collection learning, this paper proposes an uncertainty-based sample (UBS) annotation method whose basic idea is to collect the most likely misclassified samples for annotation, so as to improve the model performance. As is well known, the machine learning model (e.g., MLP, CNN, XGBoost, etc.) could easily confuse the unlabeled samples near the decision boundary, as shown in Fig. 3(a). Therefore, the first strategy aims to prioritize the annotation of these hard-to-differentiate samples, from an uncertainty-oriented perspective (i.e., annotate samples with large uncertainty).

Specifically, for the binary task of electricity theft detection, an output of the machine learning model closer to 1 indicates a higher probability of the sample being fraudulent, whereas an output closer to 0 indicates a higher probability of the sample being benign. An output of the model closer to 0.5 suggests that the model is unable to accurately classify the sample. Therefore, the machine learning model can be employed to obtain output values for unlabeled samples, and then select samples with output values closest to 0.5 for annotation. The following formula can be used to implement the strategy 1:

$$S_{\mathrm{U}}(X) = 0.5 - \left| G(X) - 0.5 \right| \tag{4}$$

where $S_{\mathrm{U}}(X)$ represents the uncertainty score of the sample X; and $G(X)$ represents the probability that the sample is identified as fraudulent by the machine learning model (i.e., the output of the machine learning model). To further clarify, $G(X)$ is an indicator of the model's confidence or likelihood that a given sample X is classified as fraudulent. Specifically, after being trained, the model calculates a numerical value for $G(X)$ based on the features and contextual information of the sample X. A higher value of $G(X)$ indicates that the model is more inclined to classify the sample as fraudulent, while a lower value suggests a higher likelihood of a normal user. Normally, $G(X)$ ranges from 0 to 1. The higher the uncertainty score $S_{\mathrm{U}}(X)$ of the sample, the more valuable it is to annotate the sample.

C. Strategy 2: Fraud Class-Based Sample Annotation

The actual electricity theft dataset often suffers from highly class imbalanced problems, which severely affects the model performance. The reason is that the number of fraudulent samples is often much smaller than that of benign samples in highly class imbalanced datasets. This results in machine learning models that are biased toward predicting benign samples. Therefore, this section proposes a new method named fraudulent class-based sample (FBS) annotation, from a class imbalance-oriented perspective, to alleviate the class imbalanced problem and improve model performance by annotating samples that are predicted to be fraudulent, as shown in Fig. 3(b).

The light blue and light green circles indicate unlabeled samples. Furthermore, the light blue circles indicate samples that are judged as fraudulent by the machine learning model, and the light green circles indicate samples that are judged as benign by the machine learning model. The core idea of strategy 2 is to prioritize the annotation of the light blue circles, i.e., the samples that are judged as fraudulent by the machine learning model.

Specifically, for the binary task of electricity theft detection, the machine learning model can be employed to obtain predicted values of unlabeled samples, and then select samples with predicted values closest to 1 for annotation. The following formula can be used to implement the strategy 2:

$$S_{\mathrm{P}}(X) = G(X) \tag{5}$$

where $S_{\mathrm{P}}(X)$ represents the fraudulent score. The higher the fraudulent score of the sample, the more valuable it is to annotate.

D. Strategy 3: Distance-Based Sample Annotation

The previous strategies are conducted based on uncertainty or class imbalance of samples, while this section presents a way named distance-based sample (DBS) annotation, from a diversity-oriented perspective, to select unlabeled samples. Its principle is to compute the distance between unlabeled samples and all labeled samples, and then select unlabeled samples with a large distance for annotation. The reason is that annotating samples that are far from the training samples can enrich the diversity of the training set and thus improve the generalization of machine learning models. As shown in Fig. 3(c), unlabeled samples inside the ellipse are preferred to be annotated, as they are farther away from labeled samples compared to unlabeled samples outside the ellipse. Specifically, the DBS involves two steps:

First, the auto-encoder model, consisting of an encoder and a decoder, is trained to reduce the computational overhead and noise by extracting the low-dimensional latent features from high-dimensional samples with the following equation:

$$E_X = \mathrm{Encoder}(X), D_X = \mathrm{Decoder}(E_X) \tag{6}$$

E_X represents the low-dimensional latent feature; D_X represents the reconstructed sample; Encoder$(\cdot)$ represents the encoder; and Decoder$(\cdot)$ represents the decoder.

Both the encoder and the decoder consist of multiple dense layers. Normally, a suitable dimensionality of latent features should not be too large or too small. The dimensionality of latent features and other parameters can be determined by the

hyper-parameter optimization [19]. For example, for the SGCC dataset, the auto-encoder consists of 5 dense layers, and their numbers of neurons are 1035, 256, 64, 256, and 1035, respectively. The activation function for the first four layers is the rectified linear unit function, and the activation function for the last layer is the sigmoid function. The optimizer is the Adam algorithm with a learning rate of 0.001. The training epoch is set to 200, and the loss function used is mean squared error.

Second, the distances between the unlabeled sample and all labeled samples are calculated by using their low-dimensional latent features:

$$S_{\mathrm{Dist},i}(X) = \sqrt[2]{\sum_{k=1}^{m}\left[E_{X(i)} - E_{X(k)}\right]^2} \tag{7}$$

where $S_{\mathrm{Dist},i}(X)$ represents the distance between the i^{th} unlabeled sample and all labeled samples; $E_{X(i)}$ represents the low-dimensional latent feature of the i^{th} unlabeled sample; and m represents the number of labeled samples. Unlabeled samples with large distances are prioritized for annotation.

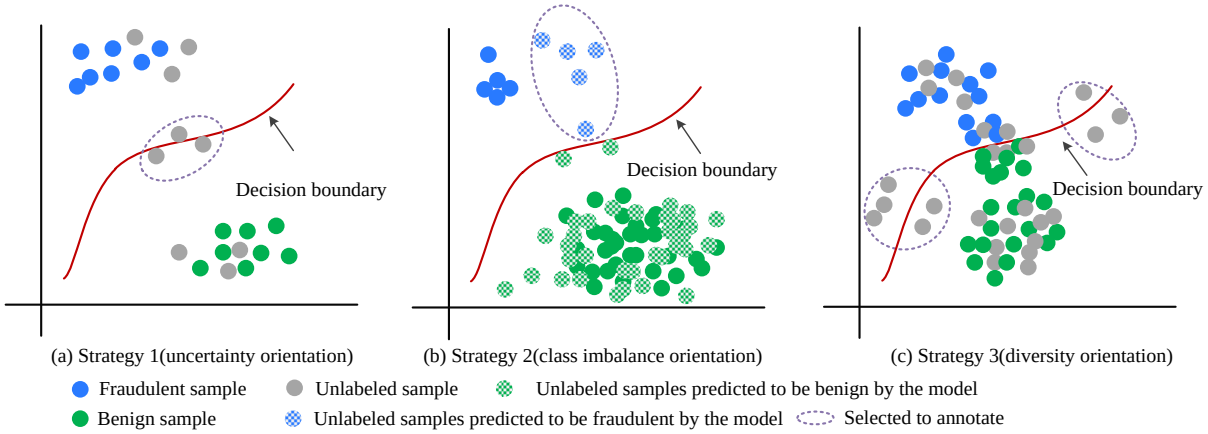

Fig. 3. A conceptual visualization of different strategies.

E. Implementation Steps

The proposed approach to annotate unlabeled samples consists of six steps. The pseudo-code is shown in Algorithm 1.

Algorithm 1: Sample Annotation

```
Input N: initial the number of labeled samples per round
Input M: initial the number of rounds
t ← 0 (Initialize the round)
for t=1,2,…, M do
    Train a classifier G and an encoder by using the training set
    Calculate S_U, S_P, S_Dist of unlabeled samples
    Training set ← labeled samples
end for
Return M×N labeled samples
```

Step 1: The parameters are initialized. For example, when the budget can support $M \times N$ annotations, the number of labeled samples N per round and the number of rounds M should be initialized. If the budget can only support a prime number of annotations, it is advisable to first transform it into a composite number before applying the proposed strategies. For example, if the budget allows for 53 annotations, one can transform it into 54, and then apply a strategy by annotating 6 samples per round for a total of 9 rounds, dropping the last sample.

Normally, a small parameter N is beneficial to improve the model performance, because the newly annotated samples improve the performance of the model, which makes it easier to find the most valuable and representative samples. However, a parameter N that is too small means that the model will be trained repeatedly, which may lead to a longer training time. Therefore, the choice of parameter N requires a combination of computational resources and model performance. A compromise parameter that balances model performance and computational resources is generally suitable.

Step 2: An auto-encoder and a classifier (i.e., machine learning model) are trained by using the labeled training set.

Step 3: A strategy is selected to annotate the unlabeled samples. For the first strategy, the pre-trained classifier is used to obtain the predicted values of the unlabeled samples, which are used to calculate the uncertainty score $S_{\mathrm{U}}(X)$ in equation (4). Other strategies can be treated in a similar manner. To reduce overfitting, the models (i.e., classifier and auto-encoder) are typically trained multiple times to obtain the mean values (e.g., the mean uncertainty score). In other words, the widely used trick called ensemble learning (also called query-by-committee) [20] is used in step 3.

Step 4: If one of the first two strategies is selected, the N unlabeled samples with the highest scores are selected for annotation. If strategy 3 is selected, the N unlabeled samples with the largest distances are selected for annotation.

Step 5: The labeled samples are added to the training set. If the M rounds have not been reached, the process returns to Step 2. Otherwise, go to step 6.

Step 6: Steps 2 to 5 are repeated M times until a certain annotation budget is reached. At this point, $M \times N$ unlabeled samples have been annotated.

In this paper, three strategies are proposed to annotate samples from different perspectives. These strategies may perform differently on different datasets. In practice, two ways can be considered to determine the most appropriate strategy for a given dataset: The first way is to use each of the three strategies to annotate a small and equal number of unknown samples. Then, the strategy that yields the most significant improvement in model performance is selected. The second way is to remove a portion of the samples from the training set and then use the three strategies to select an equivalent number of samples from the removed set. The selected samples are

reinserted into the training set. Again, the strategy that provides the greatest improvement in model performance is selected. These two methods help determine which strategy is best for a given scenario.

IV. Case Study

A. Introduction to Datasets

The SGCC dataset has been introduced in Section II [14]. To evaluate the performance of the proposed strategies on datasets with different sizes, samples are randomly selected to form three new datasets, as shown in Table II.

For example, in Dataset 2, 20% of the samples are randomly selected as the training set, 20% as the testing set, and the remaining 60% are considered as unlabeled samples.

Moreover, Dataset 1 is a small dataset, because it includes a small number of training samples. Dataset 3 is a large dataset, since it includes a large number of training samples. Dataset 2 is a medium-sized dataset.

TABLE II
The Construction of Three Datasets

Datasets	The percentage of samples		
	Training samples	Unlabeled samples	Test samples
Dataset 1	5%	75%	20%
Dataset 2	20%	60%	20%
Dataset 3	40%	40%	20%

B. Baselines and Evaluation Metrics

To demonstrate the performance of the proposed new strategies, the widely used sample selection strategies, including random sampling (RS) in [16], clustering-based sampling (CS) in [17], and density estimation-based sampling (DES) in [21], will be used as baselines for comparative analysis.

- **RS:** Unlabeled samples are randomly selected for annotation [16].
- **CS:** The clustering method (e.g., K-means) is utilized to divide unlabeled samples into multiple clusters, and then the samples closest to the cluster center in each cluster are selected for annotation [17].
- **DES:** Density estimation (e.g., Gaussian kernel density estimator) is performed on unlabeled samples, and then samples with high-density scores are selected for annotation [21].

Further, these sample selection strategies will be tested on the following popular classifiers, including MLP in [13], CNN in [14], XGBoost in [22], light gradient boosting machine (LightGBM) in [23], and random forest (RF) in [24].

The above strategies and classifiers are implemented by using the Python language with machine learning libraries, such as Tensorflow 2.0. and scikit-learn 1.2. The hyper-parameters are determined through a process of trial and error, aided by techniques, such as cross-validation and hyper-parameter optimization in [19].

To comprehensively evaluate the performance of the proposed strategies on class imbalanced electricity theft datasets, three widely adopted evaluation metrics, namely the area under curve (AUC), mean average precision (MAP), and F1 score are employed.

Specifically, AUC measures the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR) for different threshold values. The AUC ranges from 0 to 1, with a higher AUC indicating a better performance against electricity theft. In the case of electricity theft detection, the set of benign samples $\mathcal{B}$ is regarded as the negative class, while the set of fraudulent samples $\mathcal{F}$ is regarded as the positive class. The AUC represents the probability that a classifier will rank a randomly chosen fraud sample higher than a randomly chosen benign sample [10], [14]:

$$\mathrm{AUC} = \frac{\sum_{i\in\mathcal{F}} rank_i - 0.5|\mathcal{F}|\left(|\mathcal{F}|+1\right)}{|\mathcal{F}|\times|\mathcal{B}|} \tag{8}$$

where $|\mathcal{F}|$ and $|\mathcal{B}|$ are the numbers of fraudulent and benign samples, respectively.

MAP is the average precision calculated at various recall levels, evaluating the model's ability to rank fraudulent samples higher than benign. The precision of the top k electricity thieves is denoted as $P@k=Y_k/k$, where Y_k represents the number of fraudulent samples that rank in the top k. The MAP can be calculated as the average value of $P@k$ for a given number $\mathcal{D}$. In other words, $\mathrm{MAP}@\mathcal{D}$ is computed by taking the mean of the precision values $P@k$ for k=1 to $k=\mathcal{D}$.

$$\mathrm{MAP}@\mathcal{D} = \frac{1}{r}\sum_{i=1}^{r} P@k_i \tag{9}$$

where r represents the number of fraudulent samples that belong to the top N; and k_i represents the position of the i^{th} individual within the group of fraudulent samples.

As in related work [10], [14], MAP@100 and MAP@200 are used to evaluate model performance on the SGCC dataset. MAP@100 is a variant of MAP, measuring precision at the top 100 retrieved samples, respectively. Higher MAP@100 and MAP@200 indicate better model performance, showcasing the model's enhanced capability to accurately identify electricity theft instances among top predictions.

F1 score is a metric used to evaluate the performance of a binary classification model, which combines the Precision and Recall of the model:

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

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

$$\mathrm{F1\ score} = \frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{12}$$

where TP represents the proportion of actual fraudulent samples that are correctly classified as fraudulent by the classifier. FP represents the proportion of actual benign samples but is incorrectly classified as fraudulent by the classifier. FN represents the proportion of actual fraudulent samples but is incorrectly classified as benign by the classifier.

The F1 score ranges from 0 to 1, with a higher F1 score indicating a better performance against electricity theft.

The stochasticity inherent in classifiers, particularly neural networks, results in single trial detection results that may not accurately represent the true performance of the model [25], [26]. To mitigate the effects of stochasticity and increase stability, all experiments in this paper are replicated 100 times,

after which the mean value is computed.

The remaining sections present simulation and analysis, with the following specific arrangements. Section C compares the performance of different strategies by using MLP as a classifier. Section D tests the sensitivity of proposed strategies to different dataset sizes. Section E explores the sensitivity of the proposed strategies to different fraudulent ratios. Finally, Section F analyzes the generalization of proposed strategies for different classifiers.

C. Performance Comparison with Baselines

To compare the performance of proposed strategies with baselines, MLPs are treated as classifiers, and then simulations are run on dataset 2 to obtain the average metrics as shown in Table III.

Comparing the metrics in the first row with those in the other rows, it is found that additional samples annotation can help to increase AUC, MAP@100, MAP@200, and F1 score, because these samples help the model learn rich features and prior knowledge, thus improving the generalization and model performance.

For different numbers of sample annotations, the proposed strategies (i.e., UBS, FBS, and DBS) significantly outperform the baselines (i.e., RS, CS, and DES). For example, when the number of sample annotations is 500, the MAP@100 of RS is 0.839, which is the highest among the traditional methods. Compared to the traditional RS, UBS, FBS, and DBS have increased their MAP@100 by approximately 3.10%, 9.06%, and 2.74%, respectively. Similarly, the MAP@200 of UBS, FBS, and DBS improved by roughly 2.12%, 8.20%, and 1.16%, respectively, compared to the maximum value of the traditional method.

Moreover, comparing the performance of the proposed three proposed strategies (e.g., UBS, FBS, and DBS), most metrics of the FBS are the highest for different numbers of annotated samples, indicating that FBS outperforms the other three strategies for medium-sized datasets in most cases.

D. Sensitivity Analysis on Different Dataset Sizes

In the context of practical scenarios, the effectiveness of electricity detection methods can be affected by the variance in data set size encountered [27], [28]. Such variance can lead to differences in detection accuracy and stability, and ultimately affect the reliability of detection results.

To test the sensitivity of proposed strategies and baselines to different dataset sizes, MLP is still treated as a classifier, and then simulations are run on dataset 1 (i.e., a small dataset) and dataset 3 (i.e., a large dataset) to obtain the average metrics as shown in Table IV and Table V. The number of annotations ranges from 500 to 4500. In other words, we test the model performance under conditions with few annotations and under conditions with a large number of annotations.

TABLE III
THE SIMULATION RESULTS ON DATASET 2

Strategies	500 annotations				1500 annotations				3000 annotations				4500 annotations			
	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score
Without adding annotations	0.736	0.829	0.794	0.880	0.736	0.829	0.794	0.880	0.736	0.829	0.794	0.880	0.736	0.829	0.794	0.880
RS	0.738	0.839	0.802	0.882	0.742	0.845	0.809	0.884	0.747	0.860	0.825	0.888	0.750	0.889	0.847	0.893
CS	0.739	0.839	0.805	0.882	0.737	0.827	0.794	0.880	0.743	0.866	0.827	0.888	0.746	0.873	0.834	0.890
DES	0.736	0.830	0.794	0.880	0.734	0.825	0.788	0.879	0.739	0.827	0.793	0.881	0.738	0.803	0.776	0.876
UBS	0.743	0.865	0.822	0.888	0.745	0.890	0.845	0.893	0.752	0.915	0.874	0.899	0.751	0.922	0.883	0.900
FBS	0.742	0.915	0.871	0.897	0.748	0.923	0.881	0.899	0.753	0.936	0.895	0.903	0.753	0.934	0.892	0.902
DBS	0.740	0.862	0.818	0.887	0.744	0.921	0.869	0.898	0.750	0.939	0.892	0.903	0.754	0.943	0.897	0.904

TABLE IV
THE SIMULATION RESULTS ON DATASET 1

Strategies	500 annotations				1500 annotations				3000 annotations				4500 annotations			
	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score
Without adding annotations	0.704	0.681	0.661	0.875	0.704	0.681	0.661	0.875	0.704	0.681	0.661	0.875	0.704	0.681	0.661	0.875
RS	0.717	0.734	0.710	0.881	0.722	0.783	0.752	0.886	0.727	0.813	0.784	0.889	0.726	0.852	0.784	0.893
CS	0.714	0.694	0.683	0.877	0.721	0.757	0.738	0.884	0.727	0.780	0.754	0.886	0.726	0.789	0.754	0.887
DES	0.702	0.661	0.644	0.873	0.699	0.651	0.637	0.872	0.693	0.624	0.615	0.869	0.714	0.672	0.615	0.875
UBS	0.724	0.761	0.743	0.884	0.724	0.816	0.779	0.889	0.730	0.883	0.839	0.896	0.735	0.892	0.839	0.897
FBS	0.728	0.870	0.787	0.895	0.727	0.877	0.835	0.895	0.735	0.930	0.888	0.901	0.739	0.927	0.888	0.901
DBS	0.720	0.828	0.775	0.890	0.731	0.914	0.867	0.899	0.734	0.931	0.888	0.901	0.742	0.946	0.888	0.903

TABLE V
THE SIMULATION RESULTS ON DATASET 3

Strategies	500 annotations				1500 annotations				3000 annotations				4500 annotations			
	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score
Without adding annotations	0.750	0.905	0.862	0.875	0.750	0.905	0.862	0.876	0.751	0.903	0.862	0.874	0.750	0.906	0.864	0.876
RS	0.750	0.906	0.863	0.875	0.752	0.907	0.865	0.877	0.756	0.908	0.868	0.880	0.757	0.923	0.883	0.889
CS	0.752	0.907	0.864	0.877	0.753	0.912	0.870	0.881	0.757	0.916	0.874	0.885	0.757	0.918	0.877	0.886
DES	0.750	0.906	0.863	0.876	0.749	0.904	0.860	0.874	0.753	0.896	0.857	0.872	0.755	0.915	0.872	0.883
UBS	0.752	0.914	0.871	0.881	0.756	0.922	0.880	0.887	0.761	0.930	0.888	0.895	0.761	0.928	0.888	0.894
FBS	0.754	0.943	0.898	0.899	0.758	0.939	0.897	0.898	0.759	0.944	0.901	0.902	0.760	0.941	0.899	0.901
DBS	0.752	0.914	0.867	0.881	0.755	0.926	0.879	0.890	0.758	0.939	0.895	0.899	0.761	0.943	0.900	0.903

After using the traditional DES to annotate samples, most metrics of the model are smaller than those with adding annotations. This indicates that DES degrades the mode performance. Two reasons could explain this: Firstly, the unlabeled sample set may have a high density of outliers or an uneven density distribution, which can cause the selected samples to be unrepresentative and thus affect model performance. Secondly, there may be complex correlations between samples during sample selection, which can lead to insufficient diversity in the selected samples, further negatively affecting model performance.

The proposed three strategies, namely UBS, FBS, and DBS, outperform the baselines (i.e., RS, CS, and DES) for datasets of varying sizes, as demonstrated by the performance metrics shown in Table IV and Table V. This result indicates the superior efficacy of the proposed strategies for electricity theft detection tasks, regardless of dataset sizes.

Furthermore, the performance of the proposed strategies is affected by the number of annotated samples. In particular, the FBS strategy yields the best performance when the number of annotated samples is small, such as 500 annotations, as observed on Dataset 1 and Dataset 3. On the other hand, when the number of annotated samples is large, the DBS may outperform the FBS, as shown by the maximum performance improvement achieved by DBS with 4500 annotations on Dataset 1 and Dataset 3.

E. Sensitivity Analysis on Different Fraudulent Ratios

In real-world scenarios, the accuracy and stability of electricity detection methods vary significantly due to the wide range of fraudulent ratios encountered in real-world conditions. As a result, the effectiveness of detection also varies.

To test the sensitivity of proposed strategies and baselines to different fraudulent ratios, MLP and dataset 2 are still considered as examples. Note that the fraudulent ratio of dataset 2 is set to a value between 2% and 14% in the training and test sets. Then, simulations are run to obtain the average metrics, as shown in Table VI.

Regardless of whether the fraudulent ratio is high or low, the performance of the proposed strategy is superior to that of the traditional strategy, which can be verified by comparing their metrics. For example, when the fraudulent ratio is 2%, the MAP@100 of RS is the highest among the baselines, with a value of 0.651. The proposed UBS, FBS, and DBS strategies increase the MAP@100 by approximately 20.05%, 19.83%, and 19.36%, respectively, when compared to RS. Similarly, when the fraudulent ratio is 14%, the MAP@200 of RS is the highest among the baselines, with a value of 0.877. The proposed strategies, namely UBS, FBS, and DBS, achieve approximately 1.56%, 1.86%, and 3.10% improvement in MAP@200, respectively, over the RS.

Furthermore, when the fraudulent ratio is low, the performance of the three proposed strategies is relatively similar. However, as the fraudulent ratio increases, the performance of FBS gradually outperforms the other strategies. For example, FBS achieves the highest AUC, MAP@100, MAP@200, and F1 score values, when the fraudulent ratio ranges from 6% to 14%. These results indicate that FBS should be given priority in annotating unlabeled samples for datasets with a high fraudulent ratio.

F. Sensitivity Analysis on Different Classifiers

The previous sections evaluate the effectiveness of the proposed strategies by using MLP as a classifier over different dataset sizes and fraudulent ratios. This section aims to test the generalizability of the proposed strategies across other popular classifiers, including CNN [14], RF [24], XGBoost [22], and LightGBM [23].

Dataset 2 is still considered as an example. Suppose the number of annotated samples is 1500. Then, simulations are run on dataset 2 to obtain the average metrics as shown in Table VII.

TABLE VI
THE MODEL PERFORMANCE AT DIFFERENT FRAUDULENT RATIOS

Strategies	The fraudulent ratio is 2%				The fraudulent ratio is 6%				The fraudulent ratio is 10%				The fraudulent ratio is 14%			
	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score
Without adding annotations	0.718	0.608	0.585	0.970	0.741	0.746	0.723	0.911	0.727	0.823	0.796	0.858	0.735	0.886	0.863	0.811
RS	0.742	0.651	0.631	0.972	0.741	0.762	0.737	0.913	0.730	0.836	0.808	0.862	0.738	0.902	0.877	0.819
CS	0.738	0.649	0.628	0.972	0.746	0.777	0.748	0.915	0.733	0.835	0.811	0.862	0.738	0.894	0.873	0.815
DES	0.707	0.595	0.578	0.970	0.740	0.744	0.718	0.911	0.725	0.805	0.782	0.853	0.736	0.881	0.858	0.809
UBS	0.754	0.781	0.744	0.975	0.751	0.827	0.792	0.921	0.734	0.865	0.830	0.870	0.738	0.916	0.891	0.825
FBS	0.751	0.779	0.734	0.975	0.753	0.905	0.858	0.929	0.738	0.918	0.878	0.884	0.743	0.958	0.927	0.845
DBS	0.747	0.777	0.730	0.975	0.753	0.898	0.851	0.928	0.736	0.893	0.855	0.877	0.739	0.937	0.904	0.834

TABLE VII
THE SIMULATION RESULTS OF DIFFERENT MODELS

Strategies	CNN				RF				XGBoost				LightGBM			
	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score	AUC	MAP @100	MAP @200	F1 score
Without adding annotations	0.736	0.803	0.768	0.802	0.731	0.767	0.752	0.781	0.772	0.904	0.870	0.871	0.767	0.878	0.851	0.855
RS	0.742	0.855	0.816	0.831	0.743	0.851	0.815	0.830	0.766	0.910	0.883	0.871	0.770	0.910	0.875	0.873
CS	0.741	0.817	0.782	0.811	0.742	0.804	0.791	0.805	0.770	0.914	0.893	0.875	0.757	0.905	0.882	0.864
DES	0.736	0.826	0.786	0.813	0.725	0.767	0.746	0.778	0.769	0.899	0.871	0.867	0.765	0.900	0.866	0.865
UBS	0.744	0.877	0.843	0.843	0.746	0.901	0.863	0.856	0.777	0.957	0.928	0.900	0.771	0.958	0.923	0.898
FBS	0.745	0.932	0.895	0.872	0.748	0.903	0.863	0.858	0.778	0.959	0.930	0.902	0.774	0.959	0.926	0.900
DBS	0.744	0.856	0.817	0.833	0.743	0.870	0.850	0.839	0.772	0.937	0.903	0.888	0.772	0.925	0.895	0.882

No matter which classifier is used, the performance of the proposed strategy is superior to that of the traditional strategy, which can be demonstrated by comparing their metrics. For example, when CNN is used as the classifier, the MAP@100 of RS is the highest among the baselines, with a value of 0.855. The proposed UBS, FBS, and DBS strategies increase the MAP@100 by approximately 2.57%, 9.01%, and 0.12%, respectively, when compared to RS. In a similar way, the AUC, MAP@200, and F1 score of the proposed strategies are also higher than those of the baselines.

These findings suggest that the proposed strategies are effective in improving the detection performance across a range of classifiers. In particular, the FBS strategy achieves the largest improvement in AUC, MAP, and F1 score, indicating that it is highly effective in capturing fraudulent behaviors.

V. Discussion

In this paper, three strategies are proposed to annotate samples from different perspectives. From the simulation results, FBS outperforms UBS and DBS in most scenarios, but this does not mean that the proposed UBS and DBS can be discarded. Normally, these strategies may perform differently on different datasets and cases. For example, UBS slightly outperforms FBS when the fraudulent ratio is 2%, as previously discussed in Table VI. In practice, the two ways mentioned in Section III(E) can be used to determine the most appropriate strategy for a given dataset.

VI. Conclusion

To reduce the annotation effort in electricity theft detection through optimal sample selection, a general framework and three new strategies are proposed to select the most valuable and representative samples from different perspectives. After conducting simulations on the real SGCC dataset, four key conclusions have been drawn.

1) With respect to the hyper-parameters of the proposed strategy (e.g., the number of labeled samples per round and the number of rounds), these two parameters are inversely proportional, when the total number of annotated samples is fixed. Typically, a smaller value of the former parameter facilitates the improvement of the model performance, but it also leads to an increase in the number of training iterations. Therefore, the choice of parameter requires a combination of computational resources and model performance. A compromise parameter that balances model performance and computational resources is generally suitable.

2) In terms of optimal sample selection for annotation, the proposed three strategies (i.e., UBS, FBS, and DBS) significantly outperform the baselines (i.e., RS, CS, and DES) on datasets of different sizes. In particular, FBS is the best choice for annotation on medium-sized datasets.

3) Regardless of whether the fraudulent ratio of the dataset is high or low, the performance of the proposed strategies is superior to those of traditional strategies, which can be verified by comparing their metrics. This is an indication that the proposed strategies are adaptable to complex real-world application conditions with a wide range of fraudulent ratios.

4) In comparison to traditional strategies, the proposed strategies demonstrate superior performance, no matter which classifier (e.g., CNN, RF, XGBoost, and LightGBM) is used. These findings suggest that the proposed strategies are effective in improving the detection performance across a range of classifiers.

This paper only discusses optimal unlabeled sample selection by using a single strategy. The extension work can combine different strategies to select unlabeled samples.

References

[1] R. Razavi, A. Gharipour, M. Fleury, I. Akpan, "A practical feature-engineering framework for electricity theft detection in smart grids," Applied Energy, vol. 238, pp. 481-494, Mar. 2019.

[2] I. Khan, N. Javeid, C. Taylor, K. A. A. Gamage and X. Ma, "A Stacked Machine and Deep Learning-Based Approach for Analysing Electricity Theft in Smart Grids," *IEEE Trans. Smart Grid*, vol. 13, no. 2, pp. 1633-1644, Mar. 2022.

[3] B. Tahir, A. Jolfaei and M. Tariq, "Experience-Driven Attack Design and Federated-Learning-Based Intrusion Detection in Industry 4.0," *IEEE Trans. Ind. Informat*., vol. 18, no. 9, pp. 6398-6405, Sept. 2022.

[4] 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.

[5] M. Tariq and H. Poor, "Electricity Theft Detection and Localization in Grid-Tied Microgrids," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 1920-1929, May. 2018.

[6] H. Henriques, A. Barbero, R. Ribeiro, M. Fortes, W. Zanco, O. Xavier, and R. Amorim, "Development of adapted ammeter for fraud detection in low-voltage installations," *Measurement*, vol. 56, pp. 1-7, May. 2014.

[7] Z. Yan and H. Wen, "Electricity Theft Detection Base on Extreme Gradient Boosting in AMI," *IEEE Trans. Instrum. Meas*., vol. 70, pp. 1-9, Jan. 2021.

[8] S. Amin, G. Schwartz, A. Cardenas and S. Sastry, "Game-Theoretic Models of Electricity Theft Detection in Smart Utility Networks: Providing New Capabilities with Advanced Metering Infrastructure," *IEEE Control Sys. Mag*., vol. 35, no. 1, pp. 66-81, Feb. 2015.

[9] R. Qi, J. Zheng, Z. Luo and Q. Li, "A Novel Unsupervised Data-Driven Method for Electricity Theft Detection in AMI Using Observer Meters," *IEEE Trans. Instrum. Meas*., vol. 71, pp. 1-10, Jul. 2022.

[10] L. Xu, Z. Shao, and F. Chen, "A combined unsupervised learning approach for electricity theft detection and loss estimation," *IET Energy Syst. Integr.*, vol. 5, no. 2, pp 213-227. Feb. 2023.

[11] X. Wang, Z. Yao, and M. Papaefthymiou, "A real-time electrical load forecasting and unsupervised anomaly detection framework," *Appl. Energy*, vol. 330, pp 1-15. Jan. 2023.

[12] X. Cui *et al*., "Two-Step Electricity Theft Detection Strategy Considering Economic Return Based on Convolutional Autoencoder and Improved Regression Algorithm," *IEEE Trans. Power Syst.*, vol. 37, no. 3, pp. 2346-2359, May. 2022.

[13] I. U. Khan, N. Javaid, C. J. Taylor and X. Ma, "Robust Data Driven Analysis for Electricity Theft Attack-Resilient Power Grid," *IEEE Trans. Power Syst*., vol. 38, no. 1, pp. 537-548, Jan. 2023.

[14] Z. Zheng *et al*., "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.

[15] L. Cui *et al*., "A Covert Electricity-Theft Cyber-Attack against Machine Learning-Based Detection Models," *IEEE Trans. Ind. Informat*., vol. 18, no. 11, pp. 7824-7833, Nov. 2022.

[16] G. Liu and Z. Guo, "A clustering-based differential evolution with random-based sampling and Gaussian sampling," *Neurocomputing*, vol. 205, pp. 229-246, Sept. 2016.

[17] X. Yan *et al*., "A clustering-based active learning method to query informative and representative samples" *Appl. Intell*., vol. 52, pp. 13250–13267, Feb. 2022.

[18] L. Zhu, *et al*., "Deep Active Learning-Enabled Cost-Effective Electricity Theft Detection in Smart Grids," *IEEE Trans. Ind. Informat*., early access, 2023, doi: 10.1109/TII.2023.3249212.

[19] J. Wu, X. Chen, H. Zhang, L. Xiong, H. Lei, and S. Deng, " Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization," *J. Electron. Sci. Technol*., vol. 17, no.1, pp. 26-40, Mar. 2019.

[20] M. Ganaie *et al*., "Ensemble deep learning: A review" Eng. Appl. Artif. Intell., vol. 115, pp. 1-18, Oct. 2022.

[21] F. Kamalov, “Kernel density estimation based sampling for imbalanced class distribution” *Inf. Sci*., vol.512, pp. 1192-1201, Feb. 2020.
[22] Z. Yan and H. Wen, “Electricity Theft Detection Base on Extreme Gradient Boosting in AMI,” *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-9, 2021.
[23] R. Punmiya and S. Choe, “Energy Theft Detection Using Gradient Boosting Theft Detector With Feature Engineering-Based Preprocessing,” *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2326-2329, Mar. 2019.
[24] G. Lin *et al*., “Electricity Theft Detection Based on Stacked Autoencoder and the Undersampling and Resampling Based Random Forest Algorithm,” *IEEE Access*, vol. 9, pp. 124044-124058, Sept. 2021.
[25] D. Gu, Y. Gao, K. Chen, J. Shi, Y. Li and Y. Cao, “Electricity Theft Detection in AMI With Low False Positive Rate Based on Deep Learning and Evolutionary Algorithm,” *IEEE Trans. Power Syst.*, vol. 37, no. 6, pp. 4568-4578, Nov. 2022.
[26] L. Raggi, F. Trindade, V. Cunha and W. Freitas, "Non-Technical Loss Identification by Using Data Analytics and Customer Smart Meters," *IEEE Trans. Power Del.*, vol. 35, no. 6, pp. 2700-2710, Dec. 2020.
[27] W. Liao, Z. Yang, K. Liu, B. Zhang, X. Chen and R. Song, "Electricity Theft Detection Using Euclidean and Graph Convolutional Neural Networks," *IEEE Trans. Power Syst*., vol. 38, no. 4, pp. 3514-3527, July 2023.
[28] W. Liao et al., "Simple Data Augmentation Tricks for Boosting Performance on Electricity Theft Detection Tasks," *IEEE Trans. Ind Appl*., vol. 59, no. 4, pp. 4846-4858, Jul. 2023.

Wenlong Liao received the B.S. degree from China Agricultural University in 2017. He received the M.S. degree from Tianjin University in 2020. He received the Ph.D. degree from Aalborg University in 2023. He is currently a post-doctoral researcher at the Ecole Polytechnique Federale de Lausanne (EPFL). His current research interests include smart grids, machine learning, and renewable energy.

Birgitte Bak-Jensen received the M.Sc. degree and the Ph.D. degree from the Institute of Energy Technology, Aalborg University, Aalborg, Denmark, in 1986 and 1992, respectively. She is currently a Professor of intelligent control of the power distribution system, AAU Energy, Aalborg University. Her current research interests include power quality and stability in power systems and taking integration of dispersed generation and smart grid issues.

Jayakrishnan Radhakrishna Pillai received the Ph.D. degree in power systems from Aalborg University, Aalborg, Denmark, in 2011. He is currently an Associate Professor with the AAU Energy, Aalborg University. His current research interests include distribution system analysis, grid integration of electric vehicles and distributed energy resources, smart grids, and intelligent energy systems.

Xiaofang Xia received the Ph.D. degree in control theory and control engineering from the Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China, in 2019. She is currently an Associate Professor with the School of Computer Science and Technology, Xidian University, Xi’an, China. Her research interests include cyber-physical systems, smart grid security, database management systems, and anomaly detection.

Guangchun Ruan received the Ph.D. degree in electrical engineering from Tsinghua University, Beijing, China, in 2021. He is currently a Research Associate with the Tsinghua University. He is currently a post-doctoral researcher at the Massachusetts Institute of Technology. His research interests include electricity markets, demand response, machine learning, data science, and energy policy.

Zhe Yang received the B.S. degree, M.S. degree, from Northeast Electric Power University in 2017. He received the M.S. degree from North China Electric Power University in 2020. He received the Ph.D. degree from Aalborg University in 2020. He is currently a post-doctoral researcher at the Hong Kong Polytechnic University in Hong Kong. His research interests include machine learning and power protection.