Electricity Theft Detection Method Based on Ensemble Learning and Prototype Learning

Xinwu Sun, Jiaxiang Hu, Zhenyuan Zhang, *Senior Member*, *IEEE*, Di Cao, *Member*, *IEEE*, Qi Huang, *Fellow*, *IEEE*, Zhe Chen, *Fellow*, *IEEE*, and Weihao Hu, *Senior Member*, *IEEE*

***Abstract*—With the development of advanced metering infrastructure (AMI), large amounts of electricity consumption data can be collected for electricity theft detection. However, the imbalance of electricity consumption data is violent, which makes the training of detection model challenging. In this case, this paper proposes an electricity theft detection method based on ensemble learning and prototype learning, which has great performance on imbalanced dataset and abnormal data with different abnormal level. In this paper, convolutional neural network (CNN) and long short-term memory (LSTM) are employed to obtain abstract feature from electricity consumption data. After calculating the means of the abstract feature, the prototype per class is obtained, which is used to predict the labels of unknown samples. In the meanwhile, through training the network by different balanced subsets of training set, the prototype is representative. Compared with some mainstream methods including CNN, random forest (RF) and so on, the proposed method has been proved to effectively deal with the electricity theft detection when abnormal data only account for 2.5% and 1.25% of normal data. The results show that the proposed method outperforms other state-of-the-art methods.**

***Index Terms*—Electricity theft detection, ensemble learning, prototype learning, imbalanced dataset, deep learning, abnormal level.**

I. Introduction

ELECTRICITY has become essential in our daily life. However, electricity loss occurs in every process with electricity such as electricity generation, transmission, and distribution [1]. In general, these losses can be divided into two classes: nontechnical losses (NTLs) [2]-[4] and technical losses (TLs). Abnormal NTLs are usually caused by electricity theft, including tampering the circuit of the electricity meter and bypassing the electricity meter. Enormous NTLs will bring the power enterprises huge economic damage. It is reported that NTLs have accounted for 25% loss in India and this rate is 16%, 6%, 6%, and 5% in Brazil, China, American, and Australia, respectively [5].

To restraint these economic losses, power enterprises often assign their workers to check the meter of suspicious customers or update the protective device of meter. However, inevitably, these traditional methods have obvious disadvantages. For example, artificial detection relies too much on expert experience, which makes this method difficult to be applied in small enterprise. Besides, improving protective device means the iteration of smart meter, which costs much. Meanwhile, with the development of computer science, the methods of electricity theft are updating quickly such as cyber-attack for two-way communication network in smart grid without any tampering circuit [6]. However, any electricity theft will make some variables abnormal because smart grid is a physical system which satisfies many equations of state. In this case, we can take full advantages of advanced metering infrastructure (AMI), collect key state about smart grid, and conduct some data analysis for electricity theft. Through the sensors of AMI, different kinds of data can be attained such as NTLs, customer's consumption data, and the fluctuation of voltage and electricity.

There are three mainstream directions among current data-driven algorithms of electricity theft detection, including anomaly detection, state estimation, and supervised learning. Anomaly detection aims at seeking the similarity of normal samples or designating an index to judge the class of samples such as clustering, correlation analysis, principal component analysis (PCA), and local outlier algorithm. Compared with supervised learning, anomaly detection is capable of learning consumption pattern and information from unlabeled samples. Reference [7] proposed an algorithm combing density-based clustering and the maximum information coefficient to find out the correlation between NTLs and certain electricity theft. Reference [8] defined user's short-lived consumption pattern and detected ongoing electricity consumption theft. Reference [9] incorporated wavelet-based feature extraction and fuzzy *c*-means clustering. State estimation aims at looking for the anomaly measures among all measures of whole region. Compared with the smart meter in client, the observe meter is more difficult to be tampered and its reading is more credible. Therefore, some inconsis-

Manuscript received: October 18, 2022; revised: December 29, 2022; accepted: February 27, 2023. Date of CrossCheck: February 27, 2023. Date of online publication: April 10, 2023.

This work was supported by National Natural Science Foundation of China (No. 52277083).

X. Sun, J. Hu, Z. Zhang, D. Cao, and W. Hu (corresponding author) are with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China (e-mail: 202221040328@std.uestc.edu.cn; jx.hu@foxmail.com; zhangzhenyuan@uestc.edu.cn; caodi@std.uestc.edu.cn; whu@uestc.edu.cn).

Q. Huang is with Southwest University of Science and Technology, Chengdu, China (e-mail: hwong@uestc.edu.cn).

Z. Chen is with Aalborg University, Aalborg, Denmark (e-mail: zch@et.aau.dk).

DOI: 10.35833/MPCE.2022.000680

tent data between observe meter and user's meter can be detected when electricity thefts happen in this region. However, this algorithm requires high-frequency measurement, entire topological structure, and a large number of sensors installed at important places, which are difficult to be implemented in many villages and towns. Supervised learning aims at training a classifier to designate a boundary which divides abnormal data and normal data into two areas and calculate the probability of each class. In early studies, traditional machine learnings such as support vector machine (SVM) [10]-[13], random forest (RF) [13], gradient boosting model (GBM), and extreme gradient boosting (XGBoost) [14] are conducted in this area. Reference [12] primarily proposed two-step detection which analyses the anomaly NTLs of certain area and locates abnormal customers by SVM. Reference [15] preprocessed real data by decision tree (DT) to extract features and classified samples by SVM. With the increase of dataset size, people begin to pay attention to deep learning-based method which is more effective than machine learnings. Convolutional neural network (CNN) [16]-[19] and recurrent neural network (RNN) [20]-[24] are classical deep learning networks which all have been applied on time series. Reference [25] originally decomposed consumption data into multiple components, which are individually analyzed by deep neural networks. Reference [1] proposed a novel frame which extracts feature from 1-dimensional (1-D) and 2-dimensional (2-D) data. However, the imbalance of training set, which refers to the anomaly samples far less than normal samples, will restrict the performance of the model. For ameliorating this restriction, data augmentation, sampling, and neural networks are utilized in this area [19], [26]-[30]. Reference [19] utilized synthetic minority oversampling technique (SMOTE) to enlarge the number of abnormal samples. This method generates abnormal samples from raw abnormal samples and their neighbor points to increase the diversity of dataset. Reference [26] studied different sampling techniques such as random undersampling (RUS), random oversampling (ROS), and SMOTE, and obtained great improvement of the evaluation index. Reference [29] found out that the distribution of data could be improved by increasing the data located in boundary built by SVM. Therefore, they combined borderline-SMOTE-SVM and Tomek link to balance dataset and make boundary between different classes clear. However, there is no suitable index to prove whether generated samples are usual abnormal samples. Therefore, in this paper, it is the main direction to effectively excavate the feature of abnormal data instead of generating abnormal samples.

In this paper, a novel electricity theft detection model is proposed, which deals with imbalanced dataset well. Firstly, one-class support vector machine (OCSVM) is conducted on every user's consumption data to ascertain their constant electricity usage. Then, CNN, long short-term memory (LSTM), and prototype learning are employed to construct the prototype of each class. Through calculating the Euclidean distance between the sample and each prototype, the label of the sample is determined by the nearest prototype. In this process, the neural network minimizes the distance of the same class and maximizes the distance of different class to make critical features learned by the model. For the unbalanced dataset, the network is trained by different subsets of the training set. Compared with some supervised learning algorithms, the proposed method has better performance on imbalanced dataset.

The main contributions of this paper are summarized as follows.

1) Prototype learning and ensemble learning are firstly implemented in the area of electricity theft detection. In realistic world, the imbalance between abnormal users and normal users causes large imbalance in electricity consumption dataset. In this case, traditional theft detection based on artificial intelligence (AI) cannot play a role due to the risk of overfitting and the lack of feature extraction. However, the proposed method can still distinguish normal and abnormal data when other methods are unable to achieve according to the experiments.

2) Apart from the imbalance of abnormal data size, the influence of abnormal data with different abnormal levels is also considered. Slight electricity theft causes few reductions on consumption data, which reduces the charge of power and the risk of being detected. There is high similarity between abnormal and normal consumption data. Compared with traditional deep learning, the proposed method has greater performance in dealing with these samples which are difficult to be detected. The design of prototype learning significantly improves the performance of the network for this kind of samples.

3) OCSVM is utilized to further prove the constant consumption pattern of signal customers. In this case, the feature from consumption data and electricity theft dataset are reliably enough for model training. This process can be considered as the reliable proof for model learning process.

The rest of the paper is organized as follows. In Section II, the characteristics of electricity consumption and electricity theft are analyzed. In Section III, a novel electricity theft detection method is proposed. Some experiments which verify the performance of the proposed method in imbalanced dataset will be narrated in Section IV. Finally, Section V concludes this paper.

II. Problem Analysis

Electricity theft is a behavior to avoid or reduce electricity cost. All electricity theft can be summarized into three classes, including tampering, bypassing electric energy meters, and false data injection. These behaviors will leave some clues on the consumption data such as abnormal maximum value and abnormal mean value. If customers' behaviors are normal, his/her electricity usage would remain constant due to his/her fixed lifestyle. Therefore, finding out the feature of abnormal usage and normal usage is the key of detecting electricity theft. In this section, the characteristics of customers' consumption data is analyzed by OCSVM, which is utilized to prove the constant usage of most customers.

In this experiment, the public dataset containing 536 days' electricity consumption data of 4225 residential customers, released by Electric Ireland and Sustainable Energy

Authority of Ireland in January 2012 is going to be utilized [31]. Because all participators are voluntary to hand in their consumption data, all data can be assumed to be honest data. Figure 1 represents the trends of daily and weekly consumption data. Figure 1(a) shows the daily consumption data of one customer in four different days. It is easy to find that some values are different, although they have high degree of similarity. We can stretch our detection windows from one day to one week. Figure 1(b) shows four weekly consumption data of one customer in four different weeks. Compared with Fig. 1(a), there is less fluctuation and difference between these curves. We preliminarily draw a conclusion that there is periodicity in weekly electricity usage and some randomness in daily electricity usage.

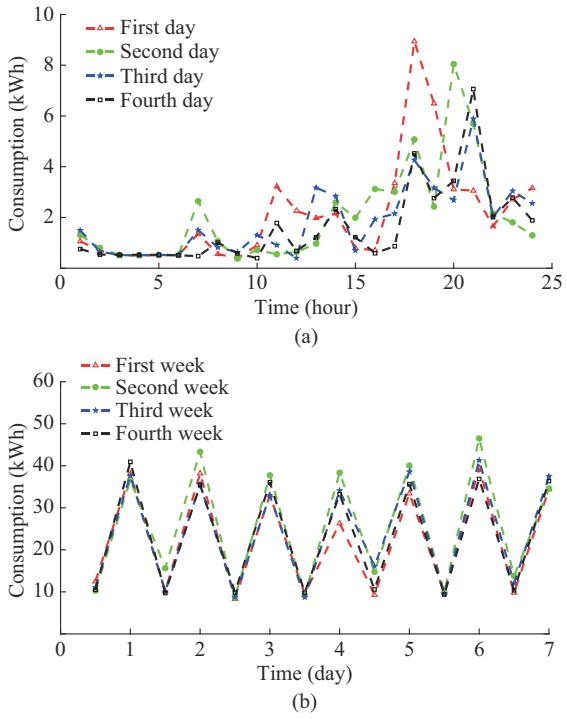

Fig. 1. Trends of daily and weekly consumption data. (a) Daily consumption data sampled in one hour. (b) Weekly consumption data sampled in twelve hours.

The above conclusions are the results of our observation for these curves without precise calculation. For verifying the constant consumption pattern of most customers, OCSVM is conducted on electricity consumption data. As a classical machine learning for novelty detection, OCSVM establishes a boundary with normal samples and distinguishes the label of samples through their position in feature space [32], [33]. For each certain customer, his/her consumption data are divided into two groups, i.e., test set and training set. If his/her electricity usage is constant, most samples from the test set should locate in the boundary. Considering the randomness of daily electricity consumption, consumption data sampled in 12 hours are set as the input of OCSVM. In dataset, all customers have 76 weekly electricity consumption data which are randomly divided into 60 samples for training and 16 samples for testing. Besides, the rate of special data in training set should be told to model. Although the trend of weekly consumption data is more stable than daily consumption data, there are still some special samples which are different with most samples. Therefore, a rough rate of special sample is given to model and this rate is 10%. Figure 2 shows the non-outlier rate of OCSVM for all customers in raw dataset. According to Fig. 2, the highest bar is located at 0.75, which accounts for over 50% of total dataset. The second-highest and third-highest bars are located at 0.65 and 0.85, respectively. This result indicates that these consumption data are classified into the same class for over 70% customers of the dataset. Considering the randomness of weekly electricity consumption, most customers have fixed electricity usage.

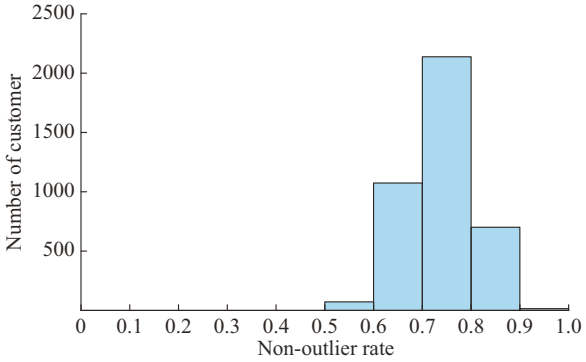

Fig. 2. Non-outlier rate of OCSVM for all customers in raw dataset.

Due to the lack of abnormal data in origin datasets, the abnormal data will be constructed based on the characteristics of real electricity theft. Tampering the circuit will permanently change the measure of smart meter such as lowering all measurements in the same proportion and setting measurements to be zero during some time. Bypassing the circuit means using electricity directly. In this case, the power meter will read the measurements of zeros all the time. Compared with the above two types of electricity theft, false data injection will bring various change on the measurements. Because of different electricity price at different time, peak-load shifting and replacing all measurements with mean value can help theft reduce the cost of used power. There is no reduction on total electricity consumption but large reduction on cost. Meanwhile, some thieves choose to add noises to these data for various fluctuations. On account of these analyses above and referring to the past abnormal function [12], five functions are utilized to construct abnormal data.

Table I lists five specific abnormal functions, where $\boldsymbol{x}$ is a vector including 48 measurements of daily electricity consumption data; and x_t is the t^{th} measurement of $\boldsymbol{x}$. $h_1(\cdot)$ multiples $\boldsymbol{x}$ with the same random value which is less than 0.8. α can be regarded as the severe degree of electricity theft. With the decrease of α, electricity theft has severer damage. $h_2(\cdot)$ sets a range where measure is set to be zero and other measurements remain constant. The function of β is similar to α and larger β means severe electricity theft. $h_3(\cdot)$ replaces all measurements with the average value of corresponding daily consumption. $h_4(\cdot)$ multiplies each measure from $h_3(\cdot)$ by a random value which is less than 0.8. $h_5(\cdot)$ reverses the or-

der of daily meter reading. By disposing original data in these functions, we can obtain five datasets including different types of electricity theft consumption data. The example of the daily consumption data in normal usage and abnormal usage is shown in Fig. 3.

TABLE I
FUNCTIONS OF DIFFERENT ABNORMAL DATA

Class	Abnormal function
Class 1	$h_1(\boldsymbol{x})=a\boldsymbol{x}$, $a=random(\alpha-0.1,\alpha+0.1)$, $\alpha=\{0.1, 0.3, 0.5, 0.7\}$, where *random* is the uniform sample operation
Class 2	$h_2(x_t)=\beta_t x_t$ If $start+duration<48$: $\beta_t=\begin{cases}0 & start\le t\le start+duration\\ 1 & \text{else}\end{cases}$ else $\beta_t=\begin{cases}0 & 0\le t\le start+duration-48\\ 0 & start\le t\le 48\\ 1 & \text{else}\end{cases}$ $start=random(0, 48)$ $\beta=duration=\{8, 16, 24\}$ where *start* means the start time of electricity theft and *duration* means the lasting time of electricity theft
Class 3	$h_3(\boldsymbol{x})=mean(\boldsymbol{x})$, where *mean* is the average of value in $\boldsymbol{x}$
Class 4	$h_4(\boldsymbol{x})=\gamma_t\cdot mean(\boldsymbol{x})$, $\gamma_t=random(0.1, 0.8)$
Class 5	$h_5(x_t)=x_{48-t}$

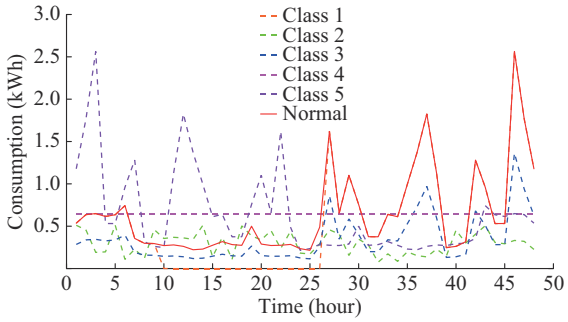

Fig. 3. Daily consumption data in normal usage and abnormal usage.

III. PROPOSED METHOD

In realistic world, the obvious characteristic of electricity consumption dataset is a small ratio of electricity theft data to normal electricity data. However, traditional supervised learnings have difficulty in dealing with this characteristic. In this case, many methods are proposed and have finite effect. Compared with those data augment, ensemble learning makes full use of existing dataset by training some weak classifiers and synthesizing their predictions. However, constructing weak classifications based on the neural network will cost amounts of time and memory. Therefore, this paper focuses on improving the accuracy of theft detection while the abnormal samples are few.

A. Batch Ensemble Learning

Weak classifier refers to the classifier whose accuracy is more excellent than random prediction. The training sets of different weak classifiers are different subsets of total training set. In this case, some trained classifiers will learn different features of training set and give contrary prediction for the same sample. Meanwhile, most classifiers will give the right prediction, which corrects the mistakes of few classifiers. Considering these weak classifiers and their predictions synthetically, a strong classifier is produced.

However, it takes long time for neural network to train its parameters. Meanwhile, the combination of multiple deep neural networks has high requirement for memory. Therefore, a deep neural network is set to replace all weak classifiers in this paper. To ensure the smooth training process of the model, the balanced subset of the total training set is extracted, which contains all abnormal samples and the same number of normal samples. In this case, the balance of training set forces the network not to prefer a certain class. At the same time, different training sets of different epochs avoid the parameters of neural network falling into local optima. After many epochs of training, all samples can be utilized fully and trained by network. Because this training method is similar to the design of batch training, it is called batch ensemble learning.

B. Data Preprocess

Before the training process, the raw data need to be preprocessed because different value ranges may influence the convergence speed and generalization performance of the model. There are two common standardization methods.

1) Zero-score Standardization

The function of zero-score standardization is to let raw data follow Gaussian distribution. The following equation is the expression of zero-score standardization:

$$f(\boldsymbol{x})=\frac{x_i-mean(\boldsymbol{x})}{std(\boldsymbol{x})}\quad x_i\in\boldsymbol{x} \tag{1}$$

where $std(\boldsymbol{x})$ represents the standard deviations of $\boldsymbol{x}$. This standardization will worsen the performance of the model if the raw data do not satisfy Gaussian distribution.

2) Min-max Scaling

The function of min-max scaling is to let raw data equal to $[0, 1]$ in equal proportion. The following equation is the expression of min-max scaling:

$$f(\boldsymbol{x})=\frac{x_i-\min(\boldsymbol{x})}{\max(\boldsymbol{x})-\min(\boldsymbol{x})}\quad x_i\in\boldsymbol{x} \tag{2}$$

where $\max(\boldsymbol{x})$ and $\min(\boldsymbol{x})$ represent the maximum and minimum values of $\boldsymbol{x}$, respectively. Compared with zero-score standardization, this method is more widespread and does not have preconditions for raw data. After the test of these two methods, we choose the second to normalize our dataset.

C. Prototype Learning

Prototype learning [34] is firstly applied in few-shot learning. Its design of the construction of prototype and prediction based on Euclidean distance has great performance. Reference [35] proved that the utilization of prototype will bring model greater robustness than the combination of *softmax* layer and cross-entropy loss function. Therefore, prototype learning is applied to extract the feature of samples in our proposed method.

Figure 4 shows the basic construction of prototype network, where L1-L3 mean the prototypes of different classes. The prototype learning consists of three main parts: set parti-

tion, feature embedding, and calculation of prototype. In set partition, the input is divided into two sets including support set (S), which is used to construct prototype, and query set (Q), which is used to optimize parameters of the network. After that, all samples in S and Q are put into feature embedding. Through feature embedding, all samples obtain their representations in feature space. Meanwhile, the prototype of corresponding class can be constructed by the representations of samples in S. The following equation is the specific calculation method:

$$\boldsymbol{c}_k = \frac{1}{N} \sum_{j \in [1, N]} f(\boldsymbol{x}_{k,j}) \tag{3}$$

where k is the class of consumption data; $f(\boldsymbol{x}_{k,j})$ is the embedded feature of the j^{th} support vector $\boldsymbol{x}$ belonging to class k; and $\boldsymbol{c}_k$ is the prototype of class k.

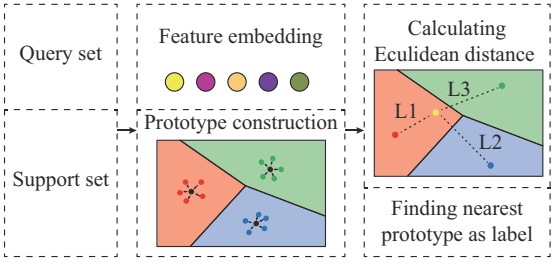

Fig. 4. Basic construction of prototype network.

Then, the representations of samples in Q are utilized to predict their class by calculating the Euclidean distance of them with all prototypes and finding the nearest prototype. According to these distances, the probability of all class can be calculated by *softmax* layer. With the help of cross-entropy function and back propagation, the parameter can be optimized in right direction. The following equation is the concrete loss function:

$$Loss = -\sum_{i \in [1, batch]} [\boldsymbol{y}_i \ln \hat{\boldsymbol{y}}_i + (1 - \boldsymbol{y}_i) \ln (1 - \hat{\boldsymbol{y}}_i)] \tag{4}$$

where $\boldsymbol{y}_i$ is the one-hot coding; and $\hat{\boldsymbol{y}}_i$ is the corresponding probability vector. According to the optimization of loss function, the distance of representations from the same class decreases while that from different classes increases.

In traditional CNN with *softmax* layer and cross-entropy function, the samples are often mapped to certain area in feature space. In this case, the distance between the features from the same class may be further than that from different classes. In this paper, the samples from the same class are mapped into certain point in feature space. This design makes the prototype representative and improves the robustness of the network.

D. CNN and LSTM

It is easy to know that the quality of prototype depends on the distribution of dataset and the ability of network. According to current deep learning framework, CNN and LSTM are utilized to extract the features of samples. In this subsection, CNN focuseses on extracting the characteristic about the periodicity of raw samples. LSTM focuses on extracting the characteristic about the global feature of raw samples. The detailed structures of two subnetworks are narrated as follows.

1) LSTM

According to the analysis of consumption data in different days, some characteristics of consumption pattern such as the maximum value, the minimum value, their corresponding time indices, and the fluctuations can be revealed. As the variant of RNN, LSTM [22]-[24], [36] is usually utilized to extract features from time series. LSTM shortens the training time and solves the problem of gradient disappearance when the length of input is too long, which is suitable for extracting the global feature of electricity data.

The construction of LSTM cell is shown in Fig. 5. And LSTM is combined with a certain number of identical LSTM cells depending on the length of data. There are three important operations in LSTM cell including forgetting information, recording information, and updating information.

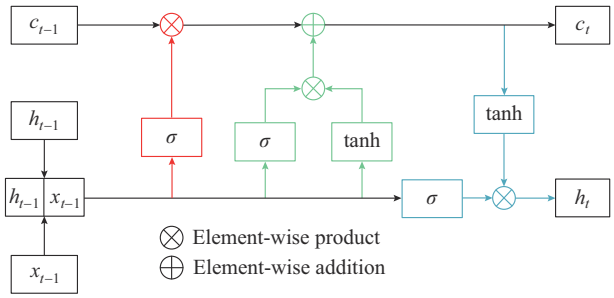

Fig. 5. Construction of LSTM cell.

The red route can be regarded as forgetting information. And forgetting signal is constructed by following formulation.

$$f_t = \sigma(\boldsymbol{W}_{if} x_{t-1} + \boldsymbol{W}_{hf} h_{t-1} + \boldsymbol{b}_f) \tag{5}$$

where f_t is the forgetting signal; and $\sigma(\cdot)$ is the *sigmoid* function which lets the number in f_t map between 0 and 1; $\boldsymbol{W}_{if}$ and $\boldsymbol{W}_{hf}$ are all trainable coefficient matrices; and $\boldsymbol{b}_f$ is a trainable bias matrix. Therefore, the element-wise product of c_{t-1} and f_t will drop some information in c_{t-1}.

The green route can be regarded as recording information. The recording signal is combined by the following formulations.

$$m_t = \sigma(\boldsymbol{W}_{im} x_{t-1} + \boldsymbol{W}_{hm} h_{t-1} + \boldsymbol{b}_m) \tag{6}$$

$$\tilde{c}_t = \tanh(\boldsymbol{W}_{ic} x_{t-1} + \boldsymbol{W}_{hc} h_{t-1} + \boldsymbol{b}_c) \tag{7}$$

where m_t is the recording signal which is similar to f_t; $\tilde{c}_t$ is the abstract feature of the current input; $\boldsymbol{W}_{im}$, $\boldsymbol{W}_{hm}$, $\boldsymbol{W}_{ic}$, and $\boldsymbol{W}_{hc}$ are all trainable coefficient matrices; and $\boldsymbol{b}_m$ and $\boldsymbol{b}_c$ are trainable bias matrices. Through the element-wise product of m_t and $\tilde{c}_t$, the information of background decays less and irrelevant information are removed.

After that, c_t remains the feature about the relationship between the past and current inputs. However, the finite metrics only retain finite information and the information from a long time ago will be covered. Therefore, through the function of blue route which filters the information of c_t, the information has been kept in h_t a long time before. In the experiment, the last h_t is utilized to represent the global feature

of the sample.

2) CNN

Through the novelty detection for weekly consumption data by OCSVM and observation of the consumption data in different weeks, the periodicity of electricity consumption for most customers can be proved. For example, the consumption data of weekends are usually higher than the consumption data of weekdays. In LSTM, the consumption data are handled in order, which will let the relation of value at interval ignored. Therefore, for extracting the periodicity of electricity usage, CNN is utilized. In this subsection, the daily electricity consumption data are folded into 2-D shape. Through sliding convolution window, we can extract features about the relation of consumption data in convolution window. The concrete CNN consists of five similar blocks, which are listed in Table II.

TABLE II
PARAMETERS OF CONCRETE CNN

Layer	Parameter	Number
Conv2d	(C, 3, 3) or (C, 5, 3)	1, 2, 3, 4, 5
ReLU	0	1, 2, 3, 4, 5
AvgPool2d	0	1
BatchNorm2d	0	1, 3

Table II lists all parameters of blocks in CNN where the C in parameter is the number of the channels. In this table, numbers mean the blocks where this layer exists. Two-dimensional convolutional layer (Conv2d) exists in all blocks for extracting feature. In general, convolution kernel of (3, 3) and (5,5) is conducive for the performance of network. Combined with the reality, the size of all convolutional kernel is (5,3) in the last Conv2d. Rectified linear unit (ReLU) following Conv2d increases the nonlinearity of network and prevents CNN from degenerating into MLP. Besides these parts, two-dimensional average pool layer (AvgPool2d) is utilized to adjust the shape of input and remains most information of input, which is beneficial to reduce the depth of network. Meanwhile, two-dimensional batch normalization layer (BatchNorm2d) is utilized to speed up the convergence rate.

3) Fully Connected Layer

After the disposal of LSTM network and CNN network, we concatenate two 1-D vectors and generate prototypes by calculating the mean of features of each class. However, the length of prototype will be too long, which will increase the cost of time. In this case, the fully connected layer will be used to adjust the length of the prototype and the proportion of two features.

E. Framework of Proposed Algorithm

The framework of the proposed algorithm is shown in Fig. 6. Firstly, sampled from raw data, training dataset consists of all abnormal samples and the same number of normal samples. After normalization, training data are divided into support set and query set. The samples in support set are utilized to construct the prototype of each class. The samples in query set are utilized to test the performance of model. After predicting the labels of samples in query set, all predicted labels are used to guide the parameters of model update. In the test process, all test samples belong to query set and training samples belong to support set. Through finding out the nearest prototype in feature space, the labels of test samples are determined.

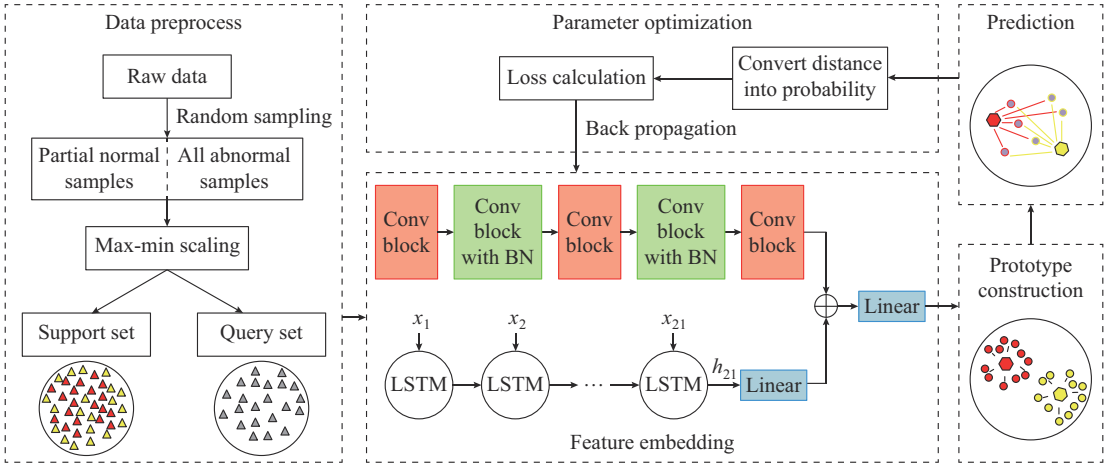

Fig. 6. Framework of proposed algorithm.

IV. RESULTS

In this section, training process and parameter optimization will be narrated in detail. To demonstrate the performance of the proposed method, some experiments are set including parameter optimization, comparing experiment, sensitivity analysis of abnormal level, and ablation experiment. Besides these, three metrics including true positive rate (TPR), false positive rate (FPR), and area under curve (AUC) are chosen to evaluate the performance of the proposed method.

A. Electricity Consumption Data

According to the abnormal functions in Table I, we have a benign dataset and five abnormal datasets whose shapes are all 4225×536×48, where 4225, 536, and 48 represent the

number of customers, days and sampling number of one day, respectively. The training set, test set, and validation set will be sampled from benign dataset and five abnormal datasets. Firstly, 2760 customers' indices are randomly chosen, including 1800 customers in training set, 480 customers in test set, and 480 customers in validation set. In the following step, the similar methods are conducted on three datasets. Taking the training set as an example, 1800 customers are randomly divided into six parts, where normal class accounts for half and each abnormal class accounts for 10% of all. Customers' electricity consumption belonging to the corresponding class is collected to assemble the training set. According to this method, normal data and the corresponding abnormal data cannot be obtained from the network at the same time, which is more practical. Meanwhile, as the generalization of model needs to be proved, different customers' future electricity consumptions are tested and validated in this paper. In the following experiments, the ability of the proposed model for imbalanced datasets is seriously concerned. Therefore, only few parts of abnormal data in the training set will be utilized. Meanwhile, untrained balanced dataset will be used to test the proposed model.

B. Performance Metrics

In the experiment, three performance metrics, i. e., *TPR*, *FPR*, and *AUC*, are considered [37]. These metrics all depend on confusion metrics. These metrics will be introduced in detail as follows.

Table III presents confusion metric, which stores the total prediction.

TABLE III
CONFUSION METRIC

Label	Prediction	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

According to this confusion, the following three metrics can be calculated, which is helpful for the calculation of *AUC*.

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

$$FPR = \frac{FP}{TN + FP} \tag{9}$$

$$Diff = TPR - FPR \tag{10}$$

where *TPR* indicates the ratio of true positive sample to all positive samples; and *FPR* indicates the ratio of false positive samples to all predicted positive samples. In electricity theft detection, our purpose is to find out all abnormal data and avoid predicting normal sample as abnormal. If *TPR* is high and FPR is low, the classifier has good performance on the dataset. However, it is difficult to let these two indices come to ideal indirection at the same time. When an algorithm gives many positive predictions, the ratio of wrong prediction will inevitably rise. In this case, *Diff* is also considered to evaluate the performance of our method.

However, even if two different methods obtain the same *TPR* and *FPR*, there are still differences between these two methods. For example, when model A gives a positive sample with the positive probability of 0.9 and model B gives the same sample with the positive probability of 0.6, all models will give the sample with positive prediction. If a random sample which is never trained needs to be predicted, model B has less confidence to give a definite prediction, which also can be regarded as the alility of the model. Therefore, *AUC* is conducted to check the confidence of the proposed method. Compared with *TPR* and *FPR*, this index accounts for the score of a randomly chosen sample. In general, an excellent method will give different scores for different classes, like the score closed to 0 for negative samples and the score closed to 1 for positive samples. Therefore, *AUC* can help us realize whether the method distinguishes the class of sample well. *AUC* is calculated by the mean of *TPR* for different thresholds from 0 to 1. Before *AUC* is calculated, a series of boundary i need to be set. When probability of the sample is less than i, the model will give a positive prediction to the sample. The following formulation is the expression of *AUC*:

$$AUC = \frac{1}{2}\sum_{k\in[1,N],k\in Z}\left[TPR\left(i=\frac{k}{N}\right)+TPR\left(i=\frac{k-1}{N}\right)\right]\cdot \left[FPR\left(i=\frac{k}{N}\right)-FPR\left(i=\frac{k-1}{N}\right)\right] \tag{11}$$

where $TPR(i=k/N)$ and $FPR(i=k/N)$ denote the values of *TPR* and *FPR* when the boundary is k/N, respectively; and N is the number of boundaries.

C. Parameter Optimization

Before the performance of the proposed method is compared with other methods, four comparing experiments are set to choose the best parameters. Due to the way of prediction which is based on the distance of feature space, we think that the number of prototype's dimensions is more important than other parameters such as batch size and learning rate. Therefore, four lengths of prototype's dimension are tested, including 16, 32, 64, and 128. Because the proportion of each class is equal, accuracy is simply chosen as the metric to compare the performance of the network. Figure 7 shows the performance of the network with different lengths of prototype.

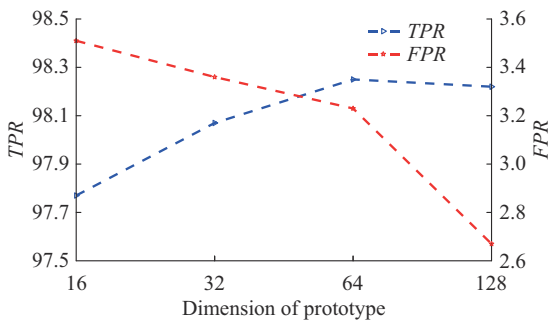

Fig. 7. Performance of network with different lengths of prototype.

According to Fig. 7, the *TPR* of test set fluctuates with

the increasing dimension and reaches 98.22% when the prototype's dimension is 128. Similar to *TPR*, *FPR* reaches the lowest value 2.67% when the number of prototype's dimension is 128. If the prototype's dimension continues to increase, it will be similar to the length of raw data, which will waste time and lower the ability of feature embedding. Therefore, 128 is determined as the length of prototype's dimension.

D. Comparing Experiment

To verify the superiority of the proposed algorithm, other five classification methods which have been used for electricity theft detection are conducted on the given training set. These five methods and corresponding concrete parameters will be introduced in the following section.

1) SVM [12], [15], [38]: it is a typical supervised machine learning which has been widely used in early research of electricity theft detection. Because the key of this method is to find out support vector which is constructed by a few samples closed to support vector, this method is not influenced by the amount of data.

2) RF [15]: this machine learning method is based on bagging, a type of ensemble learning. In this method, all weak classifiers are built in parallel. Through voting mechanism, the predictions of all weak classifiers are synthesized to determine the final prediction. Compared with single DT, RF trains many DTs with different subsets of training set, which ensures the discrepancy between DT models. Meanwhile, RF can set different weights to different classes to deal with the imbalance of dataset.

3) Adaboost [39]: Adaboost is a machine learning based on boosting, which is a type of ensemble learning. The difference of Adaboost and RF is the generation method of the weak classifiers. In Adaboost, weak classifiers are built in sequence. The weight of every classified sample is continuously revised and put into the next weak classifiers for training. In the end, different weights are assigned to all weak classifiers depending on their accuracy.

4) CNN [1], [18], [19], [25]: a five-layer CNN block is designed as the compared model which is similar to the CNN component of the proposed method. Meanwhile, different weights according to the ratio of abnormal data to normal data are given to samples.

5) Deep belief network (DBN): DBN is a probability generation model which consists of multiple restricted Boltzmann machine (RBM) and fully connected layers. Due to the unsuitably initial parameters which will make model get stuck at locally optimal value, pre-training is conducted on the RBM to obtain great mapping function and lose little information in the process of mapping. This pre-training can be regarded as the fine adjustment of initial parameters. After the process of pre-training, the DBN is trained by background propagation.

Table IV shows the concrete hyper-parameters of the compared methods. The hyper-parameters of SVM, RF, Adaboost, and DBN refer to the existing research. The hyper-parameter of CNN is the same as the CNN section of the proposed method. In general, the existing research deals with the imbalance of dataset by two methods, i.e., enlarging abnormal datasets and giving different weights to different classes. In our experiments, the second method is utilized on CNN and RF. Abnormal data are given larger weights than normal data according to the ratio of abnormal data to normal data.

TABLE IV
CONCRETE HYPER-PARAMETERS OF COMPARED METHODS

Compared method	Hyper-parameter
SVM [12], [15], [38]	Kernel is "RBF", Gamma is "auto", $C = 1.0$, and weight is the rate of abnormal data and normal data
RF [15]	The number of estimators is 40, criterion is "entropy", and random_state = 0
RF (weight) [15]	Weight is the same with SVM, and other hyper-parameters are the same with RF
Adaboost [39]	n_estimator = 100, learning_rate = 0.06, classifier is "DT"
CNN (weight) [1], [18], [19], [25]	Weight is the same with SVM, and Epoch is 100
DBN	The number of RBM is 3, and the number of neurons in each RBM is [336, 138, 32]

As stated above, there are 900 normal customers in our training set. Meanwhile, different numbers of abnormal customers in training set are utilized to form the imbalanced set, which are 10%, 5%, 2.5%, and 1.25% the size of normal data, respectively. For test section, the same balanced datasets are utilized to test all of the methods. The classifying result of all of the methods for different imbalanced datasets is shown in Table V.

Table V shows the *TPR*, *FPR*, *Diff*, and *AUC* of different methods when the imbalance of dataset is different. In this table, the previous four methods belong to statistics-based method while the last three methods are based on neural networks. Comparing *Diff* of all of the methods, it can be found out that only the proposed method and CNN (weight) succeed in distinguishing the labels of most of samples correctly when the ratio is 10%. The low *TPR* and *FPR* which are close to 0 indicate that many abnormal samples are mistakenly judged as normal samples for the previous five methods. However, when the probability that each sample belongs to a certain class is calculated, there is a clear boundary between the abnormal samples and normal samples because of high *AUC*. It may be due to that SVM, RF, and Adaboost are non-parameter methods, which seriously depend on the distribution of training samples. If the difference of abnormal samples and normal samples in training set is not obvious for machine learning, these methods fail to completely distinguish samples in test set but to judge them as normal samples with lower probability than real normal samples. As for DBN, although it has RBM and *sigmoid* activation layer to obtain the features, its small capacity makes extracting available feature and distinguishing samples difficult. Therefore, it can be observed that the performance of the previous five methods becomes worse when the ratio reduces. As a result, deep learning such as CNN (weight) and

the proposed method can deal with imbalanced dataset. The proposed method achieves better performance than CNN (weight). Due to the batch ensemble learning, only balanced subsets of training set are feed into the network at each epoch of training. In this case, balanced abstract feature is utilized by the proposed method to optimize its parameters. On the contrary, large amounts of features from normal features and little features from abnormal features are obtained by CNN (weight), which result in the overfitting of the preference for normal data. This also can be reflected from Table V, where *Diff* between CNN (weight) and the proposed method becomes larger when the ratio is reduced.

TABLE V
CLASSIFYING RESULT OF ALL METHODS FOR DIFFERENT IMBALANCED DATASETS

Method	Ratio of abnormal data to normal data is 10%				Ratio of abnormal data to normal data is 5%				Ratio of abnormal data to normal data is 2.50%				Ratio of abnormal data to normal data is 1.25%			
	TPR (%)	*FPR* (%)	*Diff* (%)	*AUC*	*TPR* (%)	*FPR* (%)	*Diff* (%)	*AUC*	*TPR* (%)	*FPR* (%)	*Diff* (%)	*AUC*	*TPR* (%)	*FPR* (%)	*Diff* (%)	*AUC*
SVM	36.82	0.58	36.24	0.8491	28.17	0.27	27.90	0.8598	13.33	0.08	13.25	0.8491	9.97	0.15	9.82	0.8367
RF	66.41	0.35	66.06	0.9788	51.78	0.07	51.71	0.9671	36.94	0.12	36.82	0.9336	16.00	0.10	15.90	0.8879
RF (weight)	61.09	0.26	60.83	0.9792	40.78	0.06	40.72	0.9681	26.01	0.05	25.96	0.9338	9.16	0.09	9.07	0.9056
Adaboost	7.67	0.13	7.54	0.9202	3.76	0	3.76	0.8999	0.31	0.01	0.30	0.9254	0	0	0	0.8729
DBN	51.85	1.13	50.72	0.8453	46.60	0.76	45.84	0.8135	0	0	0	0.5457	0	0	0	0.5342
CNN (weight)	91.91	4.94	86.97	0.9803	84.76	4.10	80.66	0.9665	83.55	5.50	78.05	0.9672	79.89	10.64	69.25	0.9253
Proposed	96.32	2.52	93.80	0.9837	92.80	2.34	90.47	0.9709	94.97	5.20	89.76	0.9654	95.50	10.34	85.16	0.9405

E. Sensitivity Analysis of Abnormal Level

In this experiment, the ability of dealing with the samples which are difficult to be detected is tested. Compared with class 3, class 4, and class 5, the abnormal levels of class 1 and class 2 are mutable for the different values of coefficients α and β. As shown in Fig. 3, while α is close to 0, there is less similarity between normal data and abnormal data of class 1. While β is close to 24, more data are set to be 0, which makes samples lose more important features like max electricity consumption. In general, these data are easier to be recognized as abnormal data by the network. Meanwhile, if α is close to 0.7 and β is close to 8, it is difficult to extract key feature and recognize abnormal pattern. Therefore, different groups of α and β are utilized to construct abnormal data. The concrete values of α and β are listed in Table VI.

Because α and β are set as research objects, the classes of abnormal data in our training set are only class 1 and class 2 which account for 50%, respectively. Meanwhile, the classes of abnormal data in the validation set and test set are the same with the training set. Besides, the ratio of abnormal data to normal data is 10% in the training set. According to twelve groups' experiment, the result is shown in Fig. 8. Figure 8 shows *AUC* and *Diff* for different groups of α and β. These bars represent *AUC* of the proposed method. With α increasing from 0.1 to 0.7, the difference of the proposed method all drops no matter how much β is. It may be due to the slight differences between abnormal samples and normal samples when raw electricity consumption is low. However, no matter what value β is, *Diff* and *AUC* do not change greatly when α is constant. Compared with the influence of β on electricity theft detection, α is a more challenging factor. It is obvious that *Diff* will reduce when α increases. According to the contrast of these two factors, we can conclude that the proposed method has great robustness for different β.

To verify the good performance of the proposed method, CNN (weight) is chosen to conduct partial experiment. ($\alpha = 0.1$, $\beta = 24$), ($\alpha = 0.5$, $\beta = 16$), and ($\alpha = 0.7$, $\beta = 8$) represent three abnormal levels. For abnormal data at these three levels, the performance of CNN (weight) and the proposed method is shown in Table VII.

TABLE VI
VALUES OF α AND β

Parameter	Value
α	{0.1, 0.3, 0.5, 0.7}
β	{8, 16, 24}

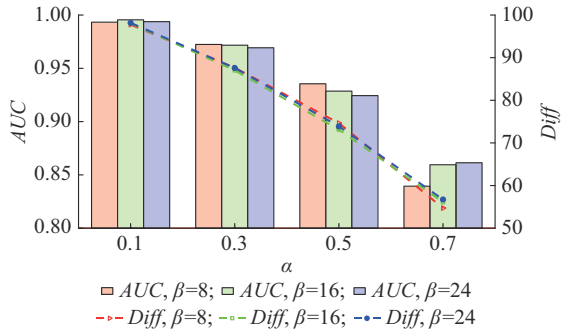

Fig. 8. *AUC* and *Diff* for different groups of α and β.

TABLE VII
PERFORMANCE OF CNN (WEHGHT) AND PROPOSED METHOD FOR ABNORMAL DATA AT THREE LEVELS

Method	($\alpha = 0.1$, $\beta = 24$)		($\alpha = 0.5$, $\beta = 16$)		($\alpha = 0.7$, $\beta = 8$)	
	Diff (%)	*AUC*	*Diff* (%)	*AUC*	*Diff* (%)	*AUC*
CNN (weight)	96.80	0.9988	68.92	0.9220	41.10	0.7806
Proposed	98.17	0.9938	73.12	0.9286	54.73	0.8394

When the abnormal level is reduced from ($\alpha=0.1$, $\beta=24$) to ($\alpha=0.5$, $\beta=16$), *Diff* of CNN (weight) decreases by 27.78%, which is about 1.1 times the reduction of the proposed method. When the abnormal level is reduced from ($\alpha=0.5$, $\beta=16$) to ($\alpha=0.7$, $\beta=8$), *Diff* of the proposed method is 1.51 times that of CNN (weight). Meanwhile, *AUC* of CNN (weight) also drops violently. While *AUC* of the proposed method fluctuates only 0.15, *AUC* of CNN (weight) drops by 0.21. When the similarity between abnormal samples and normal samples increases, the performance of CNN (weight) deteriorates faster than the proposed method. Therefore, it is concluded that the proposed method has greater robustness in dealing with abnormal data at low abnormal level.

F. Ablation Study

In this experiment, the function of prototype learning and batch ensemble learning in improving the performance of electricity theft detection will be tested when the imbalance of dataset is violent. There are three models in this experiment including CNN + LSTM, CNN + LSTM + Ensemble, and the proposed model. To avoid the influence of irrelevant variables, the used datasets including training dataset, validation dataset and test dataset are the same. While the ratio of abnormal data to normal data becomes less, the performance of every model becomes worse. Therefore, to highlight the function of models, only 2.5% abnormal data of normal data are utilized to train the model.

Table VIII shows *Diff* and *AUC* of different models when the ratio of abnormal data to normal data is 2.5%. The CNN + LSTM is set as basic model which obtains the lowest *Diff* and *AUC*. With the addition of batch ensemble learning, there are 24.55% growth on *Diff* and 0.062 growth on *AUC*. Due to the balanced subsets trained in the training process of basic model, the overfitting of model can be ameliorated. However, with the training process going on, overfitting finally happens because all normal samples have been fed into the network. With the addition of prototype learning, *Diff* increases to 89.76% while *AUC* decreases to 0.9654. It is attributed to the method of utilizing feature. Basic model prefers to extract the relevant feature to determine the labels of samples. On the contrary, prototype learning utilizes the thought of cluster to make samples belonging to the same class locate in the same position of feature space. In this process, some weak relevant features will be utilized by basic model to predict more samples correctly, which weakens the generation of the model. Prototype learning pays more attention on the similarity of feature instead of partial information.

TABLE VIII
EXPERIMENT RESULTS OF DIFFERENT MODELS

Model	*Diff* (%)	*AUC*
CNN + LSTM	60.11	0.9154
CNN + LSTM + Ensemble	84.66	0.9763
Proposed method	89.76	0.9654

V. CONCLUSION

In this paper, an electricity theft detection method based on ensemble learning and prototype learning is proposed, which has great performance on imbalanced dataset. According to feature embedding, the abstract feature of every sample is obtained to construct the prototype of each class. After that, the label of each sample is determined by searching the nearest prototype in feature space. In the training process, through extracting the balanced dataset from the total training set, the preference of the model is restrained and the generation of the model improves. To verify the performance of the proposed method on imbalanced dataset, some experiments including parameter optimization, comparing experiment, sensitivity analysis of abnormal level, and ablation study are conducted. Compared with mainstream ensemble learning and deep learning, the proposed method reflects the strongest ability of classification. When the abnormal level of abnormal data decreases, there is less impact on the proposed method while another model loses the ability of classification. Although the proposed method has great performance, there are also disadvantages such as the instability of training process compared with CNN. In our analysis, if we can obtain consumption data which come from customers with the same occupations, the proposed method can get better result using fewer abnormal data. In our opinion, the electricity theft detection should point to imbalanced dataset and how to combine the data from different sources such as occupation and permanent resident population to improve the detection model.

REFERENCES

[1] Z. Zheng, Y. Yang, X. Niu *et al*., "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1606-1615, Apr. 2018.

[2] J. I. Guerrero, I. Monedero, F. Biscarri *et al*., "Non-technical losses reduction by improving the inspections accuracy in a power utility," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 1209-1218, Mar. 2018.

[3] Z. Yang, W. Liao, Q. Zhang *et al.*, "Fault coordination control for converter-interfaced sources compatible with distance protection during asymmetrical faults," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 7, pp. 6941-6952, Jul. 2023.

[4] N. F. Avila, G. Figueroa, and C. Chu, "NTL detection in electric distribution systems using the maximal overlap discrete wavelet-packet transform and random undersampling boosting," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 7171-7180, Nov. 2018.

[5] R. Katakey and R. K. Singh, "India fights to keep the lights on," *Bloomberg Business Week*, vol. 2014, no. 4382, pp. 21-22, May 2014.

[6] S. Tufail, S. Batool, and A. I. Sarwat, "False data injection impact analysis in ai-based smart grid," in *Proceedings of SoutheastCon*, Atlanta, USA, Mar. 2021, pp. 1-7.

[7] K. Zheng, Q. Chen, Y. Wang *et al*., "A novel combined data-driven approach for electricity theft detection," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1809-1819, Mar. 2019.

[8] M. Zanetti, E. Jamhour, M. Pellenz *et al*., "A tunable fraud detection system for advanced metering infrastructure using short-lived patterns," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 830-840, Jan. 2019.

[9] R. Qi, J. Zheng, Z. Luo *et al.*, "A novel unsupervised data-driven method for electricity theft detection in AMI using observer meters," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-10, Jul. 2022.

[10] E. U. Haq, J. Huang, H. Xu *et al*., "A hybrid approach based on deep learning and support vector machine for the detection of electricity theft in power grids," *Energy Reports*, vol. 7, no. 6, pp. 349-356, Nov.

2021.
[11] X. Kong, X. Zhao, L. Chao *et al*., "Electricity theft detection in low-voltage stations based on similarity measure and DT-KSVM," *International Journal of Electrical Power & Energy Systems*, vol. 125, no, 3, p. 106544, Feb. 2021.
[12] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216-226, Jan. 2016.
[13] Z. Qu, H. Li, Y. Wang *et al*., "Detection of electricity theft behavior based on improved synthetic minority oversampling technique and random forest classifier," *Energies*, vol. 13, no. 8, p. 2039, Apr. 2020.
[14] Z. Yan and H. Wen, "Electricity theft detection base on extreme gradient boosting in AMI," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1-9, Jan. 2021.
[15] A. Jindal, A. Dua, K. Kaur *et al*., "Decision tree and SVM-based data analytics for theft detection in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1005-1016, Jun. 2016.
[16] M. Tariq and H. V. Poor, "Electricity theft detection and localization in grid-tied microgrids," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 1920-1929, May 2018.
[17] D. Yao, M. Wen, X. Liang *et al*., "Energy theft detection with energy privacy preservation in the smart grid," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7659-7669, Oct. 2019.
[18] H. Gao, S. Kuenzel, and X. Zhang, "A hybrid CONVLSTM-based anomaly detection approach for combating energy theft," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-10, Aug. 2022.
[19] M. N. Hasan, R. N. Toma, A. A. Nahid *et al*., "Electricity theft detection in smart grid systems: a CNN-LSTM based approach," *Energies*, vol. 12, no. 17, pp. 1-18, Aug. 2019.
[20] A. Takiddin, M. Ismail, M. Nabil *et al*., "Detecting electricity theft cyber-attacks in AMI networks using deep vector embeddings," *IEEE Systems Journal*, vol. 15, no. 3, pp. 4189-4198, Sept. 2021.
[21] M. Ismail, M. F. Shaaban, M. Naidu *et al*., "Deep learning detection of electricity theft cyber-attacks in renewable distributed generation," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3428-3437, Jul. 2020.
[22] S. Li, W. Hu, D. Cao *et al*., "Electric vehicle charging management based on deep reinforcement learning," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 3, pp. 719-730, May 2022.
[23] H. Zhou, Y. Zhou, J. Hu *et al*., "LSTM-based energy management for electric vehicle charging in commercial-building prosumers," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 5, pp. 1205-1216, May 2022.
[24] H. Yang, R. C. Qiu, and H. Tong, "Reconstruction residuals based long-term voltage stability assessment using autoencoders," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1092-1103, Dec. 2020.
[25] Y. Zhang, Y. Ji, and D. Xiao, "Deep attention-based neural network for electricity theft detection," in *Proceedings of 2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, Oct. 2020, pp. 154-157.
[26] J. Pereira and F. Saraiva, "A comparative analysis of unbalanced data handling techniques for machine learning algorithms to electricity theft detection," in *Proceedings of 2020 IEEE Congress on Evolutionary Computation (CEC)*, Glasgow, UK, Jul. 2020, pp. 1-8.
[27] Y. Kulkarni, S. Hussain, K. Ramamritham *et al*., "EnsembleNTLDetect: an intelligent framework for electricity theft detection in smart grid," in *Proceedings of 2021 International Conference on Data Mining Workshops (ICDMW)*, Auckland, New Zealand, Dec. 2021, pp. 527-536.
[28] R. Yao, N. Wang, W. Ke *et al*., "Electricity theft detection in unbalanced sample distribution: a novel approach including a mechanism of sample augmentation," *Applied Intelligence*, doi: 10.1007/s10489-022-04069-z
[29] A. Arif, T. A. Alghamdi, Z. A. Khan *et al*., "Towards efficient energy utilization using big data analytics in smart cities for electricity theft detection," *Big Data Research*, vol. 27, p. 100285, Feb. 2022.
[30] H. Liu, Z. Li, and Y. Li, "Noise reduction power stealing detection model based on self-balanced data set," *Energies*, vol. 13, no. 7, p. 1763, Apr. 2020.
[31] Commission for Energy Regulation (CER). (2012, Dec.). CER smart metering project - electricity customer behavior trial, 2009-2010, 1st edition, Irish social science data archive. SN: 0012-00. [Online]. Available: https://www.ucd.ie/issda/data/commissionforenergyregulationcer/
[32] B. Scholkopf, R. Williamson, A. Smola *et al.*, "Support vector method for novelty detection," in *Proceedings of Conference and Workshop on Neural Information Processing Systems (NIPS)*, Cambridge, USA, Nov. 1999, pp. 583-588.
[33] I. Parvez, M Aghili, A. I. Sarwat *et al*., "Online power quality disturbance detection by support vector machine in smart meter," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 5, pp. 1328-1339, Sept. 2019.
[34] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of Conference and Workshop on Neural Information Processing Systems (NIPS)*, Red Hook, USA, Dec. 2017, pp. 4077-4087.
[35] H. Yang, X. Zhang, F. Yin *et al.*, "Robust classification with convolutional prototype learning," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, Dec. 2018, pp. 3474-3482.
[36] K. Cho, B. V. Merriënboer, C. Gulcehre *et al*., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1724-1734.
[37] J. J. Davis and M. H. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, New York, USA, Jun. 2006, pp. 233-240.
[38] H. Zhao, Y. Gao, H. Liu *et al.*, "Fault diagnosis of wind turbine bearing based on stochastic subspace identification and multi-kernel support vector machine," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 2, pp. 350-356, Apr. 2018.
[39] Z. Qu, H. Liu, Z. Wang *et al*., "A combined genetic optimization with AdaBoost ensemble model for anomaly detection in buildings electricity consumption," *Energy and Buildings*, vol. 248, p. 111193, Oct. 2021.

Xinwu Sun received the B.E. degree in electrical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2022. He is currently working toward the M.Sc. degree in electrical engineering in the same university. His research interests include electricity theft detection and applications of deep learning in power systems.

Jiaxiang Hu received the B.E. degree in electrical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2021. He is currently working toward the M.Sc. degree in electrical engineering in the same university. His research interests include fault diagnosis and applications of machine learning in power systems.

Zhengyuan Zhang received the Ph.D. degree in electrical engineering from The University of Texas at Arlington, Arlington, USA. He is currently working in the University of Electronics Science and Technology of China, Chengdu, China, as a Researcher. His research interests include hybrid energy storage, smart grids, renewable energy, electrical safety analysis, and power system analysis.

Di Cao received the Ph.D. degree in control science and engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2021. His research interests include optimization of distribution network and applications of machine learning in power systems.

Qi Huang received the B.S. degree in electrical engineering from Fuzhou University, Fuzhou, China, in 1996, the M.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 1999, and the Ph.D. degree in electrical engineering from Arizona State University, Tempe, USA, in 2003. He is currently a Professor with Southwest University of Science and Technology (SWUST) and University of Electronic Science and Technology of China, Chengdu, China. He is President of SWUST and the Director with the Sichuan State Provincial Lab of Power System Wide-area Measurement and Control. He is an IET Fellow and IEEE Fellow Member. His current research interests include power system instrumentation, power system monitoring and control, and informatics for smart electric energy systems.

Zhe Chen received the B.Eng. and M.Sc. degrees from the Northeast China Institute of Electric Power Engineering, Jilin, China, and the Ph.D. degree from the University of Durham, Durham, UK, in 1997. He is currently a Full Professor with the Department of Energy Technology, Aalborg University, Aalborg, Denmark. He is the Leader of the Wind Power System Research Program with the Department of Energy Technology, Aalborg University, and the Danish Principle Investigator for the Wind Energy of Sino-Dan-

ish Centre for Education and Research. He has led many research projects. He has authored or coauthored more than 400 publications in his technical field. His research interests include power systems, power electronics, and electric machines, and his current research interests include wind energy and modern power systems.

Weihao Hu received the B. Eng. and M. Sc. degrees from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2007, respectively, both in electrical engineering, and the Ph. D. degree from Aalborg University, Aalborg, Denmark, in 2012. He is currently a Full Professor and the Director of Institute of Smart Power and Energy Systems (ISPES) at the University of Electronics Science and Technology of China, Chengdu, China. He was an Associate Professor at the Department of Energy Technology, Aalborg University, and the Vice Program Leader of Wind Power System Research Program at the same department. He is an Associate Editor for IET Renewable Power Generation and Journal of Modern Power Systems and Clean Energy, He was serving as the Technical Program Chair (TPC) for IEEE Innovative Smart Grid Technologies (ISGT) Asia 2019 and is serving as the Conference Chair for the Asia Energy and Electrical Engineering Symposium (AEEES 2020). He is currently serving as Chair for IEEE Chengdu Section PELS Chapter and he is an IEEE Senior Member. His research interests include artificial intelligence in modern power systems and renewable power generation.