

Received April 26, 2020, accepted June 10, 2020, date of publication June 19, 2020, date of current version June 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3003778

A Load Identification Method Based on Active Deep Learning and Discrete Wavelet Transform

LUYANG GUO^{1,2}, SHOUXIANG WANG^{1,2}, (Senior Member, IEEE),
HAIWEN CHEN^{1,2}, AND QINGYUAN SHI³

¹Key Laboratory of Smart Grid of Ministry of Education, Tianjin University, Tianjin 300072, China

²Tianjin Key Laboratory of Power System Simulation and Control, Tianjin University, Tianjin 300072, China

³School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

Corresponding author: Haiwen Chen (244755391@qq.com)

This work was supported in part by the Tianjin Graduate Research and Innovation Project under Grant 2019YJSB188, and in part by the National Innovation Program for College Students of China and Science and Technology Project of State Grid Corporation of China.

ABSTRACT Non-Intrusive Load Monitoring (NILM) makes it possible for users and energy providers to track the fine-grained energy consumption information of residential and commercial buildings. The load identification methods in NILM usually require labeling many samples for training and evaluation, which is always expensive and time-consuming. In order to reduce the labeling cost, this paper proposed a load identification method based on Active Deep Learning (ADL). In this method, Discrete Wavelet Transform (DWT) was applied to extract high-dimensional appliance features from original current signals. Then a pool-based or stream-based active deep learning model was built to learn the features and select high-value samples that worthy of labeling. A mixed dataset based on three public datasets was formed to evaluate the proposed method and three sampling approaches of active learning. The results showed that the proposed method could significantly reduce labeling cost on large datasets, and the number of samples required is 33% lower than the state-of-the-art method when the F1 score is equal. Compared with pool-based sampling approaches, the stream-based approach's benefits are that the classifier improved and the query frequency decreased with continuous input of samples.


INDEX TERMS NILM, load identification, active deep learning, semi-supervised learning, CNN, pool-based sampling, stream-based sampling.

I. INTRODUCTION

Residential and commercial buildings account for more than 40% of global energy consumption and produce more than one-third of the total carbon dioxide emissions [1]. In order to improve energy efficiency, it is important to provide fine-grained energy consumption information for demand-side management [2]. Non-Intrusive Load Monitoring (NILM) can monitor the operation status and electricity consumption of appliances by analyzing the aggregated electrical signal [3]. The concept of NILM was first proposed by Hart and has attracted wide attention due to its low cost and high flexibility [4]. As an important step in NILM, load identification facilitates the interaction between energy providers and consumers, and the implementation of energy-saving

policies. Not only can the providers perform non-intrusive monitoring and operation using fine-grained information of consumers, but also the consumers can adjust their electricity consumption behavior according to the suggestions pushed by providers [5].

Load identification is commonly used in event-based NILM systems, where the operating state transition of an appliance is called an event. When an event is detected, load signatures can be extracted by analyzing the difference of electrical signal before and after the event [6]. In the machine learning context, these signatures are called features [7]. Then a classifier is used to identify which appliance caused the event. To make the classifier more robust, we need a lot of labeled appliance samples to train the classifier. The development of AMI allows users to participate in NILM, providing appliance data and gradually forming a large dataset [8]. However, the labeling of a large number of samples is always

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

an expensive and time-consuming process, which limits the scalability of NILM systems. Usually, the NILM datasets contain much redundancy, and not all samples are equally valuable for the training [9]. So, it is desired to minimize the labeling cost by labeling as little high-value data as possible.

Active learning is a machine learning technique that the model tries to query experts, such as appliance users, to get the labels of the most valuable samples. To our best knowledge, only ref [9]–[11] have discussed the active learning method for event-based NILM. Moreover, all these studies were based on low-dimensional power variation features and the training dataset, named BLUED, only contained a small number of appliances. It is difficult for power variation features to distinguish appliances with similar steady-state power or small power consumption, while large datasets usually contain many of these appliances [12].

Reducing the labeling cost in large datasets is the primary purpose of active learning, and performance across datasets is also an important criterion for method evaluation [13]. Reference [7] proved that the features containing spectral energy distribution or transient information have the best overall performance in various datasets. The original signal contains the most information, but it usually contains a great deal of redundancy, which increases the model complexity and weakens the accuracy. Discrete Wavelet Transform (DWT) is an efficient timing signal compression and noise reduction method, and it has properties like multi-resolution and time-frequency localization.

In general, features with more information have higher dimensions. Benefit from the flexible structure and strong capability in feature extraction, deep-learning methods are more applicable for high-dimensional features than traditional methods. In the field of event-based NILM, no study has discussed the feature application problem and active deep learning methods. Also, no study has discussed the sampling approaches of active learning for NILM. According to the above analysis, this paper proposed two high-dimensional DWT-based features for active learning and an active deep learning model that apply to high-dimensional features. Moreover, two pool-based and a stream-based sampling approaches of active learning for NILM were analyzed and compared. The contributions of this study are as follows.

- 1) A load identification method based on active deep learning and discrete wavelet analysis was proposed. Compared with existing methods, the proposed method has better performance of reducing the labeling cost on large datasets.
- 2) The applicability of the stream-based sampling approaches to NILM was discussed and validated for the first time. Compared with pool-based approaches, its benefits are that the classifier improved and the query frequency decreased significantly with the continuous input of samples.
- 3) A larger mixed dataset with the same data format is made to evaluate the proposed method. The mixed

dataset has about nine times as many appliances as the BLUED dataset used in other studies.

The rest of the paper is organized as follows. Section II provides a brief review of the background and related work. Section III describes the event-based NILM method based on active deep learning. Section IV mainly describes the features extraction method and the mixed dataset we made. Section V analyses the experiments we have done. Finally, Section VI concludes the paper.

II. BACKGROUND AND RELATED WORK

A. FEATURES OF EVENT-BASED NILM

Each appliance forms its unique features and can be seen as an individual fingerprint. At present, the commonly used features mainly include conventional features such as power features and current waveform [14]. In order to improve the performance of the conventional features, some improved features such as Resolution-Enhanced Admittance (REA) [5] and Voltage–Current (V–I) [15], [16] trajectory were proposed. Apart from the conventional features and improved features, the frequency domain features have also been widely used. Analysis tools such as the S transform [17], the Time–Time (TT) transform [18], the Short-Time Fourier Transform (STFT), and the DWT [19] have been used to get the features. In order to compare the performance of various features, ref [7] made a comprehensive comparison of the existing features on public datasets. The results showed that the current waveform and transient process are the main information for distinguishing different kinds of appliances. The features that contain the spectral energy distribution or transient information have the best overall performance in various datasets.

B. DATASETS OF EVENT-BASED NILM

High-frequency datasets are essential for training and evaluating event-based NILM methods. TABLE 1 shows several commonly used high-frequency NILM datasets. The subject of these datasets is mainly divided into two types: residence or individual appliances. We can see that it is a common strategy to collect many samples of each appliance, which allows the datasets to contain more information about each appliance. However, it may also make the datasets contain much redundancy.

At present, the evaluation of event-based NILM methods mainly faces two problems. First, because of the difference in data collection equipment, subject, and data storage mode, the data format is different between each high-frequency NILM dataset. Second, the number of appliances in most datasets is small. In order to solve these problems, ref [20] designed a set of features independent of the sampling frequency so that it could be applied to high-frequency datasets with different sampling frequency. Moreover, ref [20] proved that mixed datasets were a feasible way to evaluate event-based NILM methods. However, they did not integrate the data format. To unify the data format of low-frequency datasets, *Batra et al.*

TABLE 1. Statistics of high-frequency datasets.

Dataset	Sampling frequency	Subject	Number of appliances	Number of samples
REDD[23]	15kHz	residence	82	Months
BLUED[24]	12kHz	residence	43	2300
UK-DALE[25]	16kHz	residence	53	2 years
PLAID[39]	30kHz	individual	317	1793
WHITED[26]	44kHz	individual	110	1339
COOL[27]	100kHz	individual	42	840

developed the NILMTK toolkit [21]. Unfortunately, it can only work on low-frequency datasets. At present, no toolkit can integrate the data format of high-frequency datasets.

C. ACTIVE LEARNING FOR EVENT-BASED NILM

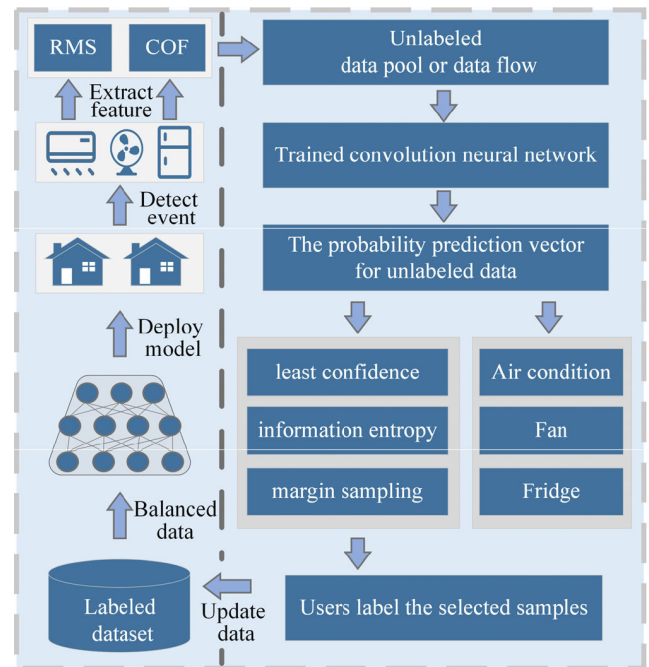
At present, there are two common methods to reduce the labeling cost for NILM: semi-supervised learning and active learning. Semi-supervised learning uses a small number of labeled samples and a large number of unlabeled samples to train the model. It exploits the most confident unlabeled samples to improve the performance of the current model [27]. References [28], [29] proved that semi-supervised learning could improve performance and reduce the labeling cost.

For datasets that contain much redundancy, active learning may be more applicable to reducing the labeling cost. It only focuses on the high-value samples and discards the low-value samples. According to the difference in sampling approaches, active learning can be divided into two types: pool-based sampling and stream-based sampling [30]. Pool-based sampling assumes that there is an unlabeled sample pool, and selects those samples that are most informative from the pool. Stream-based sampling assumes that the samples are seen by the model one by one, and selects samples according to the informativeness threshold. The selection of each sample is independent of other samples. There are three studies about active learning for event-based NILM. Reference [10] proposed an active learning framework. Reference [9] discussed the performance of different query strategies based on [10]. Reference [11] proposed an RF-based hybrid method of combining active learning with self-training semi-supervised learning. All these methods are based on low-dimensional power variation features and the BLUED dataset with a small number of appliances. Also, no study has discussed the sampling approaches of active learning in NILM.

III. THE PROPOSED ACTIVE DEEP LEARNING MODEL FOR LOAD IDENTIFICATION

A. WORKFLOW OF THE PROPOSED METHOD

The workflow of the proposed method is shown in Figure 1. The difference between pool-based sampling and stream-based sampling is the way of the model gets and queries unlabeled samples. We used CNN as the classifier for the active learning model. COF and RMS denote the DWT-

**FIGURE 1. The workflow of the proposed method.**

based features we used, and their extraction methods are in section IV. It is worth noting that we mainly focus on the training and classification process of event-based NILM. In order to avoid the influence of the event detector, we assume that there is a perfect event detector that can detect the operating state transition of an appliance and extract the voltage and current signal of the appliance from the aggregated signal. At present, many high-frequency datasets are made based on this assumption, and they only collect the voltage and current signal of a single appliance instead of the aggregated signal of all appliances.

B. ACTIVE LEARNING MODEL

1) INITIALIZATION METHOD

It is a common method to select samples randomly from the unlabeled set to form the initial labeled set. However, this method is not stable. To solve this problem, we used the AP clustering method to select samples from the unlabeled set to form the initial labeled set. Compared with conventional methods such as k-means and DBSCAN, the clustering number of AP clustering does not need to be assigned at first. When the AP clustering method is executed many times, the results are exactly the same, which means that the sample groups obtained from clustering are more stable [31]. We selected the same number of samples from each group after clustering, and a total of 200 samples were labeled as the initial labeled set. The clustering process is as follows.

First, suppose there are N samples in the unlabeled set. Obtain each sample and denote it as X_i ($i = 1, \dots, N$). Calculate the similarity (Euclidean distance in this paper) between every two samples and get the similarity matrix S of $N \times N$.

Clustering is achieved through the steps of message passing. Suppose there are two key matrices: responsibility matrix ($\mathbf{R} = [r(i, k)]_{N \times N}$) and availability matrix ($\mathbf{A} = [a(i, k)]_{N \times N}$). $r(i, k)$ denotes the degree to which sample k is suitable as the class representative point (exemplar) of sample j . $a(i, k)$ denotes the suitability of sample i to select sample k as the exemplar. The calculation formulas of \mathbf{R} and \mathbf{A} is as follows:

$$r(i, j) = s(i, j) - \max\{a(i, k) + s(i, k)\},$$

$$k \in 1, 2, \dots, N \text{ and } k \neq j \quad (1)$$

$$a(i, j) = \begin{cases} \min\{0, r(j, j) + \sum_k \max\{0, r(k, j)\}\}, \\ k \in 1, 2, \dots, N \text{ and } k \notin \{i, j\} \\ \sum_k \max(0, r(k, i)), \\ k \in 1, 2, \dots, N \text{ and } k = i \end{cases} \quad (2)$$

Alternately update \mathbf{A} and \mathbf{R} until the class representative points tend to be stable. Then get the clustering center of each sample (typical sample for each group) according to the following formula:

$$k = \operatorname{argmax}_k \{a(i, k) + r(i, k)\},$$

$$i = 1, 2, \dots, N, \quad k \in 1, 2, \dots, N \quad (3)$$

If $i = k$, the point X_i itself is the clustering center, if $i \neq k$, the X_k is the clustering center of point X_i . The number of clustering centers is equal to the number of groups, and the X_k is the typical sample corresponding to each group.

2) QUERY STRATEGY

Each query strategy uses an informativeness metric $H(x_i)$ to indicate how valuable sample x_i is for the training of the classifier. Query strategy based on uncertainty is the most commonly used. These query strategies take uncertainty as the informativeness of the sample. We used the following three uncertainty-based query strategies in this paper:

a: LEAST CONFIDENCE SAMPLING [32]

Least confidence sampling selects those samples, for which the maximum class probability is minimal, which indicates that the model has not learned enough features of such samples. The informativeness metric $H(x_i)$ is calculated as:

$$H(x_i) = 1 - P(y = y_1^* | x_i), \quad i = 1 \dots m \quad (4)$$

b: INFORMATION ENTROPY SAMPLING [32]

The information entropy is maximum when the probability of all categories is equal. The Information entropy is minimum when there is a class of probability 1. The information entropy sampling selects those samples with the maximum information entropy. The informativeness metric $H(x_i)$ is calculated as:

$$H(x_i) = - \sum_{n=1}^N p(y = n | x_i) \log[p(y = n | x_i)], \quad i = 1 \dots m \quad (5)$$

TABLE 2. The structure of CNN.

Layer type	Output dimensions	Kernel size	activation
Input	(None,143)	—	—
Conv1D	(None,141,64)	3(64)	Relu
Conv1D	(None,139,64)	3(64)	Relu
Conv1D	(None,137,64)	3(64)	Relu
Dense	(None,137,128)	128	Relu
Dropout(0.5)	(None,137,128)	—	—
Flatten	(None,17536)	—	—
Dense	(None,11)	—	Softmax

c: MARGIN SAMPLING [33]

The margin sampling selects those samples closest to the boundary of the first and second most probable class. The calculation formula is as follows: The informativeness metric $H(x_i)$ is calculated as:

$$H(x_i) = 1 - (P(y = y_1^* | x_i) - P(y = y_2^* | x_i)), \quad i = 1 \dots m \quad (6)$$

where x_i denotes the i -th sample in the unlabeled set, y denotes the class label of x_i , y_a^* denotes the class label with the a -th highest probability in the prediction probability vector, N denotes the number of classes, m denotes the number of samples of the unlabeled set, $H(x_i)$ denotes the amount of information of the i -th sample, $P(y | x_i)$ denotes the probability of sample x_i belonging to the y class.

C. STRUCTURE OF CNN

Benefit from the flexible structure and strong capability in feature extraction, deep learning is more applicable to high-dimensional features than traditional machine learning. So we used CNN as the classifier of the active learning model. Furthermore, to avoid the overfitting problem, we added a dropout layer to the hidden layer. There is no uniform standard for the structural design of the neural network. We compare many different network structures and design a lightweight network structure without reducing the model accuracy. The output layer activation function of the network is softmax, and the output is an 11-dimensional probability vector. Each value in the vector is a probability value. In the process of training CNN, we used Cross-Entropy Loss (CEL) as the loss function and adaptive moment (Adam) as the optimizer. The network structure of CNN is shown in TABLE 2.

IV. FEATURE EXTRACTION AND DATA PROCESSING METHODS

A. FEATURE EXTRACTION BASED ON DWT

DWT is an efficient timing signal compression and noise reduction method, and it has properties like multi-resolution and time-frequency localization. We proposed two methods of features extraction based on DWT:

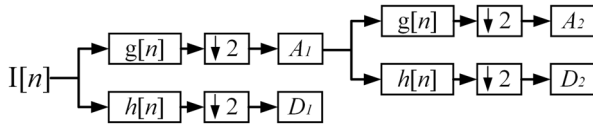


FIGURE 2. The principle of wavelet analysis.

1) APPROXIMATION COEFFICIENTS FEATURES

In this paper, the DWT approximation coefficients of the current signals are called the COF features. DWT decomposes the current signal into a coarse approximation and fine detail at different levels of decomposition. Approximation coefficients represent the large-scale, low-frequency components, while detail coefficients are small-scale, high-frequency components. Its principle is shown in Figure 2. The low-frequency components are filtered once at each level. Although increasing the decomposition level reduces the data dimensionality more, similarity to the original signal usually decreases [34].

Where A_r and D_r denote the low-frequency and high-frequency components of level r , $I[n]$ denotes the n th sampling point of the discrete current signal, $g[n]$ denotes a low-pass filter that can filter out the high-frequency part of the current signal, $h[n]$ denotes a high-pass filter that can filter out the low-frequency part of the current signal, $\downarrow 2$ denotes a second-order reduced sampling filter. The relationship between the current signal, components, and coefficients is as follows:

$$I[n] = \sum_{r=1}^R D_r^n + A_r^n \tag{7}$$

$$\begin{cases} A_r^n = \sum_f a_{rf} \varphi_n^r, & n = 1, 2, \dots, N \\ D_r^n = \sum_f d_{rf} \phi_n^r, & n = 1, 2, \dots, N \end{cases} \tag{8}$$

where r denotes the decomposition level, f denotes the number of components, A_r^n and D_r^n denote the low-frequency components and detail components of level r at the n th point, φ_n^r and ϕ_n^r denote the scale function and the mother wavelet function. a_{rf} and d_{rf} denote the low-frequency coefficient and the high-frequency coefficient. The coefficient calculation formula can be seen in the literature [35].

The current waveform of 25 periods after an event was used as the original signal. Firstly, the frequency of the original signal was reduced to 40 points per period by uniformly spaced sampling. After reducing the sampling rate, the current signal of 25 periods contains 1000 points. Then, the ‘db4’ wavelet was used to construct the current signal, and the third level approximation coefficients that contain 143 points was taken as the COF features. Figure 3 shows the comparison between the reconstructed signal using the approximation coefficients and the original signal of a vacuum. It shows that the proposed COF features can achieve efficient data compression with a small loss.

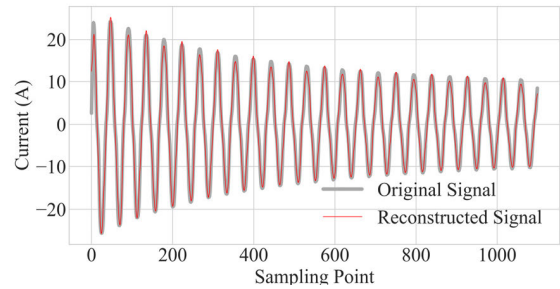


FIGURE 3. The contrast of reconstructed low-frequency components with the original waveform.

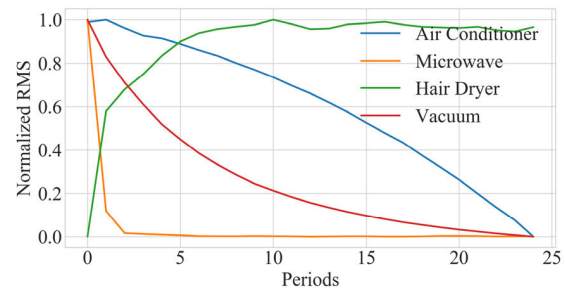


FIGURE 4. The RMS features of different types of appliances.

2) RMS FEATURES

In this paper, the Root Mean Square (RMS) of low-frequency components is called RMS features. Different types of appliances have different transient processes during state switching. The purpose of removing the high-frequency components is to improve the stability of the features and reduce the intra-class differences. First, the low-frequency components were reconstructed using the approximate coefficients extracted in the previous section. Then the RMS of each period of the low-frequency components were calculated as follows:

$$I_{P(b)} = \sqrt{\frac{\sum_{m=1}^N i_m^2}{N}} \tag{9}$$

$$COT_{25} = [I_{P(1)}, I_{P(2)}, \dots, I_{P(25)}] \tag{10}$$

where $I_{P(b)}$ denotes the RMS of low-frequency components in the period b , i_m denotes the i -th low-frequency components value in a period, and N denotes the number of sampling points in each period. Figure 4 shows the RMS features of several typical appliances, indicating a unique variation trend for different types of appliances. It should be noted that when drawing Figure 4, RMS features were normalized, but not normalized during actual use.

B. THE FUSION OF DATASETS

Reference [20] proved that mixed datasets were a feasible way to evaluate event-based NILM methods. We combined

TABLE 3. Appliance and sample information of the mixed dataset.

Type	PLAID	WHITED	COOLL	Total devices	Total samples
AC	26	1	0	27	218
CFL	44	2	0	46	240
Fan	30	6	2	38	310
Fridge	27	1	0	28	100
HD	36	7	4	47	398
Heater	15	25	0	40	334
ILB	32	7	0	39	218
Laptop	45	2	0	47	227
Microwave	32	3	0	35	259
Vacuum	14	4	7	25	253
WM	16	1	0	17	85
Total	317	59	13	389	2642

AC=Air Conditioner, CFL=Compact Fluorescent Light, HD=Hair Dryer, ILB=Incandescent Light Bulb, WM=Washing Machine.

the datasets of PLAID, COOLL, and WHITED into a larger mixed dataset with the same data format. This mixed dataset only contains commonly used household appliances. The other datasets in TABLE 1 did not collect signals of individual appliances, so they cannot be combined into this mixed dataset. TABLE 3 shows the detailed information of this dataset. Since the data storage format of each dataset is different, we did the following data processing in the combination process:

- 1) Sorting out appliance categories and label formats,
- 2) Uniform sampling frequency to 20kHz by uniformly spaced sampling,
- 3) The sampling duration is 2s after the event,
- 4) Uniform the units of current and voltage to A and V.

C. PROCESSING OF UNBALANCED DATA

One of the characteristics of this mixed dataset is that the number of samples of each class varies significantly. The CFL with the largest sample size has 398 samples, while the washing machines with the smallest sample size only has 85 samples. Classifiers may not be able to learn enough features from samples of the minority class, so the processing of unbalanced data before training can achieve better performance, which was proved in section V. We used several SMOTE-based data processing methods to expand the samples of the minority class in the dataset. The details of these methods can be seen in [36].

V. EXPERIMENT AND ANALYSIS

A. EXPERIMENT DESIGN AND EVALUATION METRIC

To evaluate the performance of the proposed method, we compared the following five methods:

- 1) Active learning method: the proposed method
- 2) Supervised method: the proposed method without the active learning process.

- 3) Semi-supervised method: replace the active learning process with commonly used semi-supervised learning based on self-training in the proposed method.
- 4) Semi-supervised method: the state-of-the-art semi-supervised learning based on co-training of DT and KNN proposed by ref [29].
- 5) Active learning method: the state-of-the-art RF-base hybrid method of combining active learning with self-training semi-supervised learning proposed by ref [11].

Since the datasets, data processing methods, and features used in these methods are different, and power variation features have been proved unable to perform well in appliance identification [12]. These differences will make the comparison unfair. Therefore, we only repeat their core methods for comparison while keeping the other experimental setups unchanged. We designed three experiments:

- 1) Experiment I used all the samples in the training set to compare the performance of different classifiers used in these methods and evaluate the effect of data processing methods. The experimental results served as the performance benchmark for subsequent experiments.
- 2) Experiment II compared the performance of these methods in reducing the labeling cost. Pool-based sampling can accurately control the number of samples per query, so we used pool-based sampling in experiment II.
- 3) Experiment III analyzed and compared three sampling approaches from the following three aspects: a. the influence of parameter selection on each approach; b. the number of samples selected between these approaches; c. the comparison of the F1 score increase speed.

In each experiment, 20% of the samples from each class were randomly selected as the test set. We repeated each experiment ten times and calculated the average of the ten experiments for comparison. We trained the CNN 300 epochs and used the weight with the best performance for comparison.

F1 score is a commonly used evaluation metric, which can reflect the performance of the method more comprehensively. Therefore, we use the F1 score to evaluate the overall performance of the method. The calculation formula is as follows:

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (11)$$

$$\text{precision} = \frac{TP}{TP + FP} \quad (12)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (13)$$

where TP denotes the correct number of samples in this class, FP denotes the number of samples of other classes predicted as this class, FN denotes the number of samples of this class predicted as other classes.

TABLE 4. Comparison of classifiers and unbalanced data processing methods.

features	strategy	DT [38]	KNN [38]	RF [11]	CNN
RMS	nothing	0.844	0.815	0.863	0.873
	random	0.842	0.848	0.876	0.893
	SMOTE	0.841	0.852	0.865	0.904
	BDS	0.843	0.852	0.874	0.904
	SVMS	0.851	0.864	0.883	0.91
	SEE	0.796	0.794	0.806	0.855
	SEO	0.829	0.835	0.88	0.888
COF	nothing	0.671	0.703	0.739	0.88
	random	0.658	0.684	0.729	0.9
	SMOTE	0.662	0.71	0.715	0.898
	BDS	0.671	0.705	0.734	0.901
	SVMS	0.666	0.709	0.746	0.903
	SEE	0.577	0.644	0.631	0.767
	SEO	0.653	0.713	0.718	0.89

BDS= Borderline-SMOTE,SVMS=SVM-SMOTE, SEE=SMOTE-EEN, SEO=SMOTE-OMEK

B. EXPERIMENT I: BENCHMARKING WHEN USING ALL SAMPLES OF THE TRAINING SET

In experiment I, we used all the samples in the training set to compare the performance of different classifiers and evaluate the effect of data processing methods. The experimental results served as the performance benchmark for subsequent experiments. The results are shown in TABLE 4.

When using the low-dimensional RMS features, all classifiers have a high F1 score. When using the high-dimensional COF features, only CNN continues to maintain a high F1 score. Experimental results prove that CNN has the best identification performance and is more applicable to high-dimensional features than traditional machine learning. Moreover, the two proposed high-dimensional features have excellent identification performance on large datasets.

In most cases, BDSMOTE and SVMSMOTE have better performance because they tend to generate samples near the boundary, which are more valuable to the classifiers. SMOTEEEN always has the worst performance because it copied some low-value samples and deleted some high-value samples, causing the classifiers confusion.

C. EXPERIMENT II: PERFORMANCE COMPARISON WITH SUPERVISED, SEMI-SUPERVISED, AND OTHER ACTIVE LEARNING METHODS

In experiment II, we compared the proposed method with other methods in reducing the labeling cost. Pool-based sampling can accurately control the number of samples per query, so we used pool-based sampling in experiment II. According to the experiment I, we processed the unbalanced data using the SVMSMOTE method.

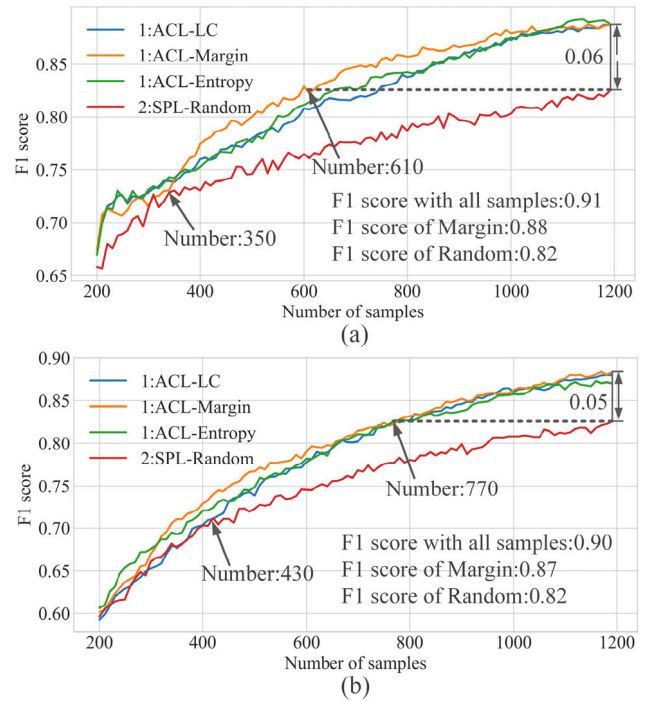


FIGURE 5. The comparison of the active deep learning method based on three query strategies with the supervised deep learning method. (a) The result of using the RMS features. (b) The result of using the COF features.

We first compared the performance of the active and supervised deep learning model, then evaluated the effect of different query strategies. We used AP clustering to select and label 200 samples from the unlabeled set as the initial labeled set. The active learning queried ten samples each time and updated the parameters of CNN. Supervised deep learning always randomly labeled the same number of samples. The experimental results are shown in Figure 5. Where ACL denotes the Active Learning, SPL denotes the Supervised Learning. The number before the legend represents the label of the five methods we compared.

It can be seen that the F1 score of active deep learning significantly higher than supervised deep learning. Moreover, the proposed method used only about 700 samples to achieve the F1 score of supervised deep learning used 1200 samples. We speculate that the reason is when the accuracy of CNN is poor, most samples are high value for the improvement of model accuracy. With the improvement of the accuracy of CNN, it is difficult to select high-value samples through random selection, while active learning can always select high-value samples. Finally, using less than half of all samples for training, the proposed method achieved 96% of the F1 score when using all samples. The results prove that the proposed method can significantly reduce the labeling cost on large datasets.

Then we compared the proposed active deep learning method with other semi-supervised and active learning methods. In [29], DT and KNN were co-trained, and each classifier used one feature. In order to compare the performance of

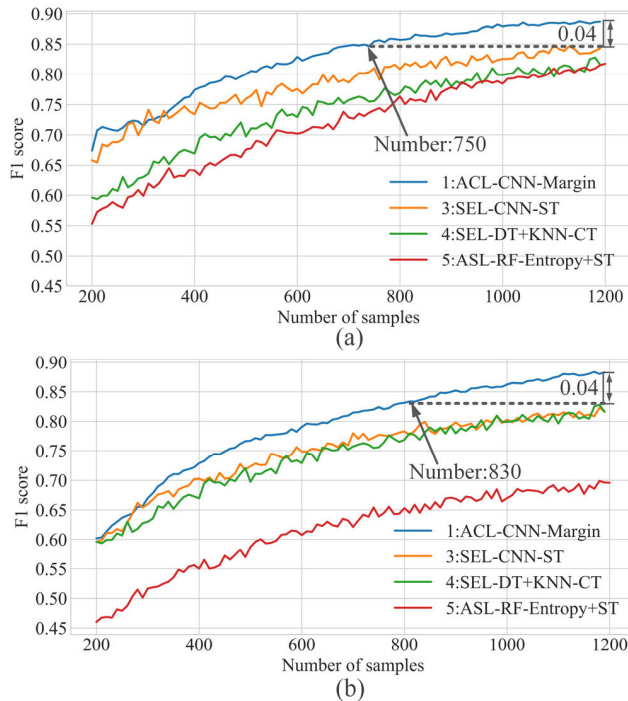


FIGURE 6. The comparison results of the F1 score growth curve between the proposed method and other state-of-the-art methods. (a) the result of using the RMS features, (b) The result of using the COF features.

these methods with the same labeling cost, semi-supervised learning randomly labeled samples with the same number of active learning, then used the rest unlabeled samples to improve the performance of the model. The margin query strategy performed well most of the time in the previous experiment, so we used the margin query strategy for comparison. The experimental results are shown in Figure 6.

Moreover, in order to evaluate these methods more comprehensively, we added two metrics: “Average F1 score” and “Ratio to the highest F1 score”. “Average F1 score” denotes the average value of the F1 score growth curve, and it reflects the average accuracy of each method during training. “Ratio to the highest F1 score” denotes the ratio of the F1 score when 1200 samples were used for training to the F1 score when all samples were used, and it reflects the relative learning rate of each method. The experimental results are shown in Figure 7. Where SEL denotes the Semi-supervised Learning, and ASL denotes the hybrid method of combining Active Learning with Semi-supervised Learning.

As can be seen from Figure 6, the F1 score of the proposed method is significantly higher than other methods, and it only used about 800 labeled samples to achieve the F1 score of the state-of-the-art method used 1200 labeled samples. The required number of samples was reduced by 33%. As can be seen from Figure 7, CNN-based methods, No. 2 and No. 3, have a higher average F1 score, but they have the lowest ratio to the highest F1 score. The hybrid method, No. 5, which combines active learning with semi-supervised learning, consistently performs poorly. However, the proposed

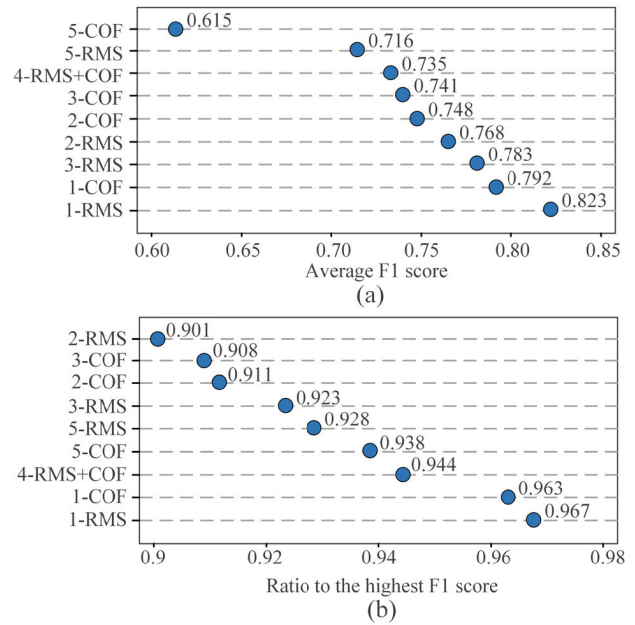


FIGURE 7. The comparison results of the two metrics between the proposed method and other state-of-the-art methods. (a) the “Average F1 score”, (b) the “Ratio to the highest F1 score.”

active deep learning method not only has the highest average F1 score but also has the highest ratio to the highest F1 score. Experiment II proves the superiority of the proposed active deep learning method, which is not determined by CNN or active learning alone, but the result of active learning and deep learning working together.

D. EXPERIMENT III: ANALYSIS OF DIFFERENT SAMPLING APPROACHES

In experiment III, we analyzed and compared three sampling approaches from the following three aspects: a. the influence of parameter selection on each approach; b. the number of samples selected between these approaches; c. the comparison of the F1 score increase speed. RMS features and COF features have similar performance in experiments I and II, so only higher dimensional COF features were used in this experiment. The margin query strategy performed well most of the time in experiment II, so we used the margin query strategy in this experiment. The three approaches are as follows:

- 1) The pool-based approach with a fixed number of samples selected (Approach 1): The entire unlabeled set is taken as a sample pool, from which a fixed number of samples are selected and labeled according to an informativeness metric. Then update the model and repeat these steps until 1000 samples are selected.
- 2) The pool-based approach with a fixed sample selection threshold (Approach 2): The entire unlabeled set is taken as a sample pool, from which all samples with an informativeness metric above a fixed threshold are selected and labeled. Then update the model and repeat these steps until 20 iterations are completed.

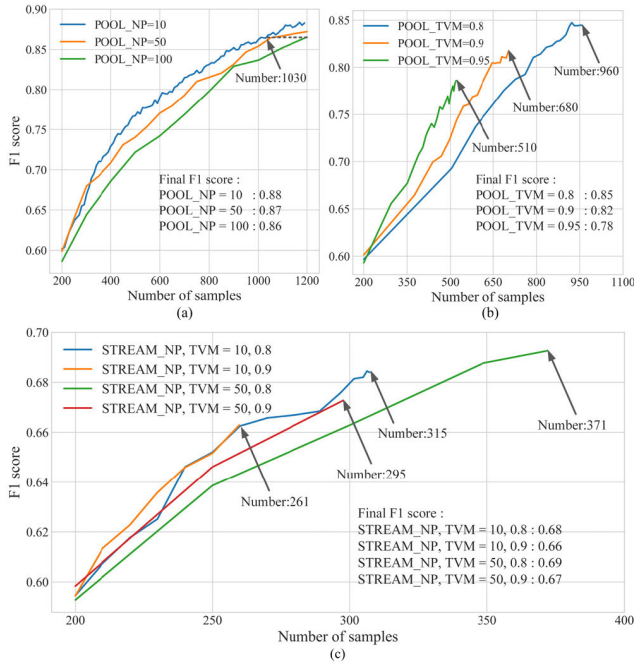


FIGURE 8. The relationship between parameters and the F1 score growth curve of each approach. (a) Approach 1, (b) Approach 2, (c) Approach 3.

3) The stream-based approach (Approach 3): The samples in the unlabeled set are input to the model one by one, and each sample is labeled or discarded according to a fixed informativeness threshold, and update the model after a fixed number of samples have been labeled. Then repeat these steps until no samples that informativeness above the threshold can be found, or all samples in the unlabeled set have been input to the model. This approach does not depend on a pool of samples but rather a stream of samples. Therefore, it is useful for settings like NILM, where new samples arrive over time.

It can be seen that these three approaches mainly have two parameters: the number of samples selected per model update (NP) and the threshold value of margin query strategy (TVM).

The relationship between parameters and the F1 score growth curve of each approach is shown in Figure 8. It can be seen that both NP and TVM influence the F1 score increase speed of each approach. With the decrease of NP or the increase of TVM, the F1 score of each approach will increase faster. The main reason is that NP determines the frequency of model updates, and TVM determines the informativeness of labeled samples. For Approach 2 and 3 with threshold parameters, the faster the F1 score increases, the lower the total number of samples selected and the F1 score reached. It proves that only high-information samples cannot make the model reach the highest accuracy, and the total number of samples selected is also an essential factor to determine model accuracy.

The relationship between parameters and the number of samples selected by each approach is shown in Figure 9.

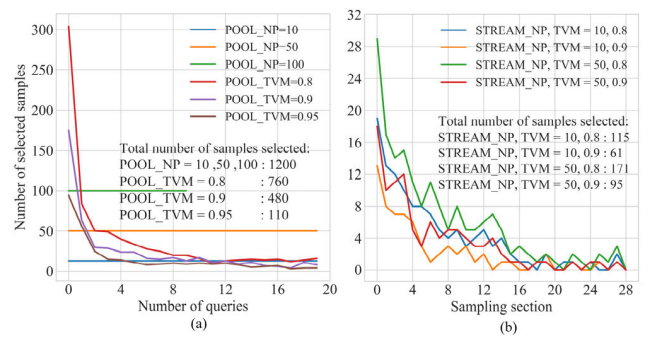


FIGURE 9. The relationship between parameters and the number of samples selected by each approach. (a) Approach 1 and Approach 2, (b) Approach 3.

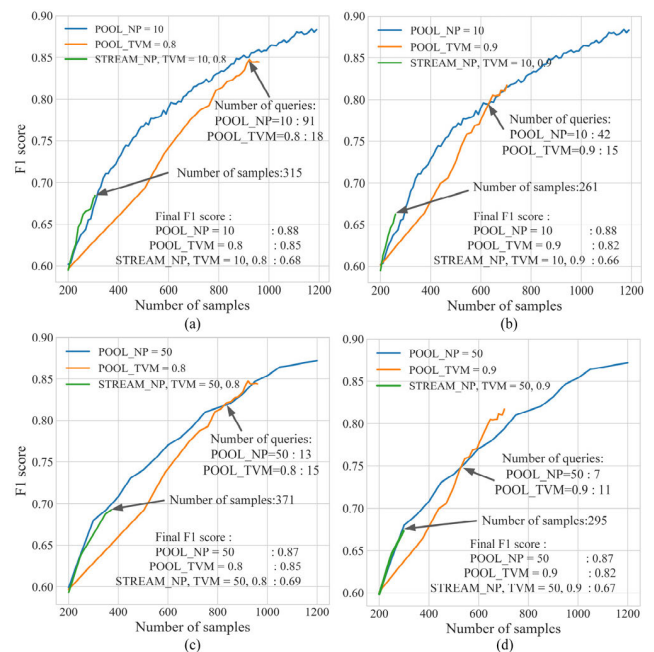


FIGURE 10. The F1 score increase speed comparison between these approaches. (a) NP = 10, TVM = 0.8, (b) NP = 10, TVM = 0.9, (c) NP = 50, TVM = 0.8, (d) NP = 50, TVM = 0.9.

Where Figure 9 (b) segmented the unlabeled samples according to the order in which they were input to the model, with 100 samples in each segment. It can be seen that the number of samples selected by Approach 2 and 3 rapidly decreased as training progressed. Moreover, the lower the NP or higher the TVM, the lower the number of samples selected at each stage.

The comparison of the F1 score growth curve between these approaches is shown in Figure 10. As can be seen from figures 9 and 10. For pool-based approaches, the F1 score of Approach 1 increased faster at the beginning of training and then gradually decreased, while the F1 score of Approach 2 increased at a more stable speed. Moreover, with the sample size grows, the F1 score of Approach 2 all reached or exceeded Approach 1, and Approach 2 used fewer iterations when NP = 10. The advantage of Approach 1 is

that when Approach 2 was difficult to select samples, it could still select a fixed number of samples to continue training the model, so it reached a higher F1 score finally.

For stream-based Approach 3, when the parameters are equal, the F1 score increase speed is closer to that of Approach 1 at the beginning of training, but the F1 score and the total number of samples selected is significantly less than other approaches. Its characteristic is suitable for the NILM system, where new samples arrive over time. Its benefits are that the classifier improved and the query frequency decreased significantly with the continuous input of samples. Moreover, the low-information samples are not saved and consequently do not consume storage space.

VI. CONCLUSION AND FUTURE RESEARCH

This paper proposed a load identification method based on active deep learning and discrete wavelet transform. We made a large mixed dataset and designed three experiments to evaluate the proposed method and different sampling approaches of active learning.

Experiment I proved that CNN has the best identification performance and is more applicable to high-dimensional features than traditional machine learning. Moreover, the two proposed high-dimensional features have excellent identification performance on large datasets. Experiment II proved that the proposed active deep learning method could significantly reduce the labeling cost on large datasets, and has the best comprehensive performance compared to similar approaches, which is the result of active learning and deep learning working together. Experiment III discussed and validated the applicability of the stream-based sampling approaches to NILM for the first time. Moreover, compared with pool-based approaches, its benefits are that the classifier improved and the query frequency decreased significantly with the continuous input of samples. In the future, we will study the practical application of the NILM system based on active learning.

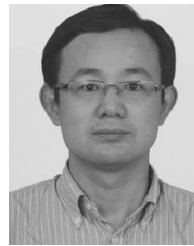
REFERENCES

- [1] S.-H. Yoon, S.-Y. Kim, G.-H. Park, Y.-K. Kim, C.-H. Cho, and B.-H. Park, "Multiple power-based building energy management system for efficient management of building energy," *Sustain. Cities Soc.*, vol. 42, pp. 462–470, Oct. 2018.
- [2] Y. Liu, X. Wang, and W. You, "Non-intrusive load monitoring by voltage-current trajectory enabled transfer learning," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5609–5619, Sep. 2019.
- [3] D. Balsamo, G. Gallo, D. Brunelli, and L. Benini, "Non-intrusive Zigbee power meter for load monitoring in smart buildings," in *Proc. IEEE Sensors Appl. Symp. (SAS)*, Zadar, Croatia, Apr. 2015, pp. 1–6.
- [4] G. W. Hart, "Non-intrusive appliance load monitoring," *Proc. IEEE*, vol. 80, no. 12, pp. 1870–1891, Dec. 1992.
- [5] Y. Liu, X. Wang, L. Zhao, and Y. Liu, "Admittance-based load signature construction for non-intrusive appliance load monitoring," *Energy Buildings*, vol. 171, pp. 209–219, Jul. 2018.
- [6] J. Gao, E. C. Kara, S. Giri, and M. Berges, "A feasibility study of automated plug-load identification from high-frequency measurements," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Orlando, FL, USA, Dec. 2015, pp. 220–224.
- [7] M. Kahl, A. U. Haq, T. Kriechbaumer, and H.-A. Jacobsen, "A comprehensive feature study for appliance recognition on high frequency energy data," in *Proc. 8th Int. Conf. Future Energy Syst.*, Hong Kong, May 2017, pp. 121–131.
- [8] A. L. Wang, B. X. Chen, C. G. Wang, and D. Hua, "Non-intrusive load monitoring algorithm based on features of V-I trajectory," *Electr. Power Syst. Res.*, vol. 157, pp. 134–144, Apr. 2018.
- [9] F. Liebgott and B. Yang, "Active learning with cross-dataset validation in event-based non-intrusive load monitoring," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Kos, Greece, Aug./Sep. 2017, pp. 296–300.
- [10] X. Jin, "An active learning framework for non-intrusive load monitoring," in *Proc. 3rd Int. Workshop Non-Intrusive Load Monit.*, Vancouver, BC, Canada, May 2016, pp. 1–7.
- [11] A. M. Fatouh, O. A. Nasr, and M. M. Eissa, "New semi-supervised and active learning combination technique for non-intrusive load monitoring," in *Proc. IEEE Int. Conf. Smart Energy Grid Eng. (SEGE)*, Oshawa, ON, Canada, Aug. 2018, pp. 181–185.
- [12] S. Djordjevic, M. Dimitrijevic, and V. Litovski, "A non-intrusive identification of home appliances using active power and harmonic current," *Facta Univ.-Ser., Electron. Energetics*, vol. 30, no. 2, pp. 199–208, Jun. 2017.
- [13] J. Kelly and W. Knottenbelt, "Metadata for energy disaggregation," in *Proc. IEEE 38th Int. Comput. Softw. Appl. Conf. Workshops*, Västerås, Sweden, Jul. 2014, pp. 578–583.
- [14] H. Liu, H. Wu, and C. Yu, "A hybrid model for appliance classification based on time series features," *Energy Buildings*, vol. 196, pp. 112–123, Aug. 2019.
- [15] X. Wu, D. Jiao, K. Liang, and X. Han, "A fast online load identification algorithm based on V-I characteristics of high-frequency data under user operational constraints," *Energy*, vol. 188, Dec. 2019, Art. no. 116012.
- [16] L. De Baets, J. Ruysinck, C. Develder, T. Dhaene, and D. Deschrijver, "Appliance classification using VI trajectories and convolutional neural networks," *Energy Buildings*, vol. 158, pp. 32–36, Jan. 2018.
- [17] H. Ahmadi and J. R. Marti, "Load decomposition at smart meters level using eigenloads approach," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3425–3436, Nov. 2015.
- [18] K. Khalid, A. Mohamed, R. Mohamed, and H. Shareef, "Nonintrusive load identification using extreme learning machine and TT-transform," in *Proc. Int. Conf. Adv. Electr., Electron. Syst. Eng. (ICAEEES)*, Nov. 2016, pp. 271–276.
- [19] Y.-C. Su, K.-L. Lian, and H.-H. Chang, "Feature selection of non-intrusive load monitoring system using STFT and wavelet transform," in *Proc. IEEE 8th Int. Conf. e-Bus. Eng.*, Beijing, China, Oct. 2011, pp. 293–298.
- [20] M. Kahl, T. Kriechbaumer, A. U. Haq, and H.-A. Jacobsen, "Appliance classification across multiple high frequency energy datasets," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Oct. 2017, pp. 147–152.
- [21] J. Kelly, N. Batra, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava, "NILMTK v0.2: A non-intrusive load monitoring toolkit for large scale data sets," in *Proc. 1st ACM Conf. Embedded Syst. Energy-Efficient Buildings (BuildSys)*, New York, NY, USA, Nov. 2014, pp. 182–183.
- [22] J. Z. Kolter and M. J. Johnson, "Redd: A public data set for energy disaggregation research," *Artif. Intell.*, vol. 25, pp. 1–6, Jan. 2011.
- [23] K. Anderson, A. F. O'Connell, D. Benitez, D. Carlson, A. Rowe, and M. Bergés, "BLUED: A fully labeled public dataset for event-based non-intrusive load monitoring research," in *Proc. 2nd KDD Workshop Data Mining Appl. Sustainability (SustKDD)*, Beijing, China, 2012, pp. 1–5.
- [24] J. Kelly and W. Knottenbelt, "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes," *Sci. Data*, vol. 2, no. 1, Mar. 2015, Art. no. 150007.
- [25] M. Kahl, A. U. Haq, T. Kriechbaumer, and H. A. Jacobsen, "WHITED—A worldwide household and industry transient energy data set," in *Proc. 3rd Int. Workshop Non-Intrusive Load Monit.*, Vancouver, BC, Canada, 2016, pp. 1–5.
- [26] T. Picon, M. N. Meziane, P. Ravier, G. Lamarque, C. Novello, J.-C. Le Bunetel, and Y. Raingeaud, "COOLL: Controlled on/off loads library, a public dataset of high-sampled electrical signals for appliance identification," 2016, *arXiv:1611.05803*. [Online]. Available: <http://arxiv.org/abs/1611.05803>
- [27] T. Ma, J. Ge, and J. Wang, "Combining active learning and semi-supervised for improving learning performance," in *Proc. 4th Int. Symp. Appl. Sci. Biomed. Commun. Technol. (ISABEL)*, Barcelona, Spain, Oct. 2011, pp. 1–5.
- [28] Y. Yang, J. Zhong, W. Li, T. A. Gulliver, and S. Li, "Semi-supervised multi-label deep learning based non-intrusive load monitoring in smart grids," *IEEE Trans. Ind. Informat.*, early access, Nov. 25, 2019, doi: 10.1109/TII.2019.2955470.

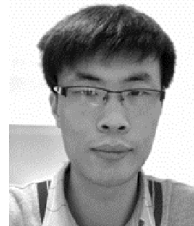
- [29] J. M. Gillis and W. G. Morsi, "Non-intrusive load monitoring using semi-supervised machine learning and wavelet design," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2648–2655, Nov. 2017.
- [30] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. 1648, 2010. [Online]. Available: <http://burrsettles.com/pub/settles.activelearning.pdf>
- [31] H. Chen, S. Wang, S. Wang, and Y. Li, "Day-ahead aggregated load forecasting based on two-terminal sparse coding and deep neural network fusion," *Electr. Power Syst. Res.*, vol. 177, Dec. 2019, Art. no. 105987.
- [32] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proc. 17th Annu. Int. ACM SIGIR*, Aug. 1994, pp. 3–12.
- [33] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden Markov models for information extraction," in *Proc. 4th Int. Symp. Intell. Data Anal.*, Berlin, Germany, Sep. 2001, pp. 309–318.
- [34] Y. Liu, "Dimensionality reduction and main component extraction of mass spectrometry cancer data," *Knowl.-Based Syst.*, vol. 26, pp. 207–215, Feb. 2012.
- [35] M. Vetterli and C. Herley, "Wavelets and filter banks: Theory and design," *IEEE Trans. Signal Process.*, vol. 40, no. 9, pp. 2207–2232, Sep. 1992.
- [36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
- [37] J. Gao, S. Giri, E. C. Kara, and M. Bergés, "PLAID: A public dataset of high-resolution electrical appliance measurements for load identification research," in *Proc. 1st ACM Conf. Embedded Syst. Energy-Efficient Buildings (BuildSys)*, 2014, pp. 198–199.



LUYANG GUO was born in Shijiazhuang, Hebei, China, in 1995. He received the B.S. degree from the Hebei University of Science and Technology, in 2018. He is currently pursuing the M.S. degree in electrical engineering with Tianjin University. He works on nonintrusive load monitoring.



SHOUXIANG WANG (Senior Member, IEEE) received the B.S. and M.S. degrees from the Shandong University of Technology, Jinan, China, in 1995 and 1998, respectively, and the Ph.D. degree from Tianjin University, Tianjin, China, in 2001, all in electrical engineering. He is currently a Professor with the School of Electrical and Information Engineering and the Deputy Director of the Key Laboratory of Smart Grid of Ministry of Education, Tianjin University. His main research interests include distributed generation, microgrids, and smart distribution systems.



HAIWEN CHEN was born in Shandong, China. He received the B.S. degree from Shandong University, in 2016. He is currently pursuing the Ph.D. degree with Tianjin University. His research interest includes big data applications in energy systems.



QINGYUAN SHI was born in Zibo, Shandong, China, in 2000. He is currently pursuing the bachelor's degree in electrical engineering with Tianjin University. His research interest includes big data in power systems.

• • •