New Appliance Detection for Nonintrusive Load Monitoring

Jianjun Zhang, Xuanqun Chen, Wing W. Y. Ng, *Senior Member IEEE*,
Chun Sing Lai, *Member IEEE* and Loi Lei Lai, *Fellow IEEE*

Abstract—Current methods for non-intrusive load monitoring problems assume that the number of appliances in the target location is known, however, this may not be realistic. In real-world situations, the initial setup of the site can be known but new appliances may be added by users after a period of time, especially in a household or non-restrictive scenarios. In this sense, current methods without detecting new appliances may not accurately monitor loads of different appliances and scenarios. In this work, a novel new appliance detection method is proposed for non-intrusive load monitoring with imbalance classification for appliances switching on or off. The prediction of an appliances being switched on or off is an important step in load monitoring and the switching on frequencies for coffee machine and air conditioning in a household are different, making the problem inherently imbalanced. Experimental results show that the proposed method yields outstanding performance against the well-known oversampling method, synthetic minority oversampling technique (SMOTE), on real non-intrusive load monitoring applications in scenarios with new appliances emerging.

Index Terms—new appliance detection, imbalance classification, multi-label classification, non-intrusive load monitoring.

I. Introduction

There are two major ways to make energy more sustainable, namely, using renewable energy (such as solar, wind, geothermal, etc.) and improving energy utilization efficiency. The generation of renewable energy is still not very stable and scalable in comparison to traditional energy generation methods, e.g. coal and nuclear power plants. With the fact that 40% of global energy is consumed by residential and commercial buildings, effective management of the energy usage is essential to improve energy efficiency [1]. Load monitoring has great potential in many useful applications, for examples, energy awareness and energy conservation, controllable load quantitative evaluation, human behavior and load prediction [2]. Load monitoring helps to understand the energy consumption of specific appliances in a house and make a more energy efficient plan. If the electricity customers are aware of the average consumption of a type of appliance, more personalized and specific energy saving models of appliances can be recommended to those who are using inefficient devices [3].

There are two major types of load monitoring, that is, intrusive load monitoring (ILM) and non-intrusive load monitoring (NILM) [4]. The ILM refers to installing sensors on each individual appliance or using intelligent sockets to directly monitor their energy use. The key drawback of the ILM is the requirement of a large number of installation and maintenance costs for installed sensors. Therefore, to make load monitoring more practical, the NILM is proposed. Ideally, the NILM only requires data from a smart meter to disaggregate appliance-level data. The NILM is cost-effective and friendly to the new installation and replacement of appliances. The disaggregation problem is usually solved by machine learning methods, e.g. sparse coding [5] and Hidden Markov Model [6].

The NILM problem can be transformed into a multi-label classification problem, such that the ON/OFF state of each appliance is classified simultaneously at each time step [7]. When the NILM problem is treated as a multi-label classification problem, it is inherently a class imbalance problem because some appliances are frequently used (e.g. refrigerators) while others may only occasionally used (e.g. coffee machines). Class imbalance is a common issue in many real-world applications, such as diagnosis of rare diseases, forecasting rare extreme returns in financial markets [8], power system dynamic stability assessment [9], fault diagnosis, and anomaly detection [10]. Class imbalance problems occur when one class severely out-represents another [11], i.e. a class consists of much more samples (i.e., majority class) than other classes (i.e., minority class). Without properly handling the class imbalance issue, classifiers learnt from the skewed dataset will be biased to the majority class and yield a low accuracy on the more important minority class. However, this issue has not been well discussed in current multi-label based learning models for NILM.

Another important issue is that current NILM models are built based on the assumption that the number of appliances during training and testing is fixed. However, in a real-world setting, this assumption seldom holds true because users may add new appliances after a period of time, especially in households and non-restrictive locations. In this sense, existing algorithms may not be able to accurately monitor loads in the

target location. With the introduction of appliances in the loads, it may imply that that the resident has gradually changed the resident behavior due to personal issue or external factors. For example, driven by the price differences, residents may prefer to purchase and use more energy-saving appliances, or users choose to use more electrical appliances during the valley load period than during the peak load period [12]. Better capturing these patterns may help to infer residents' potential interests and more personalized electricity plan and energy saving appliances could be recommended [13].

Therefore, in this work, a new NILM method using the Stochastic Sensitivity Measure-based Noise Filtering and Oversampling method, i.e. SSMNFOS, and the Multi-Label Classification-based New Appliance Detection and Training method, i.e. NADT, are proposed to tackle both the class imbalance issue and the new appliances emerging issue. The SSMNFOS is designed to handle the class imbalance problem. The NADT is designed to detect new appliance being added to the load and to update the classifier ensemble to adapt to the introduction of a new appliance in the multi-label NILM classification problem.

Major contributions of this work are:

1) Based on the knowledge of the authors, this is the first work to investigate the problem of deployment of new appliance during the NILM which is inherently unavoidable in real-world NILM problems. This proposed new research problem to the NILM may lead to a new branch of interesting researches to the load monitoring community.
2) A detection and training algorithm for emergence of new appliance is proposed to help the multi-label classifier ensemble to adapt to new load monitoring environment (i.e. with more appliances than being told).
3) For multi-label classifier ensemble training, the SSMNFOS is used to relieve the class imbalance problems among different appliance classes. The SSMNFOS denoises the training dataset before oversampling to enhance the robustness of oversampling with respect to noisy samples. In contrast to existing oversampling methods, e.g. SMOTE (Synthetic Minority Oversampling Technique), the SSMNFOS is more robust to noisy samples in the minority class. This is important to the NILM problem because the load measured by the smart meter may have noise interference from deficiency of wires, environment, and/or other factors.

Details of all abbreviations are given in TABLE I. The rest of this paper is organized as follows. Related works are introduced in Section II. Section III introduces the proposed method. Experimental setup and results are discussed in Section IV. Section V concludes this work.

TABLE I
List of Abbreviations

ILM	Intrusive Load Monitoring
NILM	Non-intrusive Load Monitoring
MLP	Multilayer Perceptron
SMOTE	Synthetic Minority Oversampling Technique
SSM	Stochastic Sensitivity Measure
SSMNFOS	Stochastic Sensitivity Measure-based Noise Filtering and Oversampling
NADT	New Appliance Detection and Training
MLCDTL	Multi-label Consistent Deep Transform Learning
MLCDDL	Multi-label Consistent Deep Dictionary Learning
MLKNN	Multi-label K-Nearest Neighborhood
KINOS	K-influential Neighborhood Oversampling
REDD	Reference Energy Disaggregation Dataset

II. Related Works

In this section, the NILM, the multi-label classification problem for NILM, and imbalance classification techniques are introduced in subsections A, B, and C, respectively.

A. NILM

Load monitoring refers to the monitoring of the usage of appliance in a house. Appliance-level models are crucial for many smart grid technologies such as demand response, energy storage, and integration of more renewables [3]. The two major categories of load monitoring are the intrusive load monitoring (ILM) and the non-intrusive load monitoring (NILM) [4]. The ILM installs a sensor on each appliance or uses intelligent socket for each appliance. It requires a large number of installation and maintenance costs. To solve this problem, Hart proposed the NILM in 1990s [4]. Ideally, it only requires data from a smart meter to disaggregate appliance-level data. Most researchers of the NILM focus on monitoring switching events on a single appliance while monitoring a set of the same type of appliances may be more meaningful [14]. Although an appliance does not consume much energy, there may be many such type of appliances (such as lights) in the house and users may switch them all ON or OFF together. In addition to smart meter data, appliance usage characteristic is another important information for load monitoring because the usage period (being turned ON) of some appliances may be relatively fixed (e.g. coffee machine in the morning but rarely in mid-night) [13]. Authors of [15] applied deep learning to the NILM while authors of [16] use a semi-supervised learning method to deal with situations of label missing in some training data.

B. Multi-label Classification for NILM

When more than one appliance is being monitored, the specific state of the appliance can be obtained by direct observation without using sensors to measure. The multi-label consistent deep transform learning (MLCDTL) and the multi-label consistent deep dictionary learning (MLCDDL) are applied to learn the multi-label classification for NILM problems by combining transform learning, dictionary learning, and deep learning [15]. The MLKNN is a multi-label classification variant being derived from the general *k*-nearest neighborhood classifier [17]. A more effective way may be training an ensemble of classifiers for the multi-label classification NILM problem with each base classifier trained particularly for each appliance [18]. The multi-label classification is then formed by concatenating all results from the classifier ensemble in a vector form. Another method is the Label Powerset which trains a classifier for each pair of class labels for better distinguishing different labels [19]. However, the problem becomes very complicated and huge when the number of labels (i.e. appliances) is large.

Current NILM researches focus on the classification of the

ON/OFF states of an appliance or a group of appliances. However, all of them assume that the number and types of appliances are previously fixed prior to the NILM. This is unrealistic because people always buy new appliances and plug them to the power network. In this common scenario, current methods will fail because of the unexpected addition of new appliances. Therefore, the detection and adaptation of addition of new appliance is an important new challenge to the NILM researches. Better identification of the introduction of new appliances would improve the overall classification performance of the multi-label learning machine.

C. Imbalance classification

When NILM is considered as a multi-label classification problem, the target is to classify if an appliance is switched ON (1) or OFF (0) at a given time step. Some appliances (e.g. air conditioning) are rarely switched off while some appliances (e.g. coffee machine) are rarely switched on. Therefore, class imbalance is unavoidable in NILM problem. Proper techniques should be employed to improve the robustness and effectiveness of these systems. Resampling is effective in handling the class imbalance problems and is one of the key elements for successful operation of many complex systems such as smart grids [9], [20].

Undersampling removes redundant majority samples while oversampling replicates or generates new minority samples in order to rebalance the class distribution. Undersampling usually refers to undersampling the majority class because minority class consists of fewer samples and removing them may lead to severe information loss. For example, the diversified sensitivity-based undersampling [21] employs clustering to maintain the distribution of both classes and introduces a stochastic sensitivity-based sampling method to select the most informative samples to create a balance dataset. This procedure is executed several times and a robust classifier is iteratively retrained using these rebalance datasets. Recent studies have been carried out to explore the potential of removing minority samples so as to eliminate minority noisy samples [22]. In [22], a *k*-nearest neighbors-based noise filter is applied to remove noises in both classes, after which an undersampling method is applied to rebalance the dataset. In this way, both class noise and class imbalance problems are handled simultaneously.

In contrast to undersampling, oversampling tries to enhance the representation of the minority class by replicating existing minority samples or generates new ones. Among many oversampling methods, the synthetic minority oversampling technique (SMOTE) is of most popularity [23], which generates new samples to increase the number of minority samples along a line connecting adjacent minority samples. Variants of the SMOTE mainly try to overcome the drawback of the SMOTE, i.e., it may introduce new noisy minority samples to the dataset or enlarge the overlapping area between classes. For example, the K-influential neighborhood oversampling (KINOS) [24] first filters the minority samples and then applies an oversampling method to the noise-filtered dataset. After that, the filtered noisy samples are added back to the rebalance dataset to avoid the loss of information.

In this work, the oversampling technique is applied to handle the class imbalance problem to maintain as much useful information as possible. However, the energy data obtained by the meter is often unstable and will fluctuate within a certain range. Thus, the obtained data contains noise, which may affect the classification results. Therefore, a noise filter based oversampling method is applied in this work to better handle the noisy imbalance problem in the non-intrusive load monitoring system.

III. NILM WITH NEW APPLIANCE DETECTION

In standard NILM problems, a classifier or an ensemble of classifier is trained using a given dataset to learn the multi-label classification of the ON/OFF states of a set of given appliances in a house. In real-world situation, new appliances may be added while existing appliances may be removed from the house. In this work, the removal of existing appliance is ignored because they can be classified as OFF without affecting the overall NILM. However, current methods ignoring the emergence of new appliance may jeopardize the NILM task and mislead the following disaggregation process.

The multi-label state classification problem is a steaming problem in which a chunk of data (readings of the smart meter) arrives in every time step. In each time step, the NADT in the proposed method detects deployment of any new appliance and learns the behavior of this new appliance for multi-label state classification. For both existing and new appliances, the multi-label state classification is learned by the SSMNFOS.

So, there are two major components in the proposed method for NILM with class imbalance and emerging appliances. The first one is the sensitivity-based noise filtering and oversampling (SSMNFOS) while the second one is the new appliance detection and training (NADT). They will be introduced in Sections III-A and III-B respectively.

A. Stochastic Sensitivity Measure-based Noise Filtering and Oversampling Method (SSMNFOS)

The classification of noisy samples will change easily if their features change slightly. Noisy samples exist in both NILM and classification problems. The robustness of the trained classifier can be enhanced by identifying and removing those noisy samples. The SSMNFOS is proposed to train classifiers for noisy and imbalance problems by identifying and removing noisy samples before oversampling to improve the classifier training. Stochastic sensitivities of training samples are evaluated using a neural network ensemble. Samples yielding stochastic sensitivity measures (Eq. 3) larger than a pre-selected threshold value are considered as noisy samples and will be removed because they are likely to be misclassified. The threshold value controls how conservative one treats a sample to be noise. The smaller the threshold, the more training samples are regarded as noises. In contrast, a too large threshold leads to very few noisy samples being recognized and the performance of the following oversampling will be hindered.

Then, the SSMNFOS oversamples the noise filtered dataset to balance the class distribution. In this way, new noises are not easily introduced to the dataset, thus the SSMNFOS enhances the performance of the oversampling and leads to a classifier with higher generalization capability. Finally, the SSMNFOS trains a classifier using the final balance denoised dataset. The

pseudocode of the SSMNFOS is given in Algorithm 1. Fig. 1 illustrates the overall procedure of the SSMNFOS.

Fig. 2 demonstrates how the SSMNFOS handles noisy imbalance dataset. A noisy dataset is given in Fig. 2(a). Fig. 2(b) shows the balance dataset created by the standard SMOTE while Fig. 2(c) shows the balance dataset created by the SSMNFOS using the SMOTE as the oversampling method. Both methods create balance datasets, but the standard SMOTE creates a noisier dataset in comparison to the original one. In contrast, the balance dataset created by the proposed method is free from noise in this case. This shows the effectiveness of the proposed method.

The stochastic sensitivity measure (SSM) of a sample is computed by the average output deviations yielded by small perturbations to its input features. If classifier outputs are severely perturbed by these small perturbations in inputs, the classifier is sensitive to this particular training sample. This sample is more likely to be a noisy sample because it is likely to be surrounded by samples in the other class.

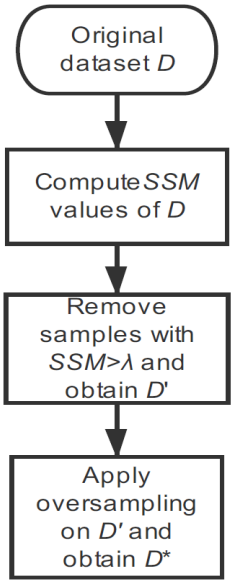


Fig. 1 Overall procedure of the SSMNFOS

In this work, the SSM of a training sample x is defined as the proportion of randomly perturbed samples with predicted labels being different from the true label y of x. The SSM is formulated as follows:

$$SSM(x,h) = \frac{\sum_{p=1}^{\beta} |y - h(x_p)|}{\beta} \quad (1)$$

where x, y, x_p, β, and $h(\cdot) \in \{0,1\}$ denote a given training sample, the true label of x, the p^{th} perturbed samples around x, the number of perturbed samples, and the predicted label by the classifier h, respectively. Perturbed samples are created via adding small perturbations to inputs of the training sample, such that they are located in a region, i.e. Q-neighborhood. The Q-neighborhood of x is defined as follows:

$$S_Q(x) = \{x_p | x_p = x + \Delta x, |\Delta x_i| \leqslant Q, i = 1, 2, \ldots, n\} \quad (2)$$

where Δx, Δx_i, Q, n denote the magnitude of perturbations to the training sample, the magnitude of perturbation to the i^{th} input feature of the training sample, the maximum magnitude of perturbation, and the number of features, respectively. For a dataset normalized to [0, 1], $Q = 0.1$ means that a maximum deviation of 10% from the training sample is allowed for perturbations. Samples located within the Q-neighborhood of a training sample are expected to be in the same class with the training sample. It is because a classifier with a good generalization capability is expected to be robust to such small perturbations.

However, evaluating the *SSM* value of a sample using only one classifier may yield a high variance. Moreover a classifier trained using an imbalance dataset may be biased to the majority class. Therefore, a neural network ensemble trained via a balance bagging method is employed in the SSMNFOS to evaluate the *SSM* values of training samples, which is formulated as follows:

$$SSM(x,H) = \frac{\sum_{t=1}^{||H||} SSM(x,H^{(t)})}{T} \quad (3)$$

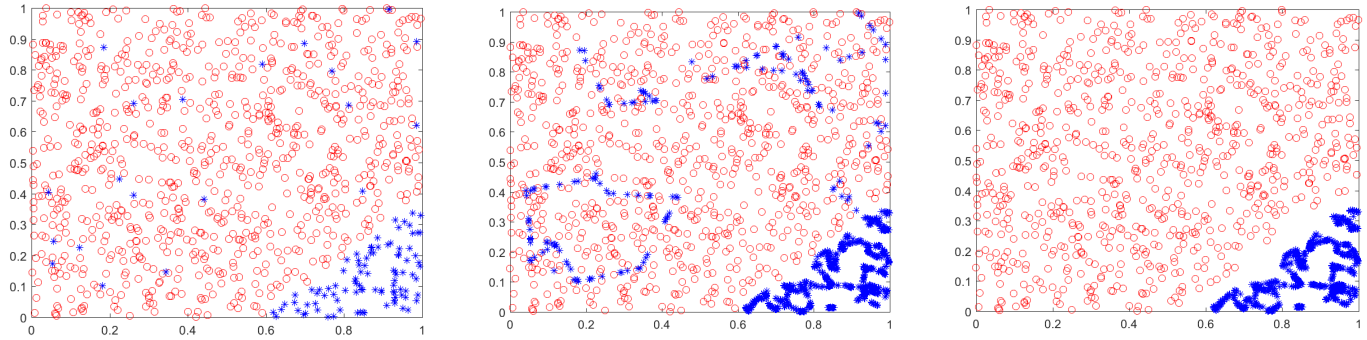
where $H^{(t)}$ and T denote the t^{th} base classifier in ensemble H and the number of base classifiers in H, respectively. The average value of the *SSM* values of each training sample yielded by all base classifiers in H is utilized as the final *SSM* value of each sample for noise evaluation. The balance bagging (Algorithm 2) is employed to create multiple balance sub-datasets to train the neural network ensemble.

Algorithm 1 SSMNFOS
Given: training dataset $\mathcal{D}$, threshold λ, oversampling method
Output: Noise-reduced and balance dataset $\mathcal{D}^*$

1. Train a neural network ensemble H based on $\mathcal{D}$ using balance bagging
2. Compute the average *SSM* value of each training sample through H using Eq. (3)
3. Remove all training samples yielding $SSM(x^{(i)},H) > \lambda$ and set the filtered dataset as $\mathcal{D}'$
4. Apply oversampling on $\mathcal{D}'$ and set the balance dataset as $\mathcal{D}^*$

Algorithm 2 Balance Bagging
Given: training dataset $\mathcal{D}$, number of base classifiers T, learning algorithm L
Output: Ensemble of base classifiers H

1. For $t = 1$ to T do
   a) Set sub-training dataset $U = \varnothing$
   b) Draw $||\mathcal{D}||/2$ samples from each class randomly with replacement and put them in U
   c) Train a base classifier $H^{(t)}$ based on U using learning algorithm L
   End for
2. $H = \arg\max_y \sum_{t:H^{(t)}(x)=y} 1$



(a) Original dataset (b) Balance dataset created by the SMOTE (c) Balance dataset created by the SSMNFOS

Fig. 2 Balance dataset created by the SMOTE and the SSMNFOS

B. New Appliance Detection and Training (NADT)

The assumption that all appliances are known at the beginning of NILM is not reasonable and unrealistic. In practice, however, new appliances may be added to the house and the number of appliances may change. Therefore, an appliance detection and training algorithm is proposed to improve the multi-label classification for NILM. In multi-label NILM problems, an ensemble of neural networks is trained with each base classifier corresponding to the ON/OFF state classification for an appliance. Each base classifier is trained using the SSMNFOS for a given appliance.

To formulate the learning problem, the data collected by the smart meter is assumed to be consecutively sent to the learning system in batches with a fixed size (for example, half an hour of energy readings). In this setting, the learning model is firstly trained using the first several data chunks and is evaluated using the later upcoming data chunks. So the learning model is evaluated over each data chunk.

If no new appliances emerge in the consecutive data chunks, the performance of the model is expected to maintain at a stable level. However, if new appliances are added to the loads, the performance of the model would be negatively affected because it can only recognize appliances learnt from the previous training data chunks. Therefore, the new appliance detection and training method (NADT) in this work detects new appliances when the overall performance of the ensemble of neural networks drops.

Algorithm 3 NADT

Given: preprocessed dataset $\mathcal{D}$, batch number of dataset M, initial training batch number M', initial appliance number R, threshold λ

Output: Classifier sets for initial appliances and new appliances H'

1. Uses M' batch data of D to train the initial R appliances and get R classifiers.
2. The average metric value of these R classifiers on the M' batch data is called the baseline B, which represents the performance of the classifiers when it is working normally (no new appliances emerge).
3. Add R initial classifiers to H'.
4. For $t = M'+1$ to M do
   a) Use R classifiers to test the t^{th} data chunk and get a metric value E of the current data chunk.
   b) If $|E-B|>\lambda$, considers that the new appliance emerges, go to 4(c), else continue.
   c) If new appliance is used on t^{th} data chunk, the algorithm is judged correctly, go to 4(d), else continue.
   d) The number of new appliances is called W. Use $(t-1)^{th}$ and t^{th} data chunks to train the classifiers of new appliances. If the size of new appliance 's minority class on these samples is less than 2, add the previous chunk. (Such that both positive and negative samples are available and there are at least two samples in each class).
   e) Add W new classifiers to H'.
   f) The average metric value of these $R+W$ classifiers on the t^{th} and $(t-1)^{th}$ data chunks is used to update the baseline B.
   g) $R=R+W$.

End for

For example, at the beginning of the experiment, there are R appliances in the house. The proposed method first trains R base classifiers for the R appliances. The performance evaluated on these training data is recorded and used as the baseline. Then, the trained model is used to evaluate on each incoming data chunk and the performance is recorded. If the performance on current data chunk differs from the baseline by more than a given threshold, a new appliance is considered to emerge in the current data chunk. This data chunk is used to train a new base classifier to monitor the emerging appliance. Detailed procedures are shown in Algorithm 3.

IV. Experimental Studies

In this section, we divide the experiment into two parts. Part A confirms the advantages of the proposed method SSMNFOS in dealing with noisy imbalance data, and part B confirms that the proposed NADT algorithm can be well combined with SSMNFOS to deal with the problem of new appliance detection in NILM. In parts A and B, the basic classifier of all the methods used is a well-known multilayer perceptron (MLP). The implementation of MLP in WEKA [25] is employed in this work and the default parameters are used.

A. Imbalance Classification with Noise

In order to evaluate the superiority, robustness and effectiveness of the SSMNFOS, experiments on imbalance classification with label noises are designed. An electricity price dataset (Australian Electricity Price Data) collected from the Australian National Electricity Market [26], a stability of electrical grid dataset, and another eight widely used imbalance datasets from the KEEL dataset repository [27] and the UCI dataset repository [28] are employed. Although some of datasets employed in this experiment are not related to electric load monitoring, without loss of generality, they can illustrate the purpose of proposing the

method. Both the Electrical Grid Stability Simulated Data and the Australian electricity price data are datasets related to smart grid. The collected Australian electricity dataset consists of electricity price data in New South Wales (NSW) and Victoria (VIC) in year 2018. The task is to determine whether the average spot price ($/MWh) in NSW is higher than that in VIC in every 30 minutes. The features used in the classification tasks are day in a week, period in a day, demand of NSW and VIC. Characteristics of datasets are given in TABLE II, where imbalance ratio (IR) is defined as the number of majority samples divided by the number of minority samples. Missing data in each dataset has been removed and the number of samples shown in TABLE II is the number of samples of the processed dataset. Each dataset is randomly split into two halves, one for training and the rest for testing. A ten-time five-fold cross validation is employed for each dataset. The mean and standard deviation values of the performance for each method over ten runs are recorded for performance evaluation.

To show the differences among different methods, the Friedman's test with a post-hoc Hochberg's test [29] at the significance level of 0.05 is applied to compare the proposed method with other methods over multiple datasets. The compared methods used in this experiment are the SMOTE and the basic MLP (with no treatment to imbalance problem). The performance metric adopted for this experiment is a new metric F derived from the $F1$ measure. The $F1$ measure provides a comprehensive consideration of precision and recall for the positive class. It is defined as follows:

$$F1 = \frac{2TP}{2TP+FP+FN} \tag{4}$$

where TP, FP, and FN denote true positive, false positive, and false negative, respectively. For imbalance problems, $F1$ is often biased to the positive class and ignores the performance of the negative class. In order to evaluate the overall performance of the classifier, F is used as the performance metric of the classifier instead which is defined as follows:

$$F1' = \frac{2TN}{2TN+FN+FP} \tag{5}$$

$$F = \frac{P}{N+P}F1 + \frac{N}{N+P}F1' \tag{6}$$

where TN, P, and N denote true negative, the number of positive samples, and the number of negative samples, respectively.

To analyze how different methods handle noisy imbalance datasets, noises are manually introduced in the training datasets because most of the datasets we used in this work may not actually contain noises. We adopt a pair-wise noise introduction scheme as follows: given a pair of classes($y1$,$y2$) and a noise level ρ, an instance with label $y1$ has a probability of ρ to be incorrectly labeled as $y2$, so does an instance with label $y2$. This mechanism was proposed by Zhu et al. [30], claiming that in realistic situations, only certain types of classes are likely to be mislabeled. Five levels of noises are introduced in the datasets, namely, 0.05, 0.1, 0.2, 0.3, and 0.4.

TABLE II
Characteristics of Datasets

Name	#Features	#Samples	IR
Australian Electricity Price Data [26]	4	17520	1.66
Electrical Grid Stability Simulated Data [28]	13	10000	1.76
Iris0 [27]	4	150	2
New-thyroid2 [27]	5	215	5.14
Pima [27]	8	768	1.87
Wisconsin [27]	9	683	1.86
Breast Cancer Wisconsin (Original) [28]	10	683	1.86
Chronic_Kidney_Disease [28]	25	158	2.67
Credit Approval [28]	15	653	1.21
Z-Alizadeh Sani [28]	56	303	2.48

Results of different methods on dataset Australian Electricity Price Data with different levels of noises are shown in Fig. 3, the rest are omitted due to space limitation and they all produce similar results.

We can draw some conclusions from it. Increasing the noise level tends to deteriorate the performances of each method. This is mainly because adding noises to the training data increases the learning complexity and skewness of the data. The proposed SSMNFOS outperforms other methods under different noise levels. It shows that SSMNFOS can filter out noisy samples and improve the performance of oversampling. The performance of SMOTE is ultimately lower than that of MLP which does not deal with imbalance problems, this is because directly oversampling the minority classes without properly handling noises introduces more noisy samples to the dataset and further increases the learning complexity. The standard deviation of SSMNFOS is smaller than that of other methods, which indicates that SSMNFOS is more stable than other methods.

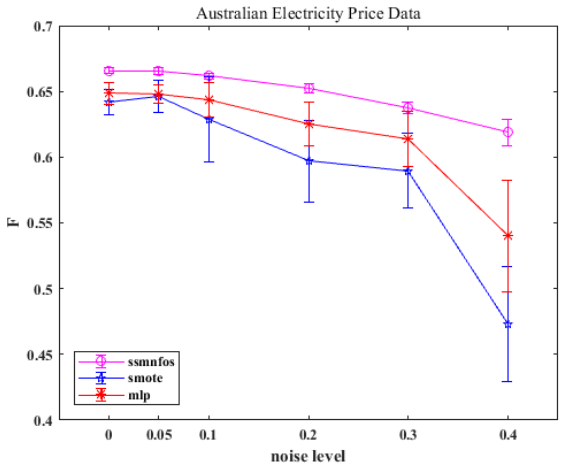


Fig. 3 Results of different methods on dataset Australian Electricity Price Data with different levels of noises. The error bar on each curve shows the mean value over 10 runs ± one standard deviation.

TABLE III
Results of Friedman's Test.

Noise level	0%		5%		10%	
Methods	rank	**p-H**	rank	**p-H**	rank	**p-H**
SSMNFOS	1.4	N/A	1.1	N/A	1.2	N/A
SMOTE	2.4	2.53E-02	2.8	1.44E-04	2.6	1.75E-03

MLP	2.2	**7.36E-02**	2.1	2.53E-02	2.2	2.53E-02
p-F	**6.08E-02**		6.76E-04		5.52E-03	
Noise level	20%		30%		40%	
Methods	rank	**p-H**	rank	**p-H**	rank	**p-H**
SSMNFOS	1	N/A	1	N/A	1	N/A
SMOTE	2.7	1.44E-04	2.7	1.44E-04	2.5	7.96E-04
MLP	2.3	3.65E-03	2.3	3.65E-03	2.5	7.96E-04
p-F	3.71E-04		3.71E-04		5.53E-04	

Results of the Friedman's test with the post-hoc Hochberg's test are given in TABLE III. The p-F and p-H mean the p-value computed by the Friedman's test and the adjusted p-value computed by the post hoc Hochberg's test respectively. From TABLE III the SSMNFOS yields the lowest average ranks in almost all cases. Almost all p-F values are much less than 0.05, which indicate that there are significant differences among the three methods. Using SSMNFOS as the control method, the p-H of SMOTE and MLP is less than 0.05 in almost all cases, which shows that the proposed method significantly outperforms the SMOTE and MLP.

Based on the above observations, it can be safely concluded that the performance of SSMNFOS on imbalance classification is significantly better than that of the compared methods under different noise levels.

B. Detection of New Appliance

The Reference Energy Disaggregation Dataset (REDD) [31] is used to evaluate the performance of the proposed new appliance detection and imbalance multi-label classification algorithm. The low frequency data in the REDD dataset is used in this work which contains both the total power and the appliance-level power data of six houses at the frequency of 1 Hz. The load monitoring analysis method is based on the load steady-state analysis method and all the features are derived from power data.

House 1 is used in our experiments because it has more appliances and data. Among them, eleven appliances in House 1, i.e. numbers 3, 4, 6, 9, 10, 12, 13, 14, 16, 17, and 20 in the dataset are used. As the NILM problem is regarded as a multi-label classification problem, the appliance-level power values are not actually needed but only states of all appliances. When the appliance-level power of an appliance is zero at a time step, it is considered as OFF state. The appliance is at ON state otherwise. In order to simulate the experimental environment in which the new appliance appears, the power of the total meter is the sum of the power of the current appliances in the house.

In the experiments, data arrives batch by batch with each batch containing half an hour of power readings. Then, each batch of data is divided into 180 10-second non-overlapping data windows. Five load identification features are extracted from multiple power data of each window, which are average, variance, minimum, maximum and median. Each window is labeled by states of all appliances in the last second of the window in a multi-label vector form. For training and testing division, the protocol in machine learning for streaming data is followed. The newly arrived data chunk serves as the testing samples for the NILM problem. After testing, this data chunk becomes a set of training samples.

When NILM is considered as a multi-label classification problem, there is no obvious bias between positive and negative classes. Therefore, the *F* measure and Accuracy are used as metrics. Owing to the fact that there are multiple classes while both F and $F1$ are designed for two-class problems only, F_{macro} and F_{micro} metrics are derived from F for multi-class problems [32]. The F_{macro} is calculated from the average F of all classes while the F_{micro} sums up metrics of all classes before computing the final metric. They are two ways to compute the average scores, and later experiments show that their difference is significant.

$$F_{macro} = \frac{1}{l}\sum_{i=1}^{l} F(TP_i, TN_i, FP_i, FN_i) \quad (7)$$

$$F_{micro} = F(\sum_{i=1}^{l} TP_i, \sum_{i=1}^{l} TN_i, \sum_{i=1}^{l} FP_i, \sum_{i=1}^{l} FN_i) \quad (8)$$

where TP_i, TN_i, FP_i, and FN_i denote TP, TN, FP, and FN of the i^{th} class, $F()$ and l denote the equation computing F in Eq. (6) and the number classes, respectively. On the other hand, *Accuracy* provides an overall performance evaluation on the classifier. It is defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

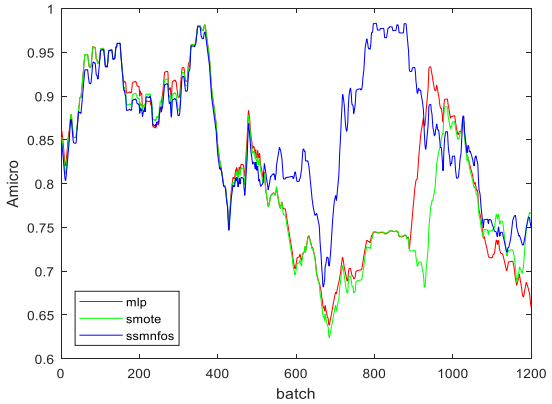
Similarly, in order to evaluate the performance of multi-label NILM problems, A_{macro} and A_{micro} are given as follows:

$$A_{macro} = \frac{1}{l}\sum_{i=1}^{l} A(TP_i, TN_i, FP_i, FN_i) \quad (10)$$

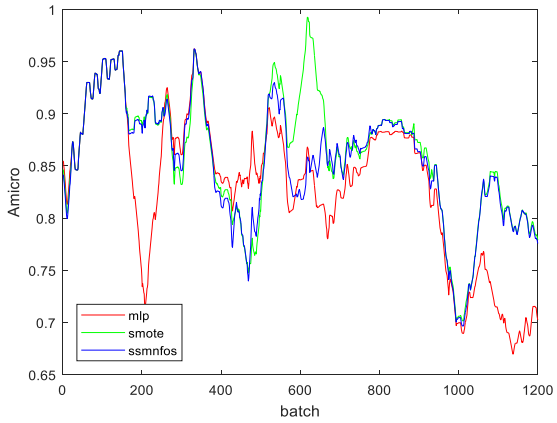
$$A_{micro} = A(\sum_{i=1}^{l} TP_i, \sum_{i=1}^{l} TN_i, \sum_{i=1}^{l} FP_i, \sum_{i=1}^{l} FN_i) \quad (11)$$

where TN_i and $A()$ denote TN of the i^{th} class and equation computing *Accuracy* as given in Eq. (9) respectively.

The ON/OFF states of an appliance can be highly imbalance, so the SSMNFOS is compared with both the widely used SMOTE and classifier trained without treatment to the imbalance issue. Multilayer Perceptron (MLP) is used as base classifier in our experiments.



(a) Scenario 1 (New appliances are added with order: 3, 4, 9, 12, 13, 14, and 20)



(b) Scenario 2 (New appliances are added with order: 3, 12, 9, 4, 14, 13, and 20)

Fig. 4 Dynamic performance of different methods in the two scenarios in terms of A_{micro}

There are two scenarios in our experiments and the data for the first three days with four appliances (6, 10, 16, and 17) in the house is used as the initial training data. Then, new appliances are added in different orders in the two scenarios with randomly selected time intervals for adding a new appliance. In Scenario 1, starting from the fourth day (testing phase), new appliances are added to the house in the following order: 3, 4, 9, 12, 13, 14, and 20. In Scenario 2, the order of adding appliances is: 3, 12, 9, 4, 14, 13, and 20. In the experiments, the new appliance detection performance of different methods is first tested and then their multi-label classification performances for NILM problems in the two scenarios are tested. For the NADT, λ is set to 0.3 for all experiments which is selected via some preliminary experiments.

In order to illustrate the dynamic performance of each method in a streaming setting, the performance value of each method in the first 1250 batches of data was recorded. Due to the limit of space, only the dynamic performance of different methods in both scenarios in terms of A_{micro} is shown in Fig. 4 in which, red line named as "MLP", green line named as "SMOTE", and blue line named as "SSMNFOS" represent the performance of off-the-shelf MLP without handling the class imbalance issue, the performance of the MLP equipped with the SMOTE, and the performance of the MLP equipped with the SSMNFOS, respectively. In order to observe the trend of Amicro, we use moving average with a window size of 50. As shown in Fig. 4, the performance of all three methods fluctuates severely as time varies. The "MLP" and the "SMOTE" show similar patterns but the "SSMNFOS" outperforms the other two in most of the time. One possible reason for the performance fluctuation is that only five simple time domain features are used to distinguish different classes, which may not be sufficient for training a very strong classifier. The major focus in this experiment is the fair comparison of different methods using the same classifiers and input features. More sophisticated input features can be employed to enhance the robustness of the classifiers, which serves as one of our future works. To get a better understanding of the performance differences among the three methods, more detailed numerical results are given in TABLEs IV to VII.

TABLEs IV and V show the new appliance detection results using the NADT in the two scenarios. There are totally 7 new appliances in each scenario. The NADT uses the same metric as the final evaluation criterion to detect the addition of new appliance. Therefore, the metric has a significant effect to the NADT.

Base classifiers are trained with different methods to deal with the imbalance issue in the NILM, namely, widely used SMOTE and the SSMNFOS in this work. TABLEs IV and V show new appliance detection results for Scenarios 1 and 2, respectively. In Scenario 1, the SSMNFOS and the SMOTE detect all 7 new appliances in 3 out of 4 cases, the MLP without treatment to imbalance problems can only detect all 7 new appliances in 1 out of 4 cases. In Scenario 2, the SSMNFOS and the SMOTE detect all 7 new appliances in 4 out of 4 cases, the MLP without treatment to imbalance problems can't detect all 7 new appliances in any case. The results show that the imbalance classification algorithm can enhance the performance of NADT in detecting new appliances.

TABLE IV
New Appliance Detection using the NADT with Different Treatment to Imbalance Problems for Scenario 1

Methods to Deal with Imbalance Issue	F_{macro}	F_{micro}	A_{macro}	A_{micro}
SSMNFOS	4	**7**	**7**	**7**
SMOTE	**7**	**7**	6	**7**
MLP	4	4	**7**	6

TABLE V
New Appliance Detection using the NADT with Different Treatment to Imbalance Problems for Scenario 2

Methods to Deal with Imbalance Issue	F_{macro}	F_{micro}	A_{macro}	A_{micro}
SSMNFOS	**7**	**7**	**7**	**7**
SMOTE	**7**	**7**	**7**	**7**
MLP	4	4	4	4

TABLE VI
Performance Evaluation using the NADT with Different Treatment to Imbalance Problems for Scenario 1

Methods to Deal with Imbalance Issue	F_{macro}	F_{micro}	A_{macro}	A_{micro}
SSMNFOS	**72.74**	**78.47**	69.69	**81.22**
SMOTE	68.43	72.12	67.93	79.98
MLP	72.06	74.56	**78.42**	79.91

TABLE VII
Performance Evaluation using the NADT with Different Treatment to Imbalance Problems for Scenario 2

Methods to Deal with Imbalance Issue	F_{macro}	F_{micro}	A_{macro}	A_{micro}
SSMNFOS	80.99	**85.09**	**85.87**	**85.56**
SMOTE	**82.07**	81.74	85.10	85.10
MLP	80.48	80.88	80.62	80.59

TABLEs VI and VII show average testing performance of the NADT combined with different treatments to imbalance issues in the NILM over all data chunks in the test phase for Scenarios 1 and 2, respectively. One can observe that the MLP method yields the best performance in terms of A_{macro} in Scenario 1, and the SMOTE performs the best in terms of F_{macro} in Scenario 2. However, the proposed method by combining the NADT with the SSMNFOS yields the best performance in 6 out of 8 cases (4 metrics and 2 scenarios), which may draw a safe conclusion that the proposed method performs better than the other two methods overall.

In summary, the four metrics provide different angles to evaluate the performance of the multi-label learning by the ensemble of neural networks for NILM. The proposed method yields the best performance in most of cases and is effective for training ensemble of neural networks for handling the NILM problems.

V. Conclusion and Future Works

Current methods for non-intrusive load monitoring (NILM) problems assume that the number of appliances in the target location is known which may be unrealistic. In the real-world, the initial settings of the site can be known, but new appliances can be added by the user after a period of time, especially in a household or unrestricted scenario. In this situation, current methods that do not detect new appliances may not accurately monitor the load on different appliances and scenarios. Therefore, a new appliance detection and a training algorithm new appliance detection and training (NADT) are proposed for multi-label classification in NILM. Then the stochastic sensitivity measure-based noise filtering and oversampling (SSMNFOS) is applied to train base classifier for an appliance to form the multi-class ensemble of neural network for the multi-label classification for the NILM. Experimental results show that the SSMNFOS yields a better performance than the widely used SMOTE for dealing with imbalance problems in the NILM.

The major contribution of this work is to propose a new research problem in the NILM. The new appliance detection method proposed in this work is primitive and more sophisticated detection methods need to be investigated to further improve the new appliance detection for NILM problems.

The proposed method focuses on the multi-label classification of ON/OFF states of appliances and the detection of new appliances. Off-the-shelf disaggregation methods may be combined with the proposed method easily to predict and monitor loads of different appliances across time. However, to fully utilize the new appliance detection, a dedicated disaggregation method may be needed. For example in a hot summer, more air conditioning can be expected to be switched on. But a demand response program may encourage customers to save energy which on the contrary leads to lower switching on frequency of appliances. The disaggregation method needs to take into account the effects of the demand response programs. This will be an important future work.

A practical application of this work is to combine the proposed method with specific knowledge and load patterns of an appliance to predict the failure or performance degradation of an appliance. Based on the classification of ON/OFF state and predicted load patterns, the depreciation model of an appliance can be found and user experience can be enhanced by more precise customer services.

References

[1] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," *Energy Buildings*, vol. 40, no. 3, pp. 394–398, 2008.

[2] W. Kong, Z. Y. Dong, D. J. Hill, J. Ma, J. H. Zhao and F. J. Luo, "A hierarchical hidden Markov model framework for home appliance modeling," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3079-3090, July 2018.

[3] W. Kong, Z. Y. Dong, J. Ma, D. J. Hill, J. Zhao and F. Luo, "An extensible approach for non-intrusive load disaggregation with smart meter data," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3362-3372, July 2018.

[4] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870-1891, 1992.

[5] E. Elhamifar and S. Sastry, "Energy disaggregation via learning powerlets and sparse coding," *AAAI*, pp. 629–635, 2015.

[6] S. Makonin, F. Popowich, I. V. Bajić, B. Gill and L. Bartram, "Exploiting hmm sparsity to perform online real-time nonintrusive load monitoring," *IEEE Transactions on Smart Grid*, vol. 7, no. 6, pp. 2575-2585, Nov. 2016.

[7] S. M. Tabatabaei, S. Dick and W. Xu, "Toward non-intrusive load monitoring via multi-label classification," *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 26-40, 2017.

[8] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalance domains," *ACM Computing Surveys*, vol. 49, no. 2, article 31, 2016.

[9] L. Zhu, C. Lu, Z. Y. Dong and C. Hong, "Imbalance learning machine-based power system short-term voltage stability assessment," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2533-2543, Oct. 2017.

[10] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, vol. 42, no. 4, pp. 463-484, 2012.

[11] C. S. Lai, et al, "A robust correlation analysis framework for imbalanced and dichotomous data with uncertainty,"

Information Sciences, vol. 470, pp. 58-77, 2019.

[12] F. Luo, G. Ranzi, W. Kong, Z. Y. Dong, S. Wang, and J. Zhao, "Non-intrusive energy saving appliance recommender system for smart grid residential users," *IET Generation Transmission and Distribution*, vol. 11, no. 7, pp. 1786-1793, 2017.

[13] M. Ma, W. Lin, J. Zhang, P. Wang, Y. Zhou and X. Liang, "Toward energy-awareness smart building: discover the fingerprint of your electrical appliances," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1458-1468, April 2018.

[14] Y. Liu, G. Geng, S. Gao and W. Xu, "Non-intrusive energy use monitoring for a group of electrical appliances," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3801-3810, July 2018.

[15] V. Singhal, J. Maggu and A. Majumdar, "Simultaneous detection of multiple appliances from smart-meter measurements via multi-label consistent deep dictionary learning and deep transform learning," *IEEE Transactions on Smart Grid*. DOI. 10.1109/TSG.2018.2815763.

[16] J. M. Gillis and W. G. Morsi, "Non- intrusive load monitoring using semi-supervised machine learning and wavelet design," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2648-2655, Nov. 2017.

[17] M.-L. Zhang and Z.-H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification," *IEEE International Conference on Granular Computing,* pp. 718-721, 2005.

[18] M. Gulati, S. S. Ram, A. Majumdar, and A. Singh, "Single point conducted emi sensor with intelligent inference for detecting it appliances," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, 2018.

[19] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification," *European Conference on Machine Learning*, pp. 406–417, 2007.

[20] W. W. Y. Ng, J. Zhang, C. S. Lai, W. Pedrycz, L. L. Lai, and X. Wang, "Cost-sensitive weighting and imbalance-reversed bagging for streaming imbalance and concept drifting in electricity pricing classification," in *IEEE Transactions on Industrial Informatics.* vol. 15, no. 3, pp. 1588-1597, March 2019.

[21] W. W. Y. Ng, J. Hu, D. S. Yeung, S. Yin, and F. Roli, "Diversified sensitivity-based undersampling for imbalance classification problems," *IEEE Transactions on Cybernetics*, vol. 45, no. 11, pp. 2402 – 2412, 2015.

[22] Q. Kang, X.-S. Chen, S.-S. Li, and M.-C. Zhou, "A noise-filtered under-sampling scheme for imbalance classification," *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4263-4274, 2017.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321-357, 2002.

[24] R. F. A. B. de Morais, and G. C. Vasconcelos, "Under-sampling the minority class to improve the performance of over-sampling algorithms in imbalance data sets," *Proceedings of International Joint Conference on Artificial Intelligence*, 2017.

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," in *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10-18, 2009.

[26] Australia Energy Market Operator, [Online]: http://www.aemo.com.au/Electricity/National-Electricity-Market-NEM/Data-dashboard#aggregated-data (Accessed 25/03/2019).

[27] I. Triguero, S. González, J. M. Moyano, S. Garcaí, J. Alcalá -Fdez, J. Luengo, A. Fernández, M. J. del Jesus, L. Sánchez, and F. Herrera, "KEEL 3.0: an open source software for multi-stage analysis in data mining," *International Journal of Computational Intelligence Systems*, vol. 10, pp. 1238-1249, 2017.

[28] D. Dua and C. Graff, "UCI Machine Learning Repository," Irvine, CA: University of California, School of Information and Computer Science, 2019. [Online]. Available: http://archive.ics.uci.edu/ml.

[29] S. Garcia, A. Fernandez, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining- experimental analysis of power," *Information Sciences*, vol. 180, pp. 2044-2064, 2010.

[30] X. Zhu, X. Wu, and Q. Chen, "Eliminating class noise in large datasets," in *Proceedings of the Twentieth International Conference on Machine Learning*, Washington DC, pp. 920-927, 2003.

[31] J. Z. Kolter and M. J. Johnson, "REDD: A public data set for energy disaggregation research, " *Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA*, vol. 25, pp. 59–62, 2011.

[32] K. Basu, V. Debusschere, S. Bacha, U. Maulik and S. Bondyopadhyay, "Nonintrusive load monitoring: a temporal multilabel classification approach," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 1, pp. 262-270, 2015.