



A Feature and Classifier Study for Appliance Event Classification

Benjamin Völker^(✉), Philipp M. Scholl, and Bernd Becker

Computer Architecture, University of Freiburg, Freiburg im Breisgau, Germany
{voelkerb,pscholl,becker}@informatik.uni-freiburg.de

Abstract. The shift towards advanced electricity metering infrastructure gained traction because of several smart meter roll-outs during the last decade. This increased the interest in Non-Intrusive Load Monitoring. Nevertheless, adoption is low, not least because the algorithms cannot simply be integrated into the existing smart meters due to the resource constraints of the embedded systems. We evaluated 27 features and four classifiers regarding their suitability for event-based NILM in a standalone and combined feature analysis. Active power was found to be the best scalar and WaveForm Approximation the best multidimensional feature. We propose the feature set $[P, \cos\Phi, TRI, WFA]$ in combination with a Random Forest classifier. Together, these lead to F_1 -scores of up to 0.98 on average across four publicly available datasets. Still, feature extraction and classification remains computationally lightweight and allows processing on resource constrained embedded systems.

Keywords: NILM · Non-Intrusive Load Monitoring · Feature evaluation · Appliance classification

1 Introduction

Reducing our electricity consumption is a vital step to achieve the goal of saving earth's energy resources. In the residential or industrial domain energy monitoring and eco-feedback help by raising the awareness of an unnecessary electricity consumption of particular devices. To pinpoint user to specific appliances that consume too much energy, appliance specific consumption data are required. These can be retrospectively provided by utilizing existing smart meter infrastructure with Non-Intrusive Load Monitoring (NILM). NILM methods disaggregate the composite load into the load of each electrical consumer by incorporating machine learning approaches. These approaches can be classified into event-based and event-less methods. The latter apply disaggregation for each new data entry, while event-based approaches apply disaggregation whenever a new appliance event was recognized in the aggregated load. After identifying events, a classifier is typically used to determine to what appliances these events belong to. The generated list of events is finally used to reconstruct the load profile of the appliance, e.g. by grouping *switch-on* and *switch-off* events and

assigning a known average consumption to times an appliance was switched on. The steps to detect and classify an appliance event include the extraction and pre-processing of the event, feature extraction, and finally classification. Classification algorithms working with a large number of features may achieve high classification performances (F_1 -score > 0.9 such as proposed in [14]). However, the deployment on smart meters is hindered by the amount of features due to the required computational resources to calculate them and a typically linear increase in complexity for most classifiers.

The contributions of this work mainly include: (1) An evaluation of 27 features and four classifiers regarding their suitability for the task of appliance classification. (2) An evaluation of several combinations of these features with the goal to find a trade-off between feature dimensionality and classification performance. (3) The proposition of the feature set $[P, \cos\Phi, TRI, WFA]$ with a Random Forest classifier for the task of appliance classification on resource constrained systems.

The remainder of this paper is organized as follows: Sect. 2 provides an overview of the NILM pipeline and lists state-of-the-art features and classifiers for event-based NILM as well as existing datasets for NILM. Section 3 introduces the event detector used to generate the training data, the features and classifiers as well as the evaluation strategy. Section 4 presents the results of the standalone feature analysis. The feature selection scheme for the combined analysis is explained in Sect. 5 and the results of the analysis are presented and discussed in Sect. 6. Finally, concluding remarks are provided in Sect. 7.

2 Related Work

The general NILM process can be divided into the two steps (1) *Data Acquisition* and (2) *Disaggregation* as shown in Fig. 1. *Data Acquisition* is comprised of measuring the required attributes (such as active- and reactive power) and performing general pre-processing steps while the *Disaggregation* step is a specially designed and often individually trained algorithm. Most of the disaggregation algorithms that have been proposed by researchers can be categorized into *event-based* and *event-less* approaches (e.g. according to [27, 33]).

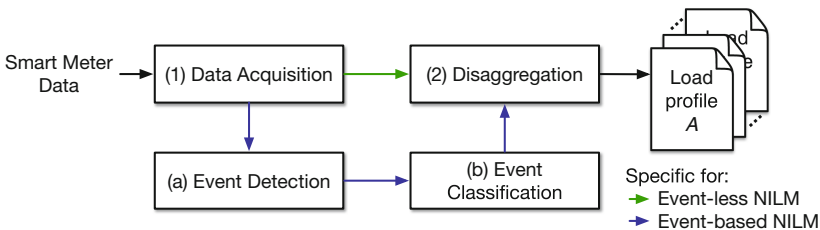


Fig. 1. General pipeline of event-less and event-based NILM systems

Event-less approaches optimize an overall system state using individually trained appliance models. These models are typically based on Hidden Markov Model (HMM) [21, 25] or Artificial Neural Network (ANN) [6, 19]. As the optimization step is recalculated for each new data input, event-less approaches typically suffer from high computational complexity and can, therefore, only be applied to lower sampling rates (<1 Hz). According to Anderson et al. [2] the *event-based* NILM process introduces two additional sub-steps as depicted in Fig. 1: (a) *Event Detection* and (b) *Event Classification*. *Event Detection* relies on the concept of the Switch Continuity Principle (SCP). The SCP was introduced by Hart [10] in 1992 and states that at a specific point in time only a single event, i.e. appliance state change, can occur and that overall, the number of events is small. This allows to treat events as signal anomalies, which need to be detected during event detection. *Event Classification* (also called appliance classification) follows the pattern recognition paradigm. Features, which are typically handcrafted by domain experts, are extracted from each event and are fed into a classifier, which outputs more details about the type of event (e.g. a specific appliance turning on). As the classification step is only applied to events, which are typically rare, event-based NILM systems are computational less expensive compared to *event-less* approaches, which perform the inference step for each new sample. The *Disaggregation* step uses the generated list of appliance events to extract estimated load profiles for each appliance (e.g. by recognizing the appliance’s state transitions such as from *on* to *off* and mapping a known average consumption to each state).

2.1 Event Classification

Over the years, several hand-crafted features, for the task of event classification, have been introduced by various researchers. The most frequently used features are surveyed e.g. by Liang et al. [24]. Kahl et al. [17] evaluated 36 features in a stand-alone feature analysis as well as their combination using a feature forward selection technique. The authors found that across all used datasets, the phase angle difference between voltage and current ($\cos\Phi$) was the best scalar feature ($F_1 = 0.49$) while Current Over Time (COT) achieved the best multi-dimensional feature performance ($F_1 = 0.8$). Different classification algorithms have been evaluated for the task of appliance classification such as Random Forests (RF) [4, 8, 26] Support Vector Machines (SVM) [16], k -Nearest Neighbour (k NN) [8, 16, 32] and more recently Artificial Neural Networks (ANN) [4, 5, 14]. Hubert et al. [13] and Kahl et al. [17] surveyed several algorithms for appliance classification. Hubert et al. [13] focused on Deep Neural Networks (DNNs) and identified higher sampling rates, the use of larger receptive fields, and an ensemble of input features, amongst others, as promising techniques to improve the performance of such networks. Kahl et al. [17] directed their focus on standard machine learning algorithms and identified that k NN performs quite well for the

task of appliance classification despite its comparable low computational complexity. It is further noted that the training of ANNs constitutes a large burden for resource constrained embedded systems such as smart meters. Depending on the system’s restrictions, a computationally lightweight algorithm such as k NN may be better suited.

2.2 Datasets

To achieve comparable results, experiments are typically carried out using pre-recorded datasets. In the domain of event-based NILM, several high-frequency datasets exist such as WHITED [15], PLAID [7], REDD [20], BLUED [1], UKDALE [18], BLOND [22], and FIRED [28]. They mainly differ in the used Data Acquisition System (DAQ). The data sampling frequencies range from 8 kHz for FIRED up to 250 kHz for BLOND-250. While WHITED and PLAID include isolated appliance events recorded in a laboratory setup, the remaining datasets include aggregated data of real world deployments.

3 Background

This section details the event detection algorithm, the extracted features, and the basic classifiers used throughout this work.

3.1 Event Detection

Event detection, often referred to as edge detection, describes the process of identifying relevant changes in a signal. We use an event definition for electrical power signals, which has been proposed by Wild et al.: “*An event is a transition from one steady state to another steady state, which definitely differs from the previous one [...] [or] a so-called active section where the signal is somehow deviating from the previous steady state*” [31]. As appliance event detection is a research field on its own (see e.g. [2,27,31]) and a deeper evaluation would go beyond the scope of this paper, we choose a relatively simple expert heuristic event detector based on work by Weiss et al. [30]. It uses a threshold-based setup, which is applied on the apparent power signal (S). At first, the signal is filtered using the combination of a median filter to remove outliers and a mean filter to further smooth the signal. Both filters have a width of 3 s. Afterwards, the absolute difference between adjacent samples of the apparent power signal is calculated (ΔS). Next, a 3 VA filter is applied to the signal, which sets all values below 3 VA to zero as

$$\Delta S_{filtered}(i) = \begin{cases} 0 & \text{if } \Delta S(i) < 3 \text{ VA,} \\ \Delta S(i) & \text{else.} \end{cases} \quad (1)$$

Each non-zero portion in the filtered signal is regarded as an event (*up* or *down*). If multiple events happen within a time window of 3 s, we only keep the first one. This ensures that fluctuations after an event are not regarded as a new event. Figure 2 shows the different stages of the event detection process for the apparent power signal of an espresso machine.

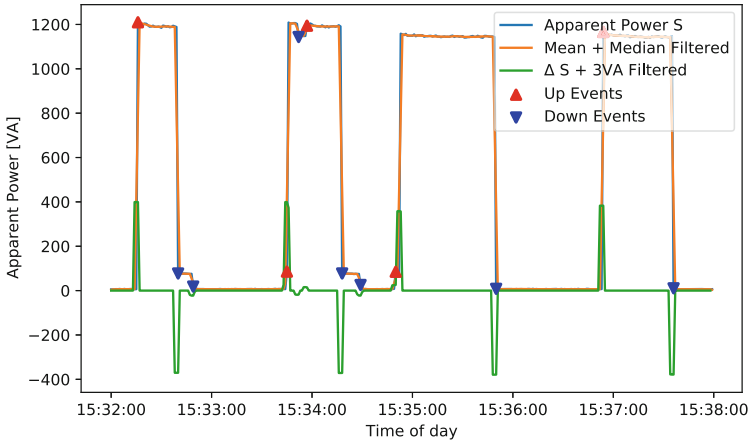


Fig. 2. Event detection applied to the 1 Hz apparent power signal of the *espresso machine* from the FIRED [28] dataset.

All significant events are clearly visible as peaks after the filtering process (green signal). The times, which are finally considered as events, are highlighted by red and blue triangles.

To be able to calculate high-frequency features for a detected event, we extract voltage and current waveforms 500 ms prior till 1 s after the timestamp of the event. We refer to this 1.5 s time interval as the Region of Interest (ROI) in the following. We further force each ROI to begin with a positive zero-crossing of the voltage measurements. All 27 features explained in the following can be extracted for each event from its corresponding ROI data.

For this evaluation, we solely use start-up events taken from individual device profiles. This means that no current is drawn in the first 500 ms. Figure 3 shows the current drawn in the ROI during a start-up event of two different appliances from the PLAID dataset.

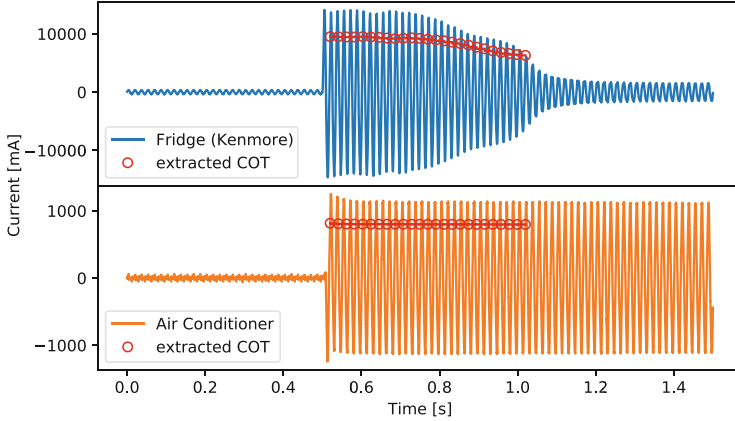


Fig. 3. Start-up transient ROI of a fridge and an air conditioner extracted from the PLAID [7] datasets. The red circles show the COT feature. (Color figure online)

3.2 Feature Selection

We have selected a set of 27 features, which have been introduced by various domain experts in related works [17, 24, 26]. All used features are summarized in Table 2 and can be extracted from the time or frequency domain of the ROI of an event. According to the Nyquist-Shannon theorem, current and voltage waveforms with a sampling rate f_s of more than $2 \cdot (18 + 1) \cdot f_0$ are required, as we analyze the signals for frequency components up to the 18th harmonic f_{18} of the grid line frequency f_0 , so $f_s > 1900$ Hz for $f_0 = 50$ Hz. To avoid aliasing artifacts, we apply a *Butterworth* low-pass filter (order = 6, $f_{\text{cutoff}} = 1$ kHz) to the current and voltage waveforms to suppress higher frequencies before extracting any feature. The feature set includes both transient and steady state features. Steady state features include several electrical measurands such as phase angle between voltage and current ($\cos\Phi$), resistance (R), admittance (Y) or active (P), reactive (Q), and apparent power (S), which can be calculated on the basis of a single main cycle. Transient features such as Current Over Time (COT) or Temporal Centroid (TC) describe the change of certain electrical characteristics (such as the current) over a certain time window. The set further includes features, which stem from excessive feature engineering such as the V-I Trajectory (VIT). The VIT was first introduced by Lam et al. [23] in 2007. The authors state that it shows a very high discriminative power, which has been proven by other researchers such as [12, 17, 29]. To calculate the VIT, the first ten periods of the current and voltage waveforms after the event are averaged and normalized. Afterwards, the averaged period is sub-sampled to 20 samples resulting in a feature vector of size 40 if voltage and current are linked together. Figure 4 shows

the VIT of six different appliances from the FIRED [28] dataset. While we can assume that most of these can be distinguished quite well (e.g. *television*, *fridge*, *vacuum cleaner*, *smartphone charger*), some devices like the *espresso machine* and the *kettle* may be difficult to keep apart using VIT as the exclusive feature.

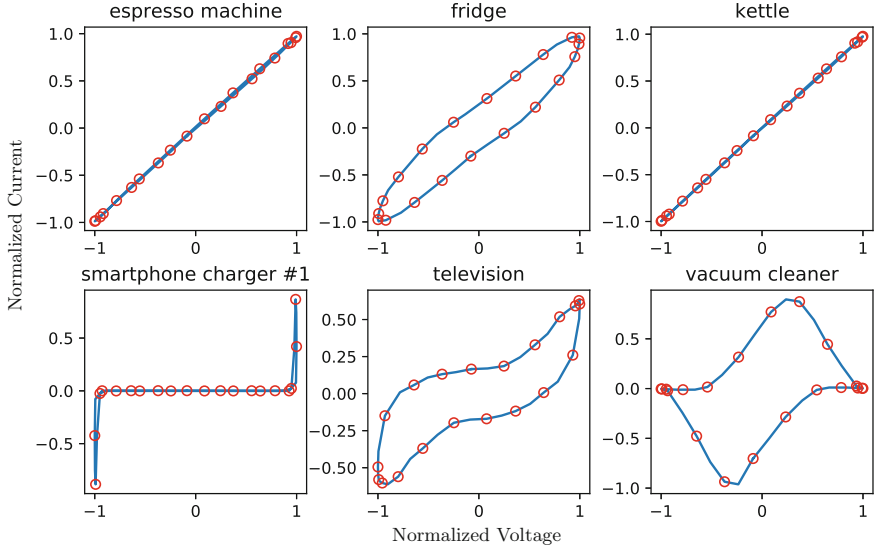


Fig. 4. Averaged and normalized VIT of six different appliances from the FIRED [28] dataset. The red dots show the sub-sampled values used in the feature vector. (Color figure online)

A second feature that stems from feature engineering is the relative Harmonic Energy Distribution (HED). The HED is a vector containing the first 18 harmonic current components normalized by the magnitude of the fundamental frequency as

$$HED = \frac{1}{x_{f_0}} \cdot [x_{f_1}, x_{f_2}, \dots, x_{f_{18}}]. \tag{2}$$

Figure 5 shows the normalized spectrum of two appliances with a strong odd-even harmonic imbalance from the BLOND [22] dataset. The extracted HED is marked with red circles.

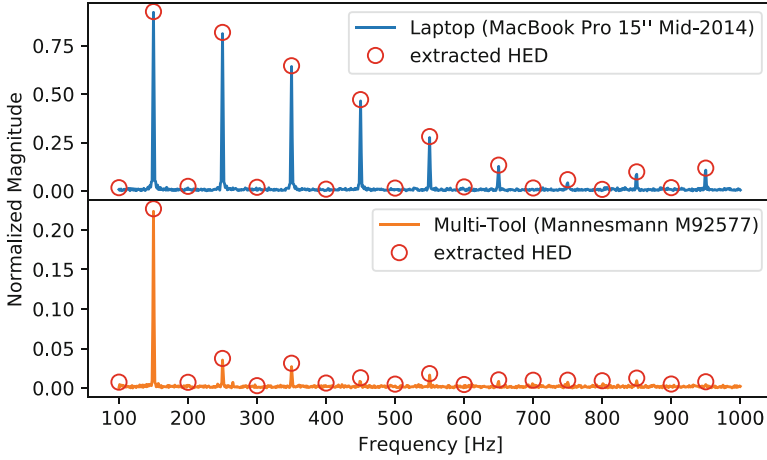


Fig. 5. The spectra of a notebook and a rotary multi-tool included in the BLOND [22] dataset, normalized to their fundamental frequency f_0 . Both devices induce a strong odd-even harmonic imbalance. The extracted HED is highlighted by red circles. (Color figure online)

The feature Current Over Time (COT) describes the amount of Root Mean Square (RMS) current in the first 25 consecutive mains cycles after an event. The mains cycle in which the event happens is not included, as its corresponding RMS current depends on the specific time the event occurred within the cycle.

$$COT = [I_{RMS(1)}, I_{RMS(2)}, \dots, I_{RMS(25)}]. \quad (3)$$

Figure 3 shows the current signal (ROI) of two appliances from the PLAID [7] dataset and the extracted COT.

For the corresponding formulas to calculate the remaining features used in this work (see Table 2), we refer to Kahl et al. [17] and Liang et al. [24]. Since we use feature combinations with different ranges, we apply feature scaling to prevent undesired feature weighting. Each dimension x in the feature space is scaled using z -score normalization by $x_{scaled} = \frac{x-\mu}{\sigma}$ with μ being the mean of all training samples and σ being the standard deviation.

3.3 Classifiers

We used four different classifiers in this work: (1) SVM, (2) k NN, (3) RF, and (4) XGBoost. These have been specifically selected for the following reasons: As will become apparent in the following, the number of training samples, i.e. appliance events, is comparably low. The used classifiers generally work quite well on smaller training sets ($< 50k$ samples) compared to e.g. ANN. The number of events differ depending on the appliance type (e.g. more fridge events than iron events) resulting in imbalanced training sets. While k NN is generally invariant to

imbalanced data, RF, SVM, and XGBoost can be adapted using class weighting or resampling strategies. Furthermore, all classifiers can be easily adapted to multi-class classification tasks and, due to their low hyper-parameter space, allow a comparably fast retraining. We applied a grid search technique to tune the parameters of each classifier based on the parameter sets listed in Table 1. For all remaining hyper-parameters, the standard values of the *scikit-learn* library are used.

Table 1. Hyper-parameter grid used while tuning each classifier.

Classifier	Parameters	Combinations
kNN	$k \in [1, 2, \dots, 20]$	20
SVM	$C \in [0.01, 0.1, 1, 10, 100, 1000]$ $\gamma \in [10000, 1000, 100, 10, 1, 0.1, 0.01]$ $kernel \in [RBF, linear]$	84
RF	$max_{depth} \in [10, 20, \dots, 100]$ $n_{estimators} \in [10, 50, 100, 1000]$	40
XGBoost	$\gamma \in [0.5, 1, 1.5]$ $n_{estimators} \in [100, 200, 1000]$	9

3.4 Metrics and Cross Validation

For each dataset, all events were shuffled and split into 80% training and 20% test samples (stratified). This allows to estimate the classification score when picking events at random as a potential NILM system would be exposed to. During grid search we applied a 5-fold random stratified split Cross Validation (CV) and averaged the results for an improved generalization estimate. During CV and for the reported scores, the confusion matrix notation in terms of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) is used to calculate Accuracy (Acc), Precision (Pre), Recall (Rec) and F_1 score (F_1) as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (4) \quad Pre = \frac{TP}{TP + FP} \quad (5)$$

$$Rec = \frac{TP}{TP + FN} \quad (6) \quad F_1 = \frac{TP}{TP + 0.5 \cdot (FP + FN)} \quad (7)$$

We use macro-averaging and calculate the unweighted means of each metric. Therefore, all classes contribute equally to the average of each metric ensuring that a class with more support in terms of the available number of samples (i.e. events) is not preferred. To simplify evaluation, we treat two different appliances of the same type (e.g. two monitors) as the same target class (\rightarrow monitor). Classes with a support of less than 5 samples are removed from the evaluation.

4 Standalone Analysis

In a first step, each feature is evaluated individually by training each classifier solely on a single feature. As Hyper-Parameter Optimization (HPO) is performed for each classifier, each dataset, and each feature individually, a total of $4 \cdot 4 \cdot 27 = 432$ different grid search instances are evaluated. The final results are reported in Table 2 and represent the F_1 -scores of the selected models applied to the test set. The results show that some features alone (e.g. VIT, WFA, COT, or HED) already show decent classification capabilities (F_1 -score > 0.8) while other features like Positive-Negative half cycle Ratio (PNR) or Periods to Steady State (PSS) stand out with exceptionally poor F_1 -scores. As found by Kahl et al. [17] among others, these features may be bad at discerning different appliances but can be used to recognize specific appliances, which exhibit certain electrical characteristics. In the time domain, e.g., the VIT already reached an F_1 -score of 0.99 and 0.95 on the laboratory datasets WHITED and PLAID, respectively. Those high scores could not be matched for the FIRED and BLOND datasets, which represent data closer to a real-world scenario. In the spectral domain, the HED achieves comparatively high scores of 0.97 on WHITED and PLAID while again not matching such performance on FIRED (0.89) and BLOND (0.8). Log Attack Time (LAT), PNR, Max-Min Ratio (MAMI), Max-Inrush Ratio (MIR), PSS, and Spectral Flatness (SPF) show a very low average F_1 -score ($\varnothing < 0.2$). As found by Kahl et al. [17] among others, these features may be bad at distinguishing a larger set of different appliances but can be used to recognize specific appliances, which exhibit certain electrical characteristics. Interestingly, those features (except MAMI) show consistent better results on BLOND and PLAID compared to FIRED and WHITED. Both BLOND and PLAID have a larger inner-class variability compared to FIRED and WHITED indicating that these features might still improve classification performance if more data are available for training.

Table 2. Classification results of a single feature applied to each dataset (WHITED, PLAID, FIRED, and BLOND) using four classifiers (k NN, SVM, RF, and XGBoost (xgb)). HPO using grid search and 5-fold CV has been applied. The features with the highest F_1 -scores for each dataset are highlighted bold in the time and spectral domain, respectively.

Feature	Dim.	WHITED				PLAID				FIRED				BLOND				\varnothing
		knn	svm	rf	xgb	knn	svm	rf	xgb	knn	svm	rf	xgb	knn	svm	rf	xgb	
Time domain																		
Active power (P)	1	0.49	0.45	0.48	0.51	0.58	0.53	0.56	0.57	0.65	0.62	0.63	0.66	0.5	0.49	0.49	0.49	0.54
Reactive power (Q)	1	0.29	0.3	0.31	0.32	0.37	0.41	0.43	0.34	0.47	0.53	0.54	0.52	0.36	0.32	0.36	0.34	0.39
Apparent power (S)	1	0.53	0.49	0.48	0.5	0.45	0.46	0.43	0.43	0.59	0.62	0.62	0.6	0.41	0.42	0.41	0.43	0.49
Resistance (R)	1	0.52	0.5	0.49	0.55	0.43	0.4	0.44	0.46	0.68	0.55	0.66	0.65	0.41	0.42	0.4	0.42	0.5
Admittance (Y)	1	0.51	0.5	0.49	0.55	0.43	0.43	0.44	0.46	0.68	0.63	0.66	0.65	0.41	0.41	0.39	0.42	0.5
Crest factor (CF)	1	0.15	0.17	0.17	0.18	0.38	0.36	0.32	0.39	0.33	0.33	0.32	0.31	0.42	0.31	0.41	0.42	0.31
Form factor (FF)	1	0.27	0.22	0.26	0.26	0.44	0.44	0.43	0.46	0.36	0.3	0.37	0.37	0.34	0.34	0.35	0.33	0.35

(continued)

Table 2. (continued)

Feature	Dim.	WHITED				PLAID				FIRED				BLOND				⊙
		knn	svm	rf	xgb	knn	svm	rf	xgb	knn	svm	rf	xgb	knn	svm	rf	xgb	
Log attack time (LAT)	1	0.05	0.05	0.05	0.05	0.12	0.16	0.15	0.15	0.1	0.09	0.09	0.09	0.19	0.22	0.23	0.19	0.12
Temporal centroid (TC)	1	0.15	0.17	0.15	0.15	0.38	0.42	0.37	0.42	0.3	0.22	0.23	0.22	0.25	0.23	0.27	0.25	0.26
Positive-negative half cycle ratio (PNR)	1	0.04	0.03	0.09	0.05	0.19	0.18	0.16	0.19	0.19	0.19	0.2	0.19	0.18	0.15	0.18	0.12	0.14
Max-min ratio (MAMI)	1	0.06	0.06	0.06	0.07	0.28	0.26	0.18	0.25	0.28	0.3	0.29	0.32	0.13	0.11	0.12	0.11	0.18
Peak-mean ratio (PMR)	1	0.2	0.19	0.16	0.19	0.4	0.3	0.43	0.37	0.38	0.37	0.37	0.39	0.4	0.38	0.42	0.38	0.33
Max-inrush ratio (MIR)	1	0.07	0.04	0.07	0.06	0.16	0.17	0.16	0.14	0.12	0.12	0.11	0.1	0.13	0.13	0.14	0.12	0.12
Mean-variance ratio (MVR)	1	0.21	0.24	0.3	0.28	0.41	0.34	0.36	0.38	0.42	0.41	0.4	0.4	0.33	0.34	0.35	0.32	0.34
Waveform distortion (WFD)	1	0.27	0.24	0.24	0.23	0.35	0.36	0.38	0.39	0.44	0.44	0.43	0.45	0.3	0.31	0.28	0.3	0.34
Period to steady state (PSS)	1	0.01	0.03	0.03	0.03	0.11	0.12	0.12	0.12	0.09	0.12	0.12	0.12	0.11	0.12	0.12	0.12	0.09
Phase angle ($\cos\Phi$)	1	0.26	0.24	0.26	0.27	0.48	0.46	0.42	0.49	0.45	0.45	0.43	0.44	0.43	0.44	0.38	0.44	0.4
Inrush current ratio (ICR)	1	0.17	0.07	0.15	0.16	0.27	0.23	0.27	0.36	0.41	0.22	0.44	0.37	0.27	0.25	0.26	0.25	0.26
Waveform approximation (WFA)	20	0.92	0.91	0.93	0.83	0.93	0.92	0.9	0.9	0.91	0.93	0.88	0.84	0.84	0.82	0.75	0.8	0.88
Current over time (COT)	25	0.8	0.84	0.93	0.86	0.81	0.72	0.86	0.87	0.88	0.93	0.95	0.94	0.8	0.81	0.83	0.83	0.85
V-I Trajectory (VIT)	40	0.93	0.99	0.95	0.89	0.91	0.93	0.95	0.88	0.7	0.82	0.77	0.72	0.82	0.85	0.71	0.78	0.85
Spectral domain																		
Total harmonic distortion (THD)	1	0.37	0.39	0.34	0.37	0.49	0.5	0.48	0.51	0.42	0.4	0.43	0.41	0.38	0.34	0.39	0.38	0.41
Spectral flatness (SPF)	1	0.06	0.07	0.09	0.1	0.2	0.19	0.21	0.22	0.17	0.17	0.15	0.18	0.23	0.21	0.19	0.21	0.17
Odd-even harmonics ratio (OER)	1	0.09	0.09	0.12	0.09	0.26	0.28	0.26	0.3	0.28	0.25	0.26	0.25	0.29	0.25	0.3	0.21	0.22
Spectral centroid (SC)	1	0.12	0.12	0.13	0.14	0.31	0.3	0.25	0.26	0.22	0.2	0.22	0.23	0.33	0.31	0.29	0.3	0.23
Tristimulus (TRI)	3	0.89	0.86	0.86	0.79	0.87	0.82	0.82	0.79	0.77	0.81	0.84	0.77	0.61	0.64	0.63	0.59	0.77
Harmonic energy distribution (HED)	18	0.97	0.93	0.97	0.83	0.97	0.88	0.94	0.93	0.85	0.85	0.89	0.88	0.7	0.8	0.77	0.77	0.87

Unsurprisingly, features showing better performance have the drawback of a high dimensionality (e.g., 40 for VIT and 20 for WFA). If the focus is shifted towards the best performing scalar features (F_1 -score > 0.4), classical electrical features such as P , S , R , Y , $\cos\Phi$, and Total Harmonic Distortion (THD) can be identified. It is argued that these features may be of choice for lightweight NILM algorithms deployed on resource constrained systems such as smart meters.

5 Feature Selection

Some of the features already performed quite well in the standalone analysis. However, it can be assumed that the combination of multiple features leads to even better classification scores. While combining all 27 features may result in better classification performance, the number of dimensions should be held small to save computational resources and to prevent performance degradation, which stem from larger feature spaces also known as *the curse of dimensionality*. Therefore, in a second analysis several feature combinations are evaluated based not only on their final classification score but also on their overall dimensionality. While the standalone feature VIT already reaches an F_1 -score of up to 0.99 in the experiments, its large dimensionality may hamper a possible application. Furthermore, it might be possible that a combination of multiple features of smaller dimensionality even outperforms VIT. Consequently, a second analysis is conducted for which the combination of several features up to a maximum dimensionality N is examined.

While Principal Component Analysis (PCA) can deliver valuable information about the expressiveness of a certain feature, it does look at each feature dimension individually and, therefore, does not account that other dimensions are already calculated for certain multidimensional features such as e.g. HED. Since an excessive evaluation that considers all possible feature combinations is not feasible ($\sum_{k=0}^{27} \binom{27}{k}$), a simple greedy heuristic i.e. a sequential selection algorithm is used. The algorithm starts by adding the best performing scalar feature ($feat^x$) to a feature set ($F_0 = \{feat^x\}$ with dimensionality $N_0 = 1$). It then evaluates all combinations of F_i with another scalar feature $feat^j$. The best performing combined set ($F_{i+1} = F_i \cup \{feat^j\}$) is stored resulting in a dimensionality of $N_{i+1} = N_i + 1$. It is then checked if any of the possible combinations of non-scalar features (F_{i+1}^{NS}), which result in the same dimensionality N_{i+1} , outperforms F_{i+1} . If this is not the case, the algorithm continues with F_{i+1} , otherwise F_{i+1}^{NS} is used. This process is repeated until a maximum dimensionality N_{max} is reached. The performance of each tested feature set is stored.

6 Combined Analysis

The selection scheme is executed for all 27 features on all datasets with a k NN, SVM, and RF classifier. XGBoost was left out due to its extensive computational requirements and comparable low performance on the standalone feature evaluation (see Table 2). The results of this experiment, which are visualized in Fig. 6, highlight that feature combinations with rather low dimensionality ($N < 10$) already lead to classification scores of over 0.98 on WHITED and PLAID. The evaluation further highlights that the performance on recordings in laboratory setups (PLAID and WHITED) is generally better and more consistent compared to more representative real-world data (FIRED and BLOND). This is, however, expected due to the lower noise-level in laboratory environments.

In this evaluation, all classifiers performed equally well. Only for the BLOND dataset, SVM classifiers outperform the others by quite a margin. Table 3 shows

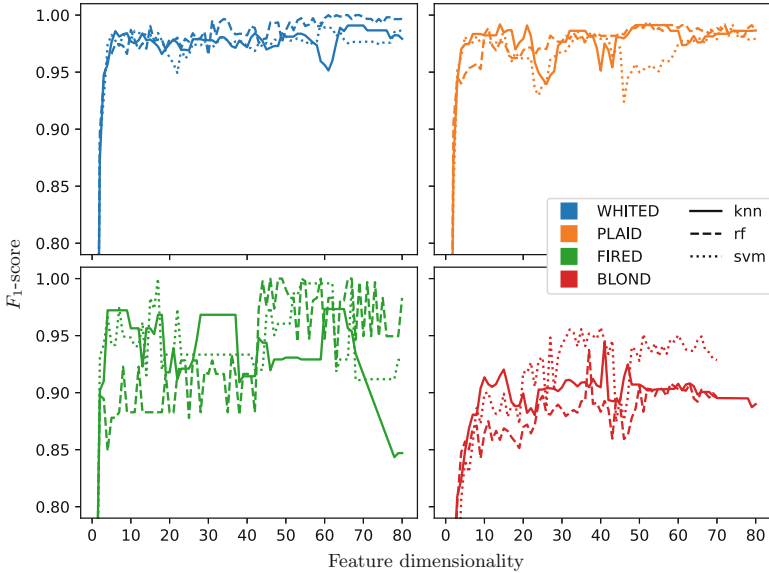


Fig. 6. Results of the proposed feature selection strategy for all classifiers (line styles) and all datasets (line colors).

the specific feature sets that have been chosen by the selection scheme for different dimensionalities N . As a tradeoff between dimensionality, performance, and computational effort, it is proposed to use features up to a dimensionality of 25. The feature set, which has been proposed by the algorithm for $N = 25$ (see Table 3), depends on the used classifier. However, it always includes the features WFA and Tristimulus (TRI). It is decided to supplement these features with P and $\cos\Phi$ resulting in the proposed feature set $[P, \cos\Phi, TRI, WFA]$. P has already been evaluated in Table 2 as being the best scalar feature with an average F_1 -score of 0.54. $\cos\Phi$ reoccurs in nearly all feature sets (see Table 3) and is added to accommodate the reactive component, which may be introduced by an appliance. TRI further showed high classification results in Table 2 and represents the only frequency domain feature in the set. TRI is preferred over the actually better performing HED (see Table 2), as it requires only three dimensions instead of 18. From the corresponding formulas, it can be seen that TRI also represents a compressed form of the HED feature. While WFA (with a dimensionality of 20) did not outperform 20 scalar features, its simple calculation and the overall best results obtained in the standalone feature analysis (see Table 2) justifies its inclusion in the set that is finally proposed.

With these four features, the proposed set is of comparatively small dimensionality, computationally lightweight enough for resource constrained systems, and still delivers decent classification results. To emphasize this, the proposed set and the combination of all 27 features was evaluated on all classifiers and datasets. The results are shown in Table 4. A slight performance increase can

even be identified if the proposed feature set is used instead of all features due to the *course of dimensionality*. With a dimension reduction from 128 to 25, the proposed set still outperforms the combination of all features, highlighting the effectiveness of the proposed feature set.

Table 3. Used features for selected dimensionalities N of the proposed feature selection strategy. The F_1 -scores for each dataset and classifier (Clf.) are shown in addition to the F_1 -scores averaged over all datasets.

N	Clf.	Featureset	WHITED	PLAID	FIRE	BLOND	$\circ F_1$
5	knn	$P, Y, THD, \cos \Phi, OER$	0.98	0.97	0.94	0.82	0.93
5	rf	$P, Y, THD, \cos \Phi, MVR$	0.96	0.96	0.93	0.77	0.91
5	svm	$P, Y, THD, \cos \Phi, MVR$	0.96	0.98	1.0	0.81	0.93
22	knn	$P, Y, THD, \cos \Phi, OER, FF, R, S,$ $PMR, WFD, TC, ICR, MVR, SC, Q, CF,$ $PNR, MAMI, SPF, LAT, PSS, MIR$	0.87	0.96	0.91	0.89	0.91
22	rf	$P, Y, THD, \cos \Phi, MVR, WFD, SC,$ $LAT, OER, SPF, ICR, CF, S, PSS,$ $MIR, PNR, R, MAMI, Q, TC, PMR, FF$	0.98	0.97	0.92	0.87	0.93
22	svm	$P, Y, THD, \cos \Phi, MVR, CF, R,$ $OER, FF, SC, Q, PMR, S, WFD, SPF,$ $ICR, TC, PNR, MAMI, LAT, PSS, MIR$	0.95	0.96	0.97	0.88	0.94
25	knn	$WFA, TRI, \cos \Phi, Y$	0.99	1.0	0.92	0.9	0.95
25	rf	WFA, TRI, LAT, S	0.98	0.98	0.91	0.88	0.94
25	svm	WFA, TRI, S, THD	0.98	0.95	0.93	0.95	0.95
32	knn	$WFA, TRI, \cos \Phi, Y, R, OER,$ MVR, THD, P, S, SC	0.99	1.0	0.93	0.89	0.95
32	svm	$WFA, TRI, S, THD, R, CF, Y,$ $\cos \Phi, MVR, FF, OER$	0.98	0.97	0.93	0.96	0.96
32	rf	$WFA, TRI, LAT, S, CF, Y, MAMI,$ SC, SPF, MIR, OER	0.98	0.97	0.91	0.89	0.94

The average F_1 -scores of the proposed set over all datasets exceed 0.94 independent of the used classifier. The RF classifier performs best with an average F_1 -score of 0.98. It is, however, noted that a computationally fairly simple k NN classifier with $k=1$ already achieves a rather high F_1 -score of 0.97 on WHITED and 0.98 on PLAID. k NN is a so called *lazy learning* algorithm that requires no internal parameter tuning except for the choice of the number of neighbors (k) to consider. During training, the complete training set is stored. During inference, a new sample is assigned to the most common class within its k -nearest neighbors. To reduce the required memory of a k NN classifier, which linearly increases with the number of training samples, the Condensed Nearest Neighbor Rule [11] can be applied. Because of its simple training and the ability to reduce the required memory, it is argued that k NN should be the classifier of choice if deployed (including training) on systems with small computational resources such as typical smart meters. However, for systems with sufficient computational power, SVM and RF should be the classifiers of choice. XGBoost has shown enormous potential by leading many machine learning competitions during the recent years [9]. Even though it exhibits the worst performance across all classifiers in the analysis at hand, it is argued that XGBoost might still outperform RF and SVM

Table 4. Classification results for all 27 features and for the proposed feature combination $[P, \cos\Phi, TRI, WFA]$. The best results are highlighted in bold.

Clf.	WHITED				PLAID				FIRED				BLOND				$\emptyset F_1$
	Pr	Re	Ac	F_1	Pr	Re	Ac	F_1	Pr	Re	Ac	F_1	Pr	Re	Ac	F_1	
Using all 27 feature; overall vector dimension: 128																	
knn	0.97	0.97	0.96	0.97	0.98	0.97	0.97	0.97	0.89	0.87	0.99	0.86	0.94	0.94	0.99	0.94	0.93
svm	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.91	0.94	1.0	0.92	0.96	0.88	0.99	0.91	0.95
rf	1.0	1.0	0.99	1.0	0.99	0.98	0.98	0.98	1.0	0.98	1.0	0.98	0.97	0.88	0.99	0.91	0.97
xgb	0.97	0.96	0.97	0.96	0.97	0.96	0.97	0.97	0.9	0.87	0.99	0.87	0.95	0.86	0.99	0.89	0.92
Using the feature set $[P, \cos\Phi, TRI, WFA]$; overall vector dimension: 25																	
knn	0.97	0.97	0.96	0.97	0.98	0.98	0.98	0.98	0.91	0.92	0.99	0.91	0.95	0.89	0.99	0.91	0.94
svm	0.98	0.97	0.97	0.97	0.96	0.96	0.95	0.96	0.92	0.95	1.0	0.93	0.94	0.96	0.99	0.95	0.95
rf	0.99	0.99	0.98	0.99	1.0	0.99	0.99	0.99	1.0	0.98	1.0	0.98	0.96	0.95	0.99	0.95	0.98
xgb	0.94	0.91	0.93	0.91	0.97	0.96	0.97	0.97	1.0	0.97	1.0	0.98	0.95	0.86	0.98	0.89	0.94

for other hyperparameter choices as the ones tested during these evaluations (see the used grid search parameters in Table 1). However, due to its large hyperparameter space and, therefore, extensive training time, RF and SVM were selected in favor, representing a tradeoff between the required training time and possible gain in classification performance. Figure 7 shows the confusion matrix of the RF classifier using the proposed feature set on the PLAID dataset (the corresponding performance metrics are shown in Table 4. Despite the overall F_1 -score of 0.98, only some appliances with rotary motors (fan, heater, and air conditioner) are confused with one another. Due to the outstanding performance of

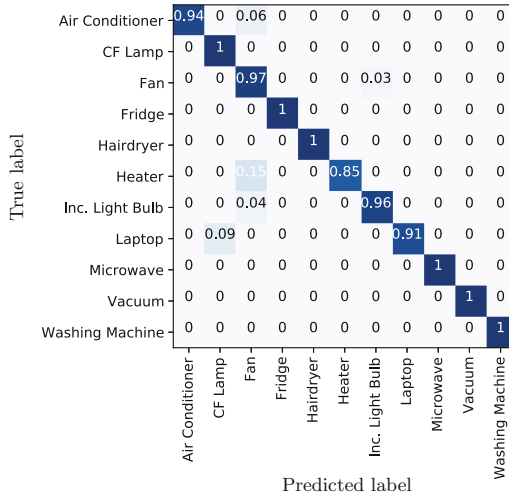


Fig. 7. Confusion matrix of a RF classifier with the feature set $[P, \cos\Phi, TRI, WFA]$ applied to the PLAID dataset.

the RF classifier with the feature set $[P, \cos\Phi, TRI, WFA]$, it is proposed to use their combination as a benchmarking algorithm when comparing novel appliance classification algorithms, similar to the low-frequency disaggregation algorithms that have been implemented as benchmarks in NILMTK [3].

7 Conclusion

In this work, we used four electricity datasets recorded at higher sampling rates to evaluate 27 features and four classifiers for the task of event classification. The best standalone features are P and WFA with corresponding F_1 -scores of 0.54 and 0.88, respectively. A feature selection algorithm revealed the feature set $[P, \cos\Phi, TRI, WFA]$ for a desired dimensionality of 25. This set achieved F_1 -scores of 0.98 on average using a RF classifier. As all classifiers appeared to be suitable, the performance of classifier ensembles should be investigated in future work.

References

1. Anderson, K., et al.: BLUED: a fully labeled public dataset for event-based non-intrusive load monitoring research. In: Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD), vol. 7. ACM (2012)
2. Anderson, K.D., Berges, M.E., Ocleanu, A., Benitez, D., Moura, J.M.: Event detection for Non Intrusive load monitoring. IECON Proceedings (Industrial Electronics Conference), pp. 3312–3317 (2012). <https://doi.org/10.1109/IECON.2012.6389367>
3. Batra, N., et al.: NILMTK: an open source toolkit for non-intrusive load monitoring. In: Proceedings of the 5th International Conference on Future Energy Systems, pp. 265–276. ACM (2014)
4. Davies, P., Dennis, J., Hansom, J., Martin, W., Stankevicius, A., Ward, L.: Deep neural networks for appliance transient classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2019)
5. De Baets, L., Ruysinck, J., Develder, C., Dhaene, T., Deschrijver, D.: Appliance classification using vi trajectories and convolutional neural networks. *Energy Build.* **158**, 32–36 (2018)
6. Faustine, A., et al.: UNet-NILM: a deep neural network for multi-tasks appliances state detection and power estimation in NILM. In: Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring (2020)
7. Gao, J., Giri, S., Kara, E.C., Bergés, M.: Plaid: a public dataset of high-resolution electrical appliance measurements for load identification research. In: proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings (2014)
8. Gao, J., Kara, E.C., Giri, S., Bergés, M.: A feasibility study of automated plug-load identification from high-frequency measurements. In: IEEE Global Conference on Signal and Information Processing (GlobalSIP) (2015)
9. GitHub: Xgboost - extreme gradient boosting (2020). <https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions>. List of Machine Learning Challenge Winners based on XGBoost

10. Hart, G.W.: Nonintrusive appliance load monitoring. *Proc. IEEE* **80**(12), 1870–1891 (1992)
11. Hart, P.: The condensed nearest neighbor rule (CORRESP). *IEEE Trans. Inf. Theory*. **14**(3), 515–516 (1968)
12. Hassan, T., Javed, F., Arshad, N.: An empirical investigation of VI trajectory based load signatures for non-intrusive load monitoring. *IEEE Trans. Smart Grid* **5**(2), 870–878 (2013)
13. Huber, P., Calatroni, A., Rumsch, A., Paice, A.: Review on deep neural networks applied to low-frequency NILM. *Energies* **14**(9), 2390 (2021). <https://doi.org/10.3390/en14092390>
14. Jorde, D., Kriechbaumer, T., Jacobsen, H.A.: Electrical appliance classification using deep convolutional neural networks on high frequency current measurements. In: *IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)* (2018)
15. Kahl, M., Haq, A.U., Kriechbaumer, T., Jacobsen, H.A.: WHITED - A worldwide household and industry transient energy data set. In: *3rd International Workshop on Non-Intrusive Load Monitoring* (2016)
16. Kahl, M., Kriechbaumer, T., Haq, A.U., Jacobsen, H.A.: Appliance classification across multiple high frequency energy datasets. In: *IEEE International Conference on Smart Grid Communications (SmartGridComm)* (2017)
17. Kahl, M., Ul Haq, A., Kriechbaumer, T., Jacobsen, H.A.: A comprehensive feature study for appliance recognition on high frequency energy data. In: *Proceedings of the 8th ACM International Conference on Future Energy Systems* (2017)
18. Kelly, J., Knottenbelt, W.: The UK-dale dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Sci. Data* **2**, 150007 (2015)
19. Kim, J.G., Lee, B.: Appliance classification by power signal analysis based on multi-feature combination multi-layer LSTM. *Energies* **12**(14), 2804 (2019)
20. Kolter, J.Z., Johnson, M.J.: REDD: a public data set for energy disaggregation research. In: *Workshop on Data Mining Applications in Sustainability (SIGKDD)*, San Diego, CA, vol. 25, pp. 59–62 (2011)
21. Kolter, Z., Jaakkola, T., Kolter, J.Z.: Approximate inference in additive factorial HMMs with application to energy disaggregation. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1472–1482 (2012). <http://people.csail.mit.edu/kolter/lib/exe/fetch.php?media=pubs:kolter-aistats12.pdf>
22. Kriechbaumer, T., Jacobsen, H.A.: BLOND, a building-level office environment dataset of typical electrical appliances. *Sci. Data* **5**, 180048 (2018)
23. Lam, H.Y., Fung, G., Lee, W.: A novel method to construct taxonomy electrical appliances based on load signatures. *IEEE Trans. Consum. Electron.* **53**(2), 653–660 (2007)
24. Liang, J., Ng, S.K., Kendall, G., Cheng, J.W.: Load signature study-part i: basic concept, structure, and methodology. *IEEE Trans. Power. Deliv.* **25**, 551–560 (2010)
25. Makonin, S., Popowich, F., Bajic, I.V., Gill, B., Bartram, L.: Exploiting HMM sparsity to perform online real-time nonintrusive load monitoring. *IEEE Trans. Smart Grid* (2015). <https://doi.org/10.1109/TSG.2015.2494592>
26. Sadeghianpourhamami, N., Ruyssinck, J., Deschrijver, D., Dhaene, T., Develder, C.: Comprehensive feature selection for appliance classification in NILM. *Energy Build.* **151**, 98–106 (2017)

27. Völker, B., Pfeifer, M., Scholl, P.M., Becker, B.: Annoticity: a smart annotation tool and data browser for electricity datasets. In: Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring, pp. 1–5 (2020)
28. Völker, B., Pfeifer, M., Scholl, P.M., Becker, B.: FIRED: A fully labeled high-frequency electricity disaggregation dataset. In: Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (2020)
29. Wang, A.L., Chen, B.X., Wang, C.G., Hua, D.: Non-intrusive load monitoring algorithm based on features of v-i trajectory. *Elect. Power Syst. Res.* **157**, 134–144 (2018)
30. Weiss, M., Helfenstein, A., Mattern, F., Staake, T.: Leveraging smart meter data to recognize home appliances. In: IEEE International Conference on Pervasive Computing and Communications (2012)
31. Wild, B., Barsim, K.S., Yang, B.: A new unsupervised event detector for non-intrusive load monitoring. In: 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 73–77 (2015)
32. Yang, C.C., Soh, C.S., Yap, V.V.: A systematic approach in appliance disaggregation using k-nearest neighbours and naive bayes classifiers for energy efficiency. *Energy Effic.* **11**(1), 239–259 (2018)
33. Zoha, A., Gluhak, A., Imran, M.A., Rajasegarar, S.: Non-intrusive load monitoring approaches for disaggregated energy sensing: a survey. *Sensors* **12**(12), 16838–16866 (2012)