Sustainable Cities and Society 62 (2020) 102399

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs

An empirical exploration of performance metrics for event detection algorithms in Non-Intrusive Load Monitoring

Lucas Pereira[a,b,*], Nuno Nunes[a,b]

[a] ITI, LARSyS, Polo Científico e Tecnológico da Madeira, floor-2, 9020-105 Funchal, Portugal
[b] Técnico Lisboa, University of Lisbon, Av. Rovisco Pais 1, Torre Norte, floor 6-8, 1049-001 Lisboa, Portugal

ARTICLE INFO

Keywords:
NILM
Event detection
Performance evaluation
Performance metrics
Empirical research

ABSTRACT

The field of Non-Intrusive Load Monitoring (NILM) gained prominence due to its promise of inferring the energy consumption of individual appliances by analyzing only the aggregated consumption. Still, despite some research efforts towards producing meaningful comparisons between approaches, it is not yet possible to find a proven and formally accepted set of metrics to do this. Against this background, this paper focuses on understanding the challenges of defining a consistent set of performance metrics for this problem. More concretely, it reports on an empirical exploration of 23 performance metrics' behavior when applied to event-detection algorithms, identifying relationships and clusters between them. The results indicate that when applied to this problem, the performance metrics will show some considerable differences in behavior compared to other, more traditional, machine-learning domains. The results also suggest that most of the differences occur due to the unbalanced nature of the event detection problem, in which the number of positive cases (True Positives and True Negatives) is much higher than the number of false situations (False Positives and False Negatives). Furthermore, the results suggest that additional research is needed to find proper domain-specific performance metrics that take full consideration of the properties of the aggregated load.

1. Introduction

The worldwide energy consumption has been steadily increasing in the past decades, emerging as one of the main contributors to climate change, and therefore one of the main targets of the climate action defined by the United Nations mainly through Sustainable Development Goals (SDGs) 7, 11, and 13.[1]

Having the ability to keep track of individual appliance consumption in households as been deemed one of the critical drivers for reducing the environmental impacts of electricity over consumption at both micro and macro scales. For example, by promoting energy-saving behaviors in individual consumers through eco-feedback, and by enabling the delivery of wide-scale personalized energy efficiency programs such as demand-response (Armel, Gupta, Shrimali, & Albert, 2013; Rolnick et al., 2019). Yet, electricity presents a particular problem to consumers and researchers alike, as unlike other utilities such as water and gas, it is an invisible resource with no visible form, flow, or weight. Thus, making it hard for people to gauge the quantity of electricity consumed by individual appliances (Attari, DeKay, Davidson, & de Bruin, 2010; Chisik, 2011; Pereira & Chisik, 2017).

In this context, the field of Non-Intrusive Load Monitoring (NILM), gained prominence due to the promise of inferring the energy consumption of individual appliances by applying advanced machine-learning and signal-processing techniques to the aggregated measurements taken at a limited number of locations in the building (Hart, 1992). This discipline greatly contrasts the more intrusive and costly monitoring technology that involves deploying multiple sensors (Berges, Goldman, Matthews, Soibelman, & Anderson, 2011).

From the early expectations that individual appliance consumption data would promote energy-saving behaviors in individuals (Armel et al., 2013; Kelly & Knottenbelt, 2016), the expectations of NILM technology quickly evolved to the supply side, serving as the backbone technology that will enable the creation of innovative smart-grid services that go beyond helping individuals reduce energy consumption (Najafi, Moaveninejad, & Rinaldi, 2018; Townson, 2016). The potential benefits of NILM include:

- Provides an inexpensive way for utilities to segment their customers (e.g., Beckel, Sadamori, Santini, & Staake, 2015; Kavousian, Rajagopal, & Fischer, 2013). It also allows the deployment and

* Corresponding author at: ITI, LARSyS, Polo Científico e Tecnológico da Madeira, floor-2, 9020-105 Funchal, Portugal.
E-mail address: lucas.pereira@tecnico.ulisboa.pt (L. Pereira).

[1] SDGs, https://sustainabledevelopment.un.org/.

https://doi.org/10.1016/j.scs.2020.102399
Received 6 January 2020; Received in revised form 28 May 2020; Accepted 11 July 2020
Available online 17 July 2020

follow-up of long-term energy efficiency programs, which are necessary to access the long-term effectiveness of such programs (e.g., Ma, Lin, & Li, 2018; Pereira & Nunes, 2019; Pereira, Quintal, Barreto, & Nunes, 2013; Quintal, Pereira, Nunes, Nisi, & Barreto, 2013).

- Enables the creation of innovative energy efficiency services. For example, the possibility of inferring and providing eco-feedback on the everyday activities (e.g., Alcal'a, Parson, & Rogers, 2015; Belley, Gaboury, Bouchard, & Bouzouane, 2013; Stankovic, Stankovic, Liao, & Wilson, 2016), and the detection of anomalies in electric loads (e.g., Munir, Stankovic, & FailureSense:, 2014; Orji et al., 2010; Rodney & Poll, 2014).
- NILM data can also play an important role in improving existing demand forecasting and demand-side flexibility estimation algorithms. These algorithms are key in optimizing the generation and distribution of electric energy, and the integration of renewable energy sources (RES) in the grid (e.g., Gajowniczek & Zkabkowski, 2017; Kong et al., 2015; Lucas, Jansen, Andreadou, Kotsakis, & Masera, 2019; Pipattanasomporn, Kuzlu, Rahman, & Teklu, 2014; Zhai, Zhou, Wang, & He, 2020).

Since its inception, NILM research has been geared towards the development of disaggregation algorithms, as suggested by the numerous reviews in the field (e.g., Esa, Abdullah, & Hassan, 2016; Najafi et al., 2018; Nalmpantis & Vrakas, 2018; Zeifman & Roth, 2011; Zoha, Gluhak, Imran, & Rajasegarar, 2012).

In general, disaggregation fall in one of two categories: (1) event-based approaches, and (2) event-less approaches (Bergés & Kolter, 2012; Pereira & Nunes, 2018). Event-based approaches seek to disaggregate the total consumption by means of detecting and classifying individual appliance transitions in the aggregated signal (e.g., Alcalá, Ureña, Hernández, & Gualda, 2017; Barsim & Yang, 2015; Berges et al., 2011; Hart, 1985). The classified transitions are then used to estimate each appliance's energy consumption employing transition matching and optimization algorithms (e.g., Giri & Bergés, 2015; He, Jakovetic, Stankovic, & Stankovic, 2018; Zhao, Stankovic, & Stankovic, 2016). On the other hand, event-less approaches attempt to match each sample of the aggregated power to the consumption of one or a combination of different appliances. This is done through methods such as Bayesian statistics, Hidden Markov Models (HMM), and deep-learning methods such as Deep Artificial Neural Networks (Deep-Nets) (e.g., Gomes & Pereira, 2020; Harell, Makonin, Bajic, & Wavenilm, 2019; Makonin, Popowich, Bajic, Gill, & Bartram, 2016; Murray, Stankovic, Stankovic, Lulic, & Sladojevic, 2019; Yan et al., 2019).

Recently, a smaller number of researchers started to address the challenges related to the lack of adequate methods to conduct meaningful performance evaluation of the developed approaches. Identified as one of the main barriers to the real-world adoption of NILM enabled smart-meters, most of the existing efforts in performance evaluation focused on five main aspects (Pereira & Nunes, 2018): (1) public datasets (e.g., Anderson, Ocneanu, et al., 2012; Kelly & Knottenbelt, 2015; Klemenjak et al., 2019; Makonin, Ellert, Bajic, & Popowich, 2016), (2) toolkit development (e.g., Batra et al., 2014, 2019; Pereira, 2017b), (3) study of performance metrics (e.g., Anderson, Bergés, Ocneanu, Benitez, & Moura, 2012; Makonin & Popowich, 2017; Pereira & Nunes, 2017), (4) evaluation frameworks (e.g., Czarnek, Morton, Collins, Newell, & Bradbury, 2015; Pereira & Nunes, 2015b; Symeonidis, Nalmpantis, Vrakas, & Benchmark, 2019), and (5) comparability and complexity (e.g., Egarter, Pöchacker, & Elmenreich, 2015; Klemenjak, Makonin, & Elmenreich, 2020; Nalmpantis & Vrakas, 2018).

1.1. Problem statement and proposed approach

The work presented in this paper focuses on performance evaluation, more concretely in the challenge of defining a consistent and widely accepted set of performance metrics to quantify, report, and benchmark NILM algorithms.

Throughout the years, *disaggregation accuracy* has been widely used by NILM researchers when referring to the performance of their algorithms. However, *disaggregation accuracy* has a very loose definition in the sense that many different metrics fit under this umbrella term. For example, in his seminal work Hart (Hart, 1985) used the total energy explained (i.e., ratio between total estimated energy and actual energy used), whereas in Liang, Ng, Kendall, and Cheng (2010) the authors defined disaggregation accuracy in terms of, *detection accuracy* (the ratio between the correctly detected events and all the detected events), *disaggregation accuracy* (the number of correctly classified events excluding the detection errors) and *overall accuracy* (the number of correct classifications including detection errors).

As of today, there is an agreement on grouping performance metrics according to two main categories. On the one hand, the *event detection performance metrics* (ED) designed to evaluate the NILM's ability to track consumption over time, i.e., event detection and classification algorithms. On the other hand, the *energy estimation performance metrics* (EE), which refer to the metrics designed to characterize and evaluate the NILM disaggregated data against the actual ground-truth (Mayhorn, Sullivan, Petersen, Butner, & Johnson, 2016). Pereira and Nunes (2018) provides a comprehensive review of 39 NILM performance metrics according to this categorization (24 for event detection, and 15 for energy estimation).

Despite these early efforts to better understand performance metrics in the scope of NILM (e.g., Makonin & Popowich, 2017; Mayhorn, Sullivan, Fu, & Petersen, 2017; Pereira & Nunes, 2017) there is not enough knowledge about this topic to enable the definition of a widely accepted set of performance metrics and their respective thresholds. For example, in Zeifman (2012), it is suggested that to be approved by homeowners, the minimum accepted *disaggregation accuracy* should be 80–90%. However, there is no information about which metrics were considered in this study. Furthermore, since many of the metrics applied in NILM are inherited from other application domains in machine-learning (e.g., the precision and recall have their origins in the information retrieval domain (Kagolovsky & Moehr, 2003)), it is challenging to establish meaningful thresholds without first gaining a deeper understanding about the behavior of such metrics when applied to the NILM problem (Nalmpantis & Vrakas, 2018; Pereira & Nunes, 2018).

Therefore, it is crucial to understand the relationships between existing performance metrics in the scope of NILM algorithms. Having access to this information would help the researchers select the metrics that are more suitable to their needs.

To this end, this paper presents an experimental comparison of performance metrics for event detection algorithms. More precisely, it analyzes the behavior of 23 performance metrics, across five algorithms and four event detection scenarios from two public NILM datasets. The following steps are proposed:

1. **Creation of Baseline Data:** In this step, baseline performance evaluation data for event detection algorithms are created. First, different event detection models are created by employing a controlled parameter sweep across five (5) event detection algorithms. Second, each model is executed against four (4) event detection scenarios from two publicly available datasets. Finally, the 23 performance metrics are calculated for each of the event detection models executed in the previous step.
2. **Calculation of Metric Pairwise Correlations and Average Correlation Matrices:** In this step, the existence of linear *Pearson* and non-linear *Spearman* pairwise correlations between the performance metrics are investigated. The former indicates the presence and direction of any linear relationships. The latter reports on the existence of monotonic relationships (i.e., the results tend to change together but not necessarily at a linear rate). Since a correlation matrix is generated for each model-dataset pair, a single cross-

dataset matrix was created by averaging the individual pairwise correlation matrices. Ultimately, the average pairwise correlations between metrics will unveil if there are performance metrics that will yield the same ranks (i.e., if they select the same models). Likewise, this will also reveal the cases in which the performance metrics rank the different models in totally different directions. Finally, metrics with low or non-existing correlation will rank the detection models in very different ways.

3. **Cluster Analysis:** Finally, after an initial analysis, in this step any existing correlations are further studied using hierarchical clustering and *dendrograms*. Ultimately, the different metric clusters will provide a better understanding concerning which metrics tend to select similar models. Furthermore, the various cluster arrangements will also provide valuable insights concerning the theoretical guarantees of such metrics, and if they hold for the event detection problem.

1.2. Related works

The selection of appropriate performance metrics is a transversal problem in machine-learning and has received the attention of many researchers through analytic and empirical approaches.

Analytic approaches focus on exploring the theoretical properties of individual metrics, and the definition of similarity and dissimilarity scores between pairs of metrics. For example, in Flach (2003), Flach studied how the output of seven binary performance metrics is affected by changes in the relative proportions of classes in evaluated problem (i.e., dataset). According to this work, metrics should distinguished by their effective skew landscapes (i.e., variation under dataset unbalance) rather than by their actual values. Furthermore, if two metrics are to be considered equivalent, they must also be skew-equivalent, i.e., they should have the same skew ratio across all the possible variations in the class distribution. Another analytical exploration of performance metrics was bone in Sokolova and Lapalme (2009). In this work, the authors analyzed the behavior of 24 metrics in binary, multi-class, multi-label, and hierarchical classification problems to understand their ability preserve they values under class-specific changes in the confusion matrix. A metric is said to be invariant to a specific confusion-matrix modification if its value does not change. Two metrics are considered similar if they have the same invariants.

Concerning empirical approaches, the general idea is to run several learning algorithms across a selection of datasets and calculate their performance using different performance metrics. These metrics are later compared to understand their behavior across the different algorithms and datasets. For instance, in Caruana and Niculescu-Mizil (2004), Ferri, Hernández-Orallo, and Modroiu (2009) the authors studied the performance of different classification algorithms across multiple problems (one or more datasets represent each problem) using a variety of metrics with the goal of understanding which were better suited in each situation. To state more concretely, Caruana and Niculescu-Mizil (2004) experimented with nine performance metrics and seven binary classification problems, using multi-dimensional scaling (MDS) and correlation for the comparison. As for Ferri et al. (2009), the authors tested with 18 metrics and 30 datasets (15 binary and 15 multi-class), using correlation and dendrograms for the comparison.

While all these works have addressed metrics similarity from many different perspectives, a key take-away message is that despite having some similarities (reflected by pairwise correlations greater than 0.5), most metrics are fundamentally different as they are designed to address different aspects of the problem. Interestingly, it was also found that except for the AUC rank-based metrics, correlations among the same family of metrics are not as high as expected (Ferri et al., 2009).

A significant result for NILM research is that the differences between the performance metrics tend to increase with the imbalance in the data. This difference is reflected in the lower pairwise correlations observed in imbalanced problems (Ferri et al., 2009). However, it is also shown that Precision is less sensitive to dataset imbalance, mostly due to its invariance to an increase in the positive cases (TP and FN) (Sokolova & Lapalme, 2009). In contrast, the value of Recall will increase with the number of TPs regardless of an increase in the amount of FPs. Thus, in imbalanced problems like NILN, the F-measure is expected to be biased towards positives (Flach, 2003).

Ultimately, it becomes evident that different performance metrics yield different trade-offs that are more or less appropriate based on the studied problem. Therefore, the correct choice of which metrics to optimize or assess a model's performance does matter and should be studied in the context of each machine learning problem. Yet, as seen in the introduction, to date the research in performance metrics for NILM is scarce and only a few works can be found in the published literature (Anderson, Bergés, et al., 2012; Makonin & Popowich, 2017; Mayhorn et al., 2016; Pereira & Nunes, 2017).

In Anderson, Bergés, et al. (2012) the authors propose four performance metrics for event detection algorithm, two of which are domain-specific. An empirical exploration using the BLUED dataset (Anderson, Ocneanu, et al., 2012) was also conducted. As expected, it was found that the optimal event detector, according to one metric, did not necessarily perform well according to other metrics. Nevertheless, the key takeaway message of this work is the importance of including the effects of power in the event detection metrics since electric appliances (and consequently power events) are not equally important in terms of energy consumption.

In Makonin and Popowich (2017) an empirical analysis of six metrics (three for classification and three for energy estimation) was performed using 10-fold cross-validation in one year of data from the AMPds dataset (Makonin, Ellert, et al., 2016). In this work, the authors show that combining classification and energy estimation in the same metric hides essential details about the algorithms' underlying performance. Instead, it is suggested that both treads are considered but using separate metrics. In either case, using one or the other will not provide enough details about the performance.

In Mayhorn et al. (2016), the authors followed an analytical approach to study 10 energy estimation metrics. More concretely, each metric's individual behavior was evaluated in three different tests against 12 NILM error scenarios (i.e., in each scenario, the algorithm was known to produce a specific error). The study finds that the metrics energy error, energy accuracy, and match rate are best suited. The key message of the paper is that regardless of the selected metrics, the difference in energy demand across the data sets requires that these metrics be carefully examined. Otherwise, wrong conclusions are likely to be drawn regarding the performance of the underlying NILM algorithms.

Finally, in Pereira and Nunes (2017) the authors analyzed the behavior of 18 performance metrics empirically when applied to classification algorithms in event-based NILM. More concretely, six classification algorithms were tested across 11 datasets, and the performance metrics compared using correlation analysis and hierarchical clustering. The obtained results show that when applied to event classification, the metrics based on the confusion matrix exhibit correlation among themselves (< 0.5), in line with what happens in other machine-learning domains like medical diagnosis and face recognition as seen above. The authors also find that probabilistic measures can provide information that is not available when using more traditional metrics since they report performance, taking into consideration not just the number of errors but also the distance to the correct values.

As for this paper, it addresses the event detection step which was only looked at by Anderson, Bergés et al. (2012). The focus of this work is to gain an in-depth understanding of how performance metrics behave when used to evaluate this problem and how this compares to the results already present in the related literature. By filling this crucial gap in the current body of knowledge, we expect to contribute to the ongoing efforts towards defining the sets of metrics that are more adequate to assess the performance of the many algorithms that

constitute a NILM solution.

1.3. Document organization

The remaining of this paper is organized as follows: Section 2 presents and describes the algorithms, datasets, and performance metrics that form the basis of this work. Section 3 thoroughly describes the experimental design. Section 4 presents the results and the respective discussion. Section 5 provides an overview of the research implications of this work and outlines future work directions. Finally, the paper concludes in Section 6.

2. Algorithms, datasets and performance metrics

This section presents and describes the different event detection algorithms, datasets, and performance metrics used in this work.

2.1. Algorithms

On an event-based NILM pipeline, event detection is the process of detecting the changes in the power load that are assumed to happen in response to appliances changing their mode of operation. This subsection provides a brief explanation of the event detection algorithms that were used in this work (listed in Table 1). For additional details, please refer to Pereira (2016), and the original references provided for each algorithm.

2.1.1. Expert heuristic detector

The expert heuristic event detector is a modified version of the algorithm presented in Meehan, McArdle, and Daniels (2014). The original algorithm is based on a sliding window that identifies changes in the root mean square (RMS) of the power signal. Power events are triggered when the following conditions are met: (i) the absolute amplitude of the RMS in the second under test is higher by a threshold value than the current RMS 4 s before, and (ii) the previous event did not happen in the last 3 s.

The original version of this algorithm was developed to perform event detection at low sample rates (≤ 1 Hz), and its parameter space Ψ is comprised of three parameters: a power threshold P_{thr}, the number of seconds before the second under evaluation G_{pre} (set initially to 4 s), and the minimum elapsed time between events T_{elap} (set initially to 3 s).

With the objective of supporting datasets with higher sampling rates (> 1 Hz), the original algorithm was extended with three additional parameters. More precisely, *pre-* and *post-event* window lengths, (ω_{pre} and ω_{post}), and an event edge E_{edge}. The *pre-* and *post-event* window lengths are used to set the number of samples to average in order to find the difference in amplitude between different instants in time (Δ_{pwr}). The event edge parameter is used to enable the evaluation of the obtained results against the ground-truth data. In other words, the event edge is the sample index inside the second where the event occurred, e.g., an event edge of zero means that the event happened in the first sample of that second (assuming zero-indexing).

The algorithm works as follows: In the first step, for each power sample, the amount of power change Δ_{pwr} is calculated by subtracting the average power before, from the average power after that sample. In the second step, power changes with a Δ_{pwr} above the predefined threshold P_{thr} (in absolute value) are flagged as possible power events. Finally, in the third step, the flagged power events that are separated by at least T_{elap} seconds are confirmed as events, whereas the others are discarded.

Table 1
Event detection algorithms under evaluation.

Algorithm	Symbol
Expert Heuristic Detector	*EHD*
Log Likelihood Ratio Detector with Voting	LLD_{Vote}
Simplified Log Likelihood Ratio Detector with Maxima	$SLLD_{Max}$
Log Likelihood Ratio Detector with Maxima	LLD_{Max}
Simplified Log Likelihood Ratio Detector with Voting	$SLLD_{Vote}$

2.1.2. Probabilistic detectors

The probabilistic detectors used in this work are extensions of the Generalized Likelihood Ratio event detector (*GLR*) presented in Luo, Norford, Leeb, and Shaw (2002), in the sense that they rely on the log-likelihood ratio test to calculate the likelihood of a potential change in the mean value of two sequential windows (*pre-* and *post-event* windows, respectively).

However, unlike the *GLR* algorithm that sets a threshold to the detection statistic (*DS*) in order to find power events, (i.e., a power event is signaled whenever a pre-defined threshold in the *DS* is reached), the *LLD* (Anderson, Bergés, et al., 2012; Berges et al., 2011) and *SLLD* (Pereira, 2017a; Pereira, Quintal, Gonc calves, & Nunes, 2014) event detectors rely on specific algorithms to extract the power events from the detection statistics signal. To state more concretely, each of the probabilistic event detectors used in this work consist of two separate algorithms. The first referred to as *Detection Statistic*, is used to calculate the power event likelihood. The second referred to as *Detection Activation*, is used to extract the power events from the signal generated by the former.

2.1.2.1. Detection statistics. In the case of the *LLD*, the detection statistic is given by Eq. (1), where $\mu_{0,n}$, $\sigma^2_{0,n}$, $\mu_{1,n}$ and $\sigma^2_{1,n}$ are the sample mean and variance of the pre- and post-event windows, respectively. $P(x)$ is the power at the xth sample.

$$\mathrm{ds}(x) = \ln\left(\frac{\sigma_{0,n}}{\sigma_{1,n}}\right) + \frac{(P(x)-\mu_{0,n})^2}{2\times\sigma^2_{0,n}} - \frac{(P(x)-\mu_{1,n})^2}{2\times\sigma^2_{1,n}} \tag{1}$$

With respect to the *SLLD*, the detection statistic is given by Eq. (2), where $\mu_{0,n}$ and $\mu_{1,n}$ are the mean of the pre- and post-event windows respectively, σ^2_n is the variance of the detection window, and $P(x)$ is the power of the xth sample.

$$\mathrm{ds}(x) = \frac{\mu_{1,n}-\mu_{0,n}}{\sigma^2_n} \times -\left|P(x) - \frac{\mu_{0,n}+\mu_{1,n}}{2}\right| \tag{2}$$

2.1.2.2. Detection activation. Two detection algorithms were developed. The first one is a voting algorithm that works by sliding a voting window (w_V) across the log-likelihood $l[n]$ and assigning a vote to the sample with the higher absolute magnitude at each shift of the window. Then, for each sample in the log-likelihood, the votes are accumulated, and the samples with a number of votes greater than a voting threshold (V_{thr}) are signaled as being power events. This algorithm was first presented in Berges (2010)

The second algorithm works by sliding a window (w_M) across the absolute value of the log-likelihood $l[n]$ looking for the local maxima. The length of the window is equal to twice the maxima precision plus one ($2M_{pre}+1$). For each shift of the sliding window, the sample in the middle is signaled as a power event if its absolute value is larger than the absolute value of all the M_{Pre} samples to the left and right.

2.2. Datasets

At the time of this research, BLUED (Anderson, Ocneanu, et al., 2012) was the only dataset that contained labeled power events by default. The BLUED dataset, released in 2012, consists of one week of whole-house current and voltage measurements (at 12 kHz) and real and reactive power (at 60 Hz) from one house in the state of Pennsylvania, USA. For each power event with an absolute power change of 30 W, the timestamp and responsible appliance name are also provided. A total of 43 loads were considered, 34 of which have labeled power

Table 2
Summary of the active power change and elapsed time between power events in the event detection datasets.

Dataset	P.E.	Power change (W)				Elapsed time (S)			
		Mean	25%	50%	75%	Mean	25%	50%	75%
UK-DALE H1	5440	268	48	100	273	111	4	7	28
UK-DALE H2	2842	365	45	74	137	212	6	15	172
BLUED PA	887	274	84	116	582	690	18	294	892
BLUED PB	1562	351	40	170	428	383	7	35	83

events.

Consequently, to proceed with this work, it was necessary to identify the power events of additional datasets. More concretely, we identified the transitions of one week of data from House 1 (23/06/2013 to 29/06/2013) and House 2 (27/05/2013 to 02/06/2013) of the UK-DALE dataset (Kelly & Knottenbelt, 2015). The UK-DALE dataset, released in 2014, is a record of electric energy consumption from five homes in the United Kingdom. Overall, the dataset contains aggregated and appliance-level active power, measured every 6 s. In three houses, aggregated current and voltage are available at 16 kHz, as well as real power, reactive power, and voltage at 1 Hz. Disaggregated consumption is made available for 54 and 20 appliances in House 1 and House 2, respectively.

The power changes were identified and labeled following the semi-automatic labeling approach presented in Pereira and Nunes (2015a). Furthermore, since only power events with a minimum absolute power change of 30 W were considered in BLUED, the same was done for UK-DALE. Note that it is essential to keep this consistent across datasets. Otherwise, BLUED would have an advantage since none of the missed transitions below ± 30 W would be considered False Negatives.

Ultimately, this resulted in four scenarios from two datasets. In each scenario, the sampling rate is that of the line frequency, i.e., 60 Hz in BLUED and 50 Hz in UK-DALE. The datasets were then converted to the Energy Monitoring and Disaggregation Data Format (EMD-DF) (Pereira, 2017b; Pereira, Nunes, & Bergés, 2014), thus avoiding the need to write different code to interface with each dataset.

Table 2 summarizes the datasets used in this experiment. **P.E.** is the number of power events in the dataset, **Power Change (W)** is a summary of the distribution of the power events in terms of mean, and the 25%, 50%, and 75% percentiles, and **Elapsed Time (S)** is a summary of the difference in time between the power events in the same terms as the Power Change column. Fig. 1 shows a graphical representation of the information in Table 2. It can be observed that the minimum absolute power change is the same in each dataset (30 W). As for the number of seconds between power events, it is possible to see that in BLUED A, the power events are much more spread than in the other cases.

2.3. Performance metrics

A characteristic of residential power data is that appliance activity is sparsely distributed. As a consequence, for an event detector with reasonable performance, it is expected that the number of true negatives (TN) will be much higher than the number of true positives (TP), false positives (FP) and false negatives (FN) (Anderson, Bergés, et al., 2012). This effect is not exclusive to the NILM problem. It is, for example, common in information retrieval problems where the number of irrelevant documents than can be returned after a specific query is much higher than the number of actual relevant items. As such, it is not unexpected that NILM researchers have adapted performance metrics used in the information retrieval domain to evaluate event detection algorithms, like for example, precision and recall (Kagolovsky & Moehr, 2003).

This work considers these and other confusion matrix and rank-based metrics that are commonly used to evaluate this class of problems. Furthermore, this work also explores performance metrics specifically created for event detection problems (Anderson, Bergés, et al., 2012), i.e., Domain-Specific Metrics. The different performance metrics are briefly described next. Please refer to Appendix A for the respective equations.

2.3.1. Confusion matrix based metrics

As the name suggests, confusion matrix based metrics are derived from the values in the confusion matrix. Table 3 lists the confusion matrix based performance metrics that were used, where **Best** and **Worst** refer to the best and worst values that each metric can report.

Table 3
Summary of confusion matrix based metrics.

Metric	Symbol	Best	Worst
Accuracy	A	1	0
Error-rate	E	0	1
Precision	P	1	0
Recall	R	1	0
False Positive Rate	FPR	0	1
False Positive Percentage	FPP	0	[a]
$F_{0.5}$-Measure	$F_{0.5}$	1	0
F_1-Measure	F_1	1	0
F_2-Measure	F_2	1	0
Standardized MCC	$SMCC$	1	0
Distance to Perfect Score P-R	DPS_{PR}	0	2
Distance to Perfect Score TPR-FPR	DPS_{Rate}	0	[a]
Distance to Perfect Score TPP-FPP	DPS_{Perc}	0	[a]

[a] Since the FPP metric can return a value greater than one it is not possible to define a lower bound.

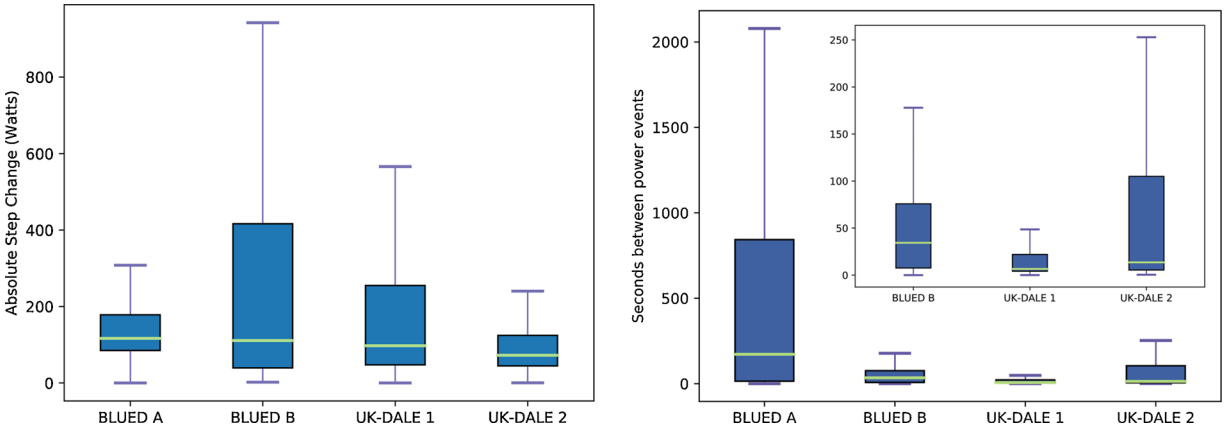

Fig. 1. Summary of dataset characteristics: absolute step change (left); seconds between power events (right).

Table 4
Summary of rank based metrics.

Metric	Symbol	Best	Worst
Wilcoxon statistics based AUC	*WAUC*	1	0
Wilcoxon statistics based AUC Balanced	*WAUCB*	1	0
Geometric Mean AUC	*GAUC*	1	0
Biased AUC	*BAUC*	1	0

2.3.2. Rank based metrics

Rank (or ordering) metrics can be thought of as summaries of the performance of a learned model over varying decision criteria. One of such measures is the area under the ROC curve (AUC), which is drawn by changing the discrimination threshold of a classifier and calculated using the trapezoidal rule.

However, for discrete algorithms where fixed labels are produced (the case of event detection), the AUC should not be measured by employing the trapezoidal rule since the eventual presence of outliers could lead to distorted results (Iba, Hasegawa, & Paul, 2009). Instead, the non-parametric Wilcoxon statistic should is used, as suggested by Hanley and McNeil (1982).

Table 4 summarizes the rank metrics used in this work. A more detailed explanation of each metric is provided in Iba et al. (2009).

2.3.3. Domain-specific metrics

Domain-specific metrics for event detection were introduced in Anderson, Bergés et al. (2012) motivated by the fact that performance metrics based solely on the confusion matrix implicitly assume that all power events are of equal importance. An assumption that, as argued by the authors, is not fair since different appliances have different consumption levels and consequently, more or less weight in the final energy estimation. Table 5 summarizes the domain specific metrics under study in this work.

3. Experimental design

This section thoroughly describes the experimental design. It starts with a description of the training and testing procedures, which is followed by an explanation of how the different performance metrics were calculated. Next, the processes for creating the pairwise correlation matrices and the performance metrics clusters are described.

3.1. Training and testing

To get event detection results, a controlled parameter sweep was conducted for each of the five event detection algorithms. A parameter sweep refers to a controlled variation of some parameters in a particular algorithm (i.e., structural changes) and provides insights into how the different parameters affect the final results.

Ultimately, the parameter sweep returned 47,950 distinct event detection models across the five algorithms. Each model was tested against the four event detection scenarios for a total of 109,800 model-dataset pairs, as summarized in Table 6. It was decided to fix the power threshold (P_{thr}) to 30 W, since this was the minimum power change for which there were labeled events. The remaining parameters of each algorithm were changed as described next.

Table 5
Summary of domain specific metrics.

Metric	Symbol	Best	Worst
Total Power Change – False Positives	TPC_{FP}	0	[a]
Total Power Change – False Negatives	TPC_{FN}	0	[a]
Average Power Change – False Positives	APC_{FP}	0	[a]
Average Power Change – False Negatives	APC_{FN}	0	[a]
Distance to Perfect Score TPC	DPS_{TPC}	0	[b]
Distance to Perfect Score APC	DPS_{APC}	0	[b]

[a] The worst result is proportional to the number of events and size of the erroneous events; thus it is not possible to define a lower bound.
[b] Since it is not possible to define a lower bound to the individual metrics it is also not possible to set a lower bound to the DPS metric.

Table 6
Number of different models that will be evaluated across datasets.

Algorithm	Individual models	Model-dataset pairs
MEH	4.95 k	19.8 k
LLD_{Max}	1 k	4 k
LLD_{Vote}	11 k (50 Hz); 9.5 k (60 Hz)	220 k; 19 k
$SLLD_{Max}$	1 k	4 k
$SLLD_{Vote}$	11 k (50 Hz); 9.5 k (60 Hz)	220 k; 19 k
	47,950	**109,800**

Table 7
Parameter ranges for the expert heuristic event detector.

Parameter	Min	Max	Increment
G_0	0	5	1 (s)
w_0	1	5	1 (s)
w_1	1	5	1 (s)
T_{elap}	0	5	0.5 (s)
E_{edge}		1, $0.5F_s$, F_s	

3.1.1. MEH

For the *MEH*, it was decided to change all the remaining parameters using the range of values presented in Table 7.

3.1.2. LLD and SLLD with voting activation

The *LLD* and *SLLD* with *voting* activation were tested according to the parameter ranges defined in Table 8. It is important to note that the number of models that result from this parameter sweep varies with the sampling rate of the dataset. For example, in a 60 Hz dataset for each half-a-second increment it is always possible to increment twice the voting threshold by fifteen samples, which is not always true in the case of 50 Hz datasets.

3.1.3. LLD and SLLD with maxima activation

The *LLD* and *SLLD* with *maxima* activation were tested according to the parameter ranges defined in Table 9.

Table 8
Parameter ranges for log-likelihood and simplified log-likelihood detectors with voting activation.

Parameter	Min	Max	Increment
w_0	0.5	5	0.5 (s)
w_1	0.5	5	0.5 (s)
w_v	0	5	0.5 (s)
v_{thr}	5	*	15 (votes)

Table 9
Parameter ranges for log-likelihood and simplified log-likelihood detectors with maxima activation.

Parameter	Min	Max	Increment
w_0	0.5	5	0.5 (s)
w_1	0.5	5	0.5 (s)
M_{pre}	0.5	5	0.5 (s)

3.2. Performance metric calculation

In this step, the 23 performance metrics were calculated for each of the models returned by the parameter sweep. To this end, the number of true positives (*TP*), false positives (*FP*), true negatives (*TN*) and false negatives (*FN*) of each model had to be counted. This was done by comparing the events triggered by each model with the actual events in the corresponding dataset (i.e., the ground-truth).

To accomplish this, a tolerance interval in which the detected events should fall to be considered correct detections was defined. This interval is defined by Eq. (3) and is based on a tolerance value that was added to account for eventual ambiguity when determining exactly where an event occurs during the labeling process (Anderson, 2014).

$$\Omega = [\text{ground\_truth} - \text{tolerance}, \text{ground\_truth} + \text{tolerance}] \quad (3)$$

In previous work on this topic (Anderson, 2014) this parameter was set to vary between one and 6 s, but no significant changes were found for more than 3 s. Consequently, it was decided to set this parameter to range between zero and 3 s with variable steps, as defined by the set τ in (Eq. (4)), where F_s is the sampling rate of the dataset.

$$\tau = \{0, 1, 5, 15, F_s, 1.5 \times F_s, 2 \times F_s, 2.5 \times F_s, 3 \times F_s\} \quad (4)$$

Fig. 2 shown an example of three ground-truth power events and the respective *TP* ranges for the tolerance intervals given by F_s, $2F_s$, and $3F_s$. The figure also presents the events detected by two different algorithms, exemplifying how different the detection positions can be.

Regarding the process of creating the confusion matrix, an algorithm that follows the logic presented in Fig. B.1 of Appendix B had to be developed. Given a list of detected events and another with the ground-truth data, the algorithm works as follows:

1. For each ground-truth event, if there are detections that fall within the interval Ω given by Eq. (3), the event that is closer to the ground-truth position (in absolute distance) or the one that was detected first (in the case of equidistant detections) is considered a *TP*, whereas the others must be compared with the next ground-truth event. Otherwise, if no detections happened within the specified interval, a *FN* is added.
2. The detected events that do not fall within any of the possible intervals Ω (one per each ground-truth event) are considered *FP*.
3. When all the detected and ground-truth events have been processed, the *TN* are calculated by subtracting the *TP*, *FN* and *FP* from the number of samples in the dataset. I.e., all the positions where an event could have happened.

Table 10
Confusion matrices obtained for the event detection results presented in Fig. 2.

	Event Detector 1				Event Detector 2			
	0	F_s	$2F_s$	$3F_s$	0	F_s	$2F_s$	$3F_s$
TP	0	3	3	3	0	2	2	2
FP	3	0	0	0	4	2	2	2
FN	3	0	0	0	3	1	1	1
TN	1754	1757	1757	1757	1753	1755	1755	1755

To demonstrate the algorithm, Table 10 shows the results when the algorithm is applied to the data from Fig. 2. As can be observed, with a zero-tolerance, none of the detected power events is considered a *TP*. Likewise, it can also be observed that in the case of event detector number 2, there are only two *TP*s even though more than one event falls whit-in the tolerance intervals.

3.3. Calculation of the metric pairwise correlations

In this step the linear (*Pearson*) and rank (*Spearman*) pairwise correlations between the 27 performance metrics (including the *TP*, *FP*, *TN* and *FN*) were calculated. This returned a total of $(27 \times 26)/2 = 351$ unique pairwise correlations per coefficient.

It is important to remark that unlike the *Pearson* coefficient that is based on the metric values, the *Spearman* coefficient is based on the ranking obtained from the metric value. In this work, the ranks were calculated according to the *competition ranking strategy*, in which metrics with equal value receive the same ranking number and a gap is left in the rank values, thus guarantying that ties do not modify the ranks given to the remaining metrics.

The selection of this ranking strategy is essential since we are trying to assess the consistency of the ranks across models and datasets. As such, it is necessary that the range of the rankings (worst to best) remains consistent independently of the number of ties that may happen. This would not happen if, for example, we had chosen the *dense ranking strategy* in which the next element always receives the immediately following ranking number independently of ties.

3.4. Calculation of the average correlation matrices

Since each model is evaluated ten times in each of the four datasets, there was a total of 200 correlation matrices per correlation coefficient (*10 tolerance values* × *5 algorithms* × *4 datasets*). From these 200

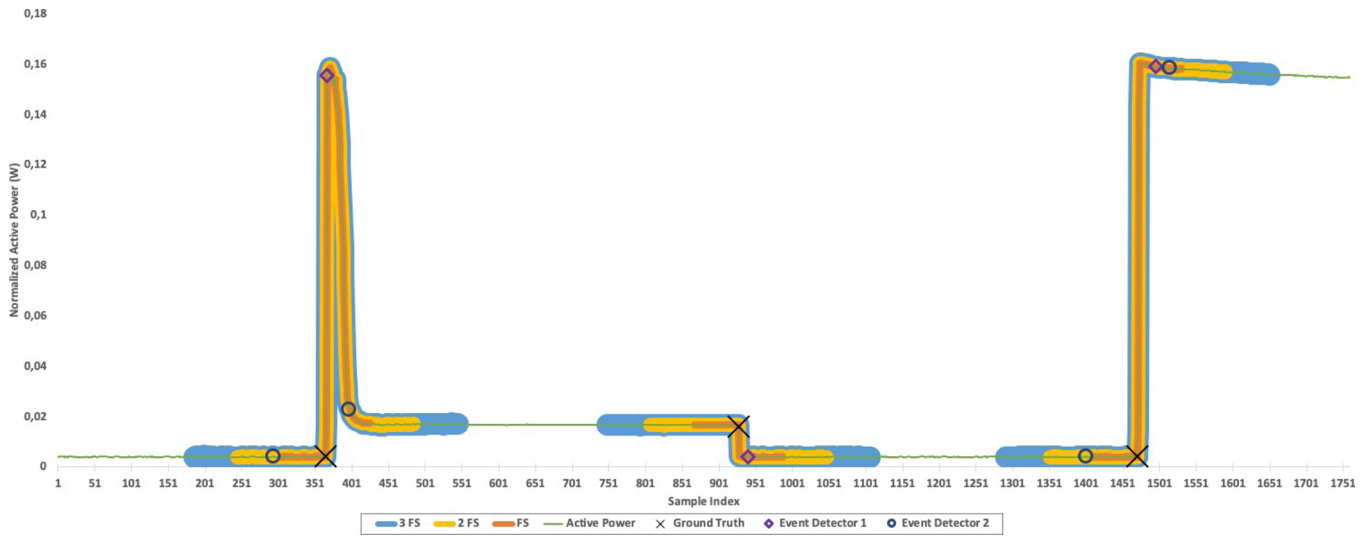

Fig. 2. Event detection example showing three ground-truth events and three of the possible *TP* ranges.

matrices, a cross-dataset correlation matrix was calculated for each coefficient. The process was as follows:

1. The 200 matrices were averaged (arithmetically) based on the algorithm, which resulted in 40 correlation matrices (*10 tolerance values* × *4 datasets*).
2. These 40 matrices are averaged by tolerance to create one matrix per dataset. This resulted in four matrices.
3. The final cross-dataset correlation matrix was calculated by averaging the correlation matrices of each dataset.

It is important to remark that under no circumstances, the evaluation results of the different model-dataset pairs are merged. Instead, only the pairwise metric correlations are averaged, thus avoiding biased conclusions due to the possibility that good results in one dataset can compensate for poor results in others and vice-versa. Furthermore, given the very high number of models (47,950), it is not expected that high pairwise correlations in some models are shadowed by low correlation in others. Another option would be to use the median, but this would report only the central tendency, therefore shadowing the effects of very low and very high pairwise correlation in some values.

3.5. Hierarchical clustering

In this step, the clusters from the resulting cross-dataset correlation matrices were built employing hierarchical clustering. To do so, it was necessary to define the dissimilarity and linkage functions. The former is used to define the distance between two clusters (or metrics), whereas the latter is used to join (i.e., group) the different pairs of metrics and clusters.

Regarding the former, this work used the dissimilarity function defined in Eq. (5), where D is the distance and $|C|$ is the absolute value of the correlation between metric pair.

$$D = 1 - |C| \tag{5}$$

This dissimilarity measure is known to discriminate well between all correlated pairs, independently of the direction, since the pairs with "stronger" correlation are ordered correctly from the bottom ($|C| = 1.0$) to the top ($|C| = 0.0$). Thus this measure is more suitable for graphical representation using dendrograms (Glynn, 2005).

As for the linkage function, the average-group distance was used. This function joins an existing group to the element (or group) whose average distance to the group in minimum. This method is also known as Un-weighted Pair Group Method with Arithmetic Mean (*UPGMA*) and the distance between two groups, A and B, is given by Eq. (6).

$$D_{AB} = \frac{1}{|A||B|} \times \sum_{a \in A} \sum_{b \in B} d(a, b) \tag{6}$$

Where d is a distance function (in our case the Euclidean distance) and $|A|$ and $|B|$ are the size of groups A and B, respectively.

4. Results and discussion

This section presents the obtained results and the respective discussion. This is done in two steps, first, for the individual pairwise correlations, and after this, for the metrics clusters that emerge from the correlation matrices.

4.1. Pairwise correlations

The average pairwise correlations across the four event detection scenarios are presented in Fig. 3 with the rank and linear correlations in the lower and upper triangles, respectively. The average correlations per metric are also presented in Fig. 4. A heat map was used to color-code the pairwise correlations, from green (higher) to red (lower).

For additional information, the respective standard deviation values are also made available in Appendix C. A different heat map was used to color-code the standard deviation values, from green (lower) to red (higher).

When examining the correlation results shown in Fig. 3, a first observation is that the Average Power Change (*APC*) metrics do not correlate well with any of the other metrics. Furthermore, considering that APC_{FP} and APC_{FN} are just the TPC_{FP} and TPC_{FN} metrics normalized by the number of events it is possible to conclude that *APC* metrics will evidence substantial variations depending on the number and the amplitude of the power events. For example, if an event detector A fails to detect (*FN*) all the power events under 50 W but only fails to detect one power event of 100 W, it will still have a APC_{FN} of about 50 W. On the other hand, an event detector B that only misses one event with 100 W will have a APC_{FN} of 100 W. Hence, under this metric algorithm, A would be considered better than B, even though it misses more events.

Another general observation concerns to the relatively strong correlation (> 0.5 in absolute value) between most of the other metrics in Fig. 4. The only exceptions to this trend are F_1, F_2, DPS_{PR} and *SMCC* with an average correlation of only 0.46. This is particularly interesting since three out of the four metrics were designed to balance Precision and Recall (F_1, F_2 and DPS_{PR}), and still they do not correlate well with their "parent" metrics. For example, the F_2 metric does not have any pairwise correlation above 0.65, and perhaps even more surprising, it does not correlate (< 0.5) with either P or R. This result contrasts the findings in Pereira and Nunes (2017), in which for event classification, all these metrics appear highly correlated among themselves (≥ 0.9).

The results in Fig. 3 also reveal that all the four rank-based metrics have robust correlations between themselves (0.99), as well as with R (0.99). However, this is just a reflection of the fact that the *Specificity* (or True Negative Rate – *TNR* –) is always close to 1 since the number of *TN* is much higher than the amount of *FP*. Consequently, the *AUC* metrics are only reflecting variations in R. Moreover, it is possible to observe that the DPS_{Rate} metric is also very well correlated with the *AUC* metrics in both coefficients. In this case, this is a reflection of the fact that the *FPR* is always close to 0, meaning that it is also fully controlled by R. Again, this is a result that contrasts those obtained for event classification, where AUC metrics appear strongly correlated among themselves (≥ 0.94) and all the remaining performance metrics (≥ 0.83).

A more specific observation concerns the firm (0.92) pairwise rank and linear correlations between TPC_{FN} and R. A possible explanation for this is that most of the missed power events (*FN*) have a similar power change value (possibly near to the minimum power threshold), hence the strong linear and non-linear correlations. Similarly, if we consider the TPC_{FP}, it is possible to observe a relatively strong correlation (0.63) in the non-linear coefficient that is not followed by a strong linear relationship. Hence, it is expected that some TPC_{FP} ranks will be relatively close to those obtained with the other metrics, in particular, those derived from the *FP*. Still, the lack of a robust linear correlation is a good indicator that the amount of power change of the FPs is heavily dependent on the dataset characteristics and not so much on the algorithm configuration (i.e., parameters setup).

4.2. Metric clusters

Overall, it is possible to find 44 metric pairs where at least one of the coefficients is above 0.9 in absolute value. These are summarized in Fig. 5, where it is possible to identify three groups that cover 15 of the 23 studied metrics:

1. R, *WAUC*, TPC_{FN}, DPS_{Rate}
2. *FPR*, *FPP*, DPS_{Perc}, A, E
3. P, $F_{0.5}$, F_1, *SMCC*, DPS_{PR}

Additionally, it is possible to observe that the metrics in the first group are all correlated with the *TP* and *FN*, whereas the metrics in the second

Linear

Non Linear	TP	FP	TN	FN	FPP	FPR	A	E	P	R	F05	F1	F2	SMCC	DPSpr	DPSrate	DPSperc	WAUC	WAUCB	GAUC	BAUC	TPCfn	TPCfp	DPStpc	APCfn	APCfp	DPSapc
TP		0,55	-0,55	-1,00	0,55	0,55	-0,47	0,47	-0,36	1,00	-0,29	-0,11	0,33	-0,02	0,02	-0,98	0,43	1,00	1,00	1,00	1,00	-0,92	0,49	0,25	-0,28	0,17	-0,22
FP	-0,66		-1,00	-0,55	1,00	1,00	-0,99	0,99	-0,81	0,55	-0,80	-0,74	-0,44	-0,67	0,64	-0,48	0,95	0,55	0,56	0,53	0,54	-0,54	0,85	0,71	-0,30	0,15	-0,22
TN	-0,66	1,00		0,55	-1,00	-1,00	0,99	-0,99	0,81	-0,55	0,80	0,74	0,44	0,67	-0,64	0,48	-0,95	-0,55	-0,56	-0,53	-0,54	0,54	-0,85	-0,71	0,30	-0,15	0,22
FN	1,00	-0,66	-0,66		-0,55	-0,55	0,47	-0,47	0,36	-1,00	0,29	0,11	-0,33	0,02	-0,02	0,98	-0,43	-1,00	-1,00	-1,00	-1,00	0,92	-0,49	-0,25	0,28	-0,17	0,22
FPP	-0,66	1,00	1,00	-0,66		1,00	-0,99	0,99	-0,81	0,55	-0,80	-0,74	-0,44	-0,67	0,64	-0,48	0,95	0,55	0,56	0,53	0,54	-0,54	0,85	0,71	-0,30	0,15	-0,22
FPR	-0,66	1,00	1,00	-0,66	1,00		-0,99	0,99	-0,81	0,55	-0,80	-0,74	-0,44	-0,67	0,64	-0,48	0,95	0,55	0,56	0,53	0,54	-0,54	0,85	0,71	-0,30	0,15	-0,22
A	-0,46	0,93	0,93	-0,46	0,93	0,93		-1,00	0,81	-0,47	0,82	0,78	0,51	0,71	-0,68	0,40	-0,96	-0,47	-0,48	-0,46	-0,47	0,47	-0,85	-0,73	0,28	-0,15	0,20
E	-0,46	0,93	0,93	-0,46	0,93	0,93	1,00		-0,81	0,47	-0,82	-0,78	-0,51	-0,71	0,68	-0,40	0,96	0,47	0,48	0,46	0,47	-0,47	0,85	0,73	-0,28	0,15	-0,20
P	-0,37	0,85	0,85	-0,37	0,85	0,85	0,84	0,84		-0,36	0,99	0,93	0,60	0,89	-0,86	0,29	-0,68	-0,35	-0,37	-0,34	-0,35	0,39	-0,65	-0,50	0,28	-0,12	0,22
R	1,00	-0,66	-0,66	1,00	-0,66	-0,66	-0,46	-0,46	-0,37		-0,29	-0,11	0,33	-0,02	0,02	-0,98	0,43	1,00	1,00	1,00	1,00	-0,92	0,49	0,25	-0,28	0,17	-0,22
F05	-0,27	0,80	0,80	-0,27	0,80	0,80	0,85	0,85	0,98	-0,27		0,96	0,67	0,93	-0,91	0,22	-0,69	-0,29	-0,30	-0,28	-0,29	0,33	-0,65	-0,52	0,25	-0,14	0,19
F1	-0,05	0,62	0,62	-0,05	0,62	0,62	0,77	0,77	0,86	-0,05	0,92		0,83	0,99	-0,97	0,04	-0,67	-0,11	-0,13	-0,10	-0,11	0,16	-0,61	-0,55	0,16	-0,17	0,09
F2	0,38	0,21	0,21	0,38	0,21	0,21	0,42	0,42	0,49	0,38	0,59	0,79		0,87	-0,85	-0,39	-0,45	0,33	0,32	0,34	0,33	-0,26	-0,36	-0,44	-0,05	-0,14	-0,11
SMCC	0,02	0,57	0,57	0,02	0,57	0,57	0,72	0,72	0,82	0,02	0,90	0,99	0,83		-0,98	-0,05	-0,60	-0,02	-0,03	-0,01	-0,02	0,07	-0,54	-0,50	0,12	-0,17	0,05
DPSpr	0,01	0,56	0,56	0,01	0,56	0,56	0,71	0,71	0,81	0,01	0,88	0,98	0,81	0,99		0,05	0,58	0,02	0,03	0,01	0,02	-0,07	0,52	0,48	-0,11	0,21	-0,03
DPSrate	1,00	-0,65	-0,65	1,00	-0,65	-0,65	-0,46	-0,46	-0,36	1,00	-0,27	-0,04	0,38	0,02	0,01		-0,37	-0,98	-0,97	-0,98	-0,98	0,89	-0,42	-0,19	0,25	-0,12	0,22
DPSperc	-0,48	0,91	0,91	-0,48	0,91	0,91	0,99	0,99	0,81	-0,48	0,82	0,75	0,40	0,70	0,70	-0,47		0,43	0,44	0,42	0,43	-0,42	0,82	0,75	-0,24	0,15	-0,17
WAUC	1,00	-0,65	-0,65	1,00	-0,65	-0,65	-0,46	-0,46	-0,36	1,00	-0,27	-0,04	0,38	0,02	0,01	1,00	-0,47		1,00	1,00	1,00	-0,92	0,49	0,25	-0,28	0,17	-0,22
WAUCB	1,00	-0,67	-0,67	1,00	-0,67	-0,67	-0,47	-0,47	-0,37	1,00	-0,28	-0,05	0,37	0,01	0,00	1,00	-0,49	1,00		0,99	1,00	-0,92	0,50	0,27	-0,29	0,18	-0,22
GAUC	1,00	-0,65	-0,65	1,00	-0,65	-0,65	-0,46	-0,46	-0,36	1,00	-0,27	-0,04	0,38	0,02	0,01	1,00	-0,48	1,00	1,00		1,00	-0,91	0,48	0,24	-0,28	0,17	-0,22
BAUC	1,00	-0,65	-0,65	1,00	-0,65	-0,65	-0,46	-0,46	-0,36	1,00	-0,27	-0,04	0,38	0,02	0,01	1,00	-0,47	1,00	1,00	1,00		-0,92	0,49	0,25	-0,28	0,17	-0,22
TPCfn	0,92	-0,70	-0,70	0,92	-0,70	-0,70	-0,54	-0,54	-0,44	0,92	-0,35	-0,14	0,28	-0,08	-0,09	0,92	-0,55	0,92	0,92	0,92	0,92		-0,48	-0,23	0,48	-0,16	0,37
TPCfp	-0,64	0,86	0,86	-0,64	0,86	0,86	0,79	0,79	0,74	-0,64	0,69	0,54	0,16	0,50	0,49	-0,63	0,78	-0,63	-0,64	-0,63	-0,63	-0,66		0,89	-0,31	0,49	-0,08
DPStpc	0,04	0,27	0,27	0,04	0,27	0,27	0,43	0,43	0,28	0,04	0,33	0,45	0,51	0,44	0,41	0,04	0,42	0,04	0,03	0,04	0,04	0,07	0,39		-0,14	0,40	0,04
APCfn	0,24	-0,35	-0,35	0,24	-0,35	-0,35	-0,36	-0,36	-0,28	0,24	-0,26	-0,17	0,03	-0,13	-0,13	0,24	-0,34	0,24	0,24	0,24	0,24	0,44	-0,31	-0,08		-0,10	0,77
APCfp	-0,17	-0,04	-0,04	-0,17	-0,04	-0,04	-0,03	-0,03	-0,05	-0,17	-0,02	0,02	0,02	0,01	0,06	-0,17	0,00	-0,17	-0,17	-0,17	-0,17	-0,15	0,32	0,39	-0,10		0,32
DPSapc	0,16	-0,30	-0,30	0,16	-0,30	-0,30	-0,29	-0,29	-0,25	0,16	-0,23	-0,12	0,06	-0,09	-0,07	0,16	-0,27	0,16	0,16	0,16	0,16	0,30	-0,12	0,14	0,77	0,26	

Fig. 3. Non linear (bottom-left triangle) and linear (top-right triangle) correlation results for all four event detection scenarios. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

group show strong correlations with the *TN* and *FP*.

To further understand the possible metric arrangements, we performed a clustering analysis of the correlation values. Fig. 6 shows the dendrograms obtained from the ranks (non-linear) and linear pairwise correlations.

The resulting dendrograms were then split at different cut-off values (i.e., distances) to determine the number of clusters that the metrics would form. More precisely, after visual inspection, the dendrograms were cut at the distances of 0.05 and 0.1. Table 11 lists the obtained clusters.

A first general observation is the fact that linear and rank correlations return very similar clusters. The only difference is the absence of *DPS* metrics in the linear correlation clusters, which is not necessarily surprising given the quadratic nature of these metrics.

Likewise, it is also possible to observe that *A* and *E* appear in the same cluster. Still, it should be stressed that this result has its peculiarities. More concretely, these two metrics report based on the number of *FP* and *TN* alone (see Fig. 5). Therefore, considering the much larger number of *TN* when compared to the number of *FP* cases, two situations will occur:

1. Accuracy and error-rate will always have values very close to 1 and 0, respectively. Hence, making these two metrics of no use to report results.
2. Accuracy and error-rate will report mostly based on the variations in the number of FP. Hence, the considerably strong pairwise correlations with *FPP* and *FPR* (0.92).

Given the high similarity between the linear and non-linear correlation clusters, the following discussion considers only the groups obtained from the non-linear correlations.

A first specific observation is that with a cut-off distance of 0.05, only 8 out of 20 metrics will belong to a cluster. More precisely, only performance metrics with a pairwise correlation of at least 0.99 get clustered together. Correspondingly, with a cut-off distance of 0.1, only performance metrics with a pairwise correlation of at least 0.985 get clustered, and so on until all the metrics are clustered at a maximum distance of around 2.5. For example, with a cut-off distance slightly below 0.25 F_1 joins *SMCC* and DPS_{PR} in cluster 4, and with a cut-off distance of about 0.3 *P* and $F_{0.5}$ are joined in a 5th cluster.

Lastly, it is important to remark that F_2 remains isolated until very late in the clustering process and that the same happens with all the domain-specific metrics except for the TPC_{FN} metric.

5. Research implications and way forward

This section summarizes the findings and significant research implications of this work. The limitations of this work are also presented, as well as an outline of possible research directions.

5.1. Research implications

The main implications of the obtained results are threefold: (i) *uncovering important behaviors of the performance metrics when applied to the event detection problem*, (ii) *highlighting niches for which additional performance metrics should be studied, and new ones created*, and (iii) *implications for energy estimation performance evaluation.*

Regarding the former, this work clearly shows that the extremely unbalanced nature of the problem (towards *TN*) has several implications in the behavior of the performance metrics:

		TP	FP	TN	FN	FPP	FPR	A	E	P	R	F05	F1	F2	SMCC	DPSpr	DPSrate	DPSperc	WAUC	WAUCB	GAUC	BAUC	TPCfn	TPCfp	DPStpc	APCfn	APCfp	DPSapc
Linear	All	0,54	0,66	0,66	0,54	0,66	0,66	0,64	0,64	0,57	0,54	0,55	0,48	0,43	0,42	0,41	0,50	0,59	0,54	0,54	0,53	0,54	0,53	0,59	0,45	0,27	0,19	0,21
	No DSM	0,58	0,72	0,58	0,58	0,71	0,72	0,70	0,70	0,63	0,58	0,61	0,53	0,45	0,48	0,46	0,55	0,64	0,58	0,60	0,58	0,58						
	No Base					0,64	0,64	0,62	0,62	0,57	0,50	0,55	0,49	0,44	0,44	0,43	0,46	0,57	0,50	0,50	0,49	0,49	0,50	0,58	0,45	0,27	0,19	0,21
	No Base or DSM					0,70	0,64	0,69	0,69	0,64	0,53	0,63	0,56	0,51	0,51	0,50	0,40	0,63	0,53	0,54	0,53	0,53						
Non Linear	All	0,55	0,66	0,66	0,55	0,66	0,66	0,62	0,62	0,57	0,55	0,54	0,43	0,37	0,40	0,39	0,55	0,62	0,55	0,55	0,55	0,55	0,57	0,61	0,24	0,27	0,11	0,22
	No DSM	0,61	0,73	0,73	0,61	0,73	0,73	0,68	0,68	0,65	0,61	0,61	0,50	0,43	0,46	0,44	0,60	0,68	0,60	0,59	0,60	0,60						
	No Base					0,63	0,63	0,61	0,56	0,56	0,50	0,54	0,44	0,39	0,42	0,41	0,50	0,60	0,50	0,49	0,50	0,50	0,52	0,58	0,25	0,27	0,12	0,22
	No Base or DSM					0,71	0,71	0,68	0,66	0,65	0,55	0,62	0,52	0,52	0,50	0,48	0,55	0,68	0,55	0,55	0,55	0,55						

Fig. 4. Non linear and linear correlations averaged by metric for the four event detection scenarios.

		Linear	Non-Linear
TP	FN	-1,00	1,00
TP	R	1,00	1,00
TP	DPSrate	-0,97	0,99
TP	WAUC	1,00	0,99
TP	TPCfn	-0,91	0,92
FP	TN	-1,00	1,00
FP	FPR	1,00	1,00
FP	FPP	1,00	0,90
FP	DPSperc	0,95	0,92
FP	A	-0,99	0,92
FP	E	0,99	0,92
TN	FPR	-1,00	1,00
TN	FPP	-1,00	1,00
TN	DPSperc	0,95	0,90
TN	A	-0,99	0,92
TN	E	-0,99	0,92
FN	R	-1,00	1,00
FN	DPSrate	0,97	0,99
FN	WAUC	-1,00	0,99
FN	TPCfn	0,91	0,92
Absolute Average		0,98	0,96

		Linear	Non-Linear
R	WAUC	-1,00	0,99
R	TPCfn	0,91	0,92
R	DPSrate	-0,97	0,99
DPSrate	WAUC	-0,97	1,00
DPSrate	TPCfn	0,89	0,92
WAUC	TPCfn	-0,91	0,92
FPR	FPP	1,00	1,00
FPR	DPSperc	0,95	0,91
FPR	A	-0,99	0,92
FPR	E	0,99	0,92
FPP	DPSperc	0,95	0,91
FPP	A	-0,99	0,92
FPP	E	0,99	0,92
DPSperc	A	-0,95	0,98
DPSperc	E	0,95	0,98
A	E	-1,00	1,00
P	F05	0,99	0,98
P	F1	0,92	0,85
SMCC	F1	0,98	0,98
F05	F1	0,96	0,92
F05	DPSpr	-0,90	0,88
F05	SMCC	0,93	0,89
DPSpr	SMCC	-0,98	0,99
DPSpr	F1	0,96	0,97
Absolute Average		0,96	0,94

Fig. 5. List of metric pairs with pairwise correlations above 0.9 in at least of one the coefficients.

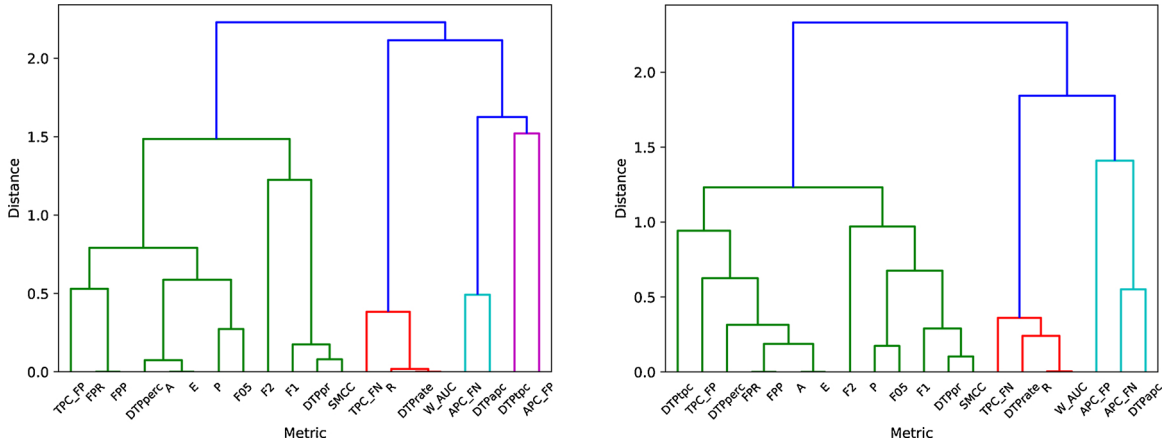

Fig. 6. Dendrograms showing rank (left) and linear (right) correlations of the performance metrics across datasets.

Table 11
Clusters formed after cutting the dendrograms of the cross dataset non-linear and linear correlations.

Distance	Rank	Linear
0.05	1. [*R*, *WAUC*, DPS_{Rate}] 2. [*FPP*, *FPR*] 3. [*A*, *E*]	1. [*R*, *WAUC*] 2. [*FPP*, *FPR*] 3. [*A*, *E*]
0.1	1. [*R*, *WAUC*, DPS_{Rate}] 2. [*FPP*, *FPR*] 3. [*A*, *E*, DPS_{Perc}] 4. [*SMCC*, DPS_{PR}]	1. [*R*, *WAUC*] 2. [*FPP*, *FPR*] 3. [*A*, *E*]

1. Since the number of *TN* is much higher than everything else (i.e., $TN \ggg TP + FP + FN$) the *Accuracy* and $Error - rate$ will always yield very positive results (very close to 1 and 0 respectively). As such, these metrics are of no use in this problem, as it was already suggested in previous literature (Anderson, Bergés, et al., 2012).
2. There is very little correlation between *P*, *R*, and any of the F_β measures. This is particularly evident in the case of F_2, which does not have a strong correlation with any of the studied metrics. Ultimately, this is another reflection of the unbalanced nature of the problem that affects the event detection results. For instance, it is possible to have detectors with very high *P* and very low *R* (i.e., *conservative detectors*) and detectors with very high *R* at the expense of very low *P* (i.e., *liberal detectors*). Consequently, using *P* and *R* alone can lead to extreme conclusions, since in the case of event detection, these two metrics tend to report the performance based solely on minimizing the number of *FP* or maximizing the number of *TP*, respectively.
3. F_1, *SMCC*, and DPS_{PR} have a very low average correlation with the remaining metrics (< 0.5). Still, they are strongly correlated between themselves (≥ 0.97). Hence, it is expected that these metrics will select the same models most of the time. Nevertheless, it is likely that *SMMC* and DPS_{PR} (0.99 pairwise correlation) are less

affected by the unbalance of the data than F_1 since the mathematical formulation of the latter (i.e., the harmonic mean between Precision and Recall), will make it closer to the smaller value. In this case, Precision since the number of FP is likely to be higher than the number of FN.
4. With respect to $F_{0.5}$ and F_2, as expected, the former shows a powerful tendency to ranks the results in the same way as Precision (pairwise rank correlation of 0.98). Of course, this tendency is aggravated by the fact the F_β– Measure tends to the smaller value, as explained in the previous point. As for the latter, in contrast to what should be expected, there is no correlation with Recall. Instead, there are slight correlations with DPS_{PR}, *SMCC*, and F_1, which implies that despite putting twice more importance in Recall, F_2 is still very affected by the Precision score.
5. Since the $FPR \approx 0$ and $TNR \approx 1$, the studied rank metrics are entirely dominated by *Recall* and do not add any new information. I.e., ROC metrics will rank the algorithms in the same way as *Recall*.
6. Regarding the *DSM*, it is clear that these metrics rely heavily on the data under test. This is particularly evident in the case of the *APC* metrics that depend both on the number of power events and respective amplitudes.

The only exception to this is the high correlation (both linear and rank) between the TPC_{FN} and *R*, which indicates that most of the missed events have similar amplitudes (in absolute value).

As for the second research implication, this work also highlights some areas in which other state-of-the-art metrics should be studied and possibly new ones created.

1. This study revealed that the metrics that balance *P* and *R* are profoundly affected by *P* and not so much by *R*. Therefore, it is necessary to study other metrics related to the Precision-Recall curve (*PRC*), which as per the literature, are more suitable to deal with moderate to substantial imbalanced datasets (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). This includes, for example, the Mean Average Precision (*MAP*), which is equivalent to the area under the *PR* (*AUPRC*), and the break-even point (*BEP*), which is the point in the curve where *P* and *R* are the same (Manning, Raghavan, & Schütze, 2008) (chapter 8).
   Another possibility would be to re-define Precision and Recall to account for the time dimension of the event detection problem. One option would be to discard the samples during steady-state operation (i.e., when no appliances are changing the mode of operation (Pereira, Ribeiro, & Nunes, 2017)). Since the majority of the TNs lie in these regions, such an approach would help reduce the unbalanced nature of the problem.
2. This work has also highlighted the potential of *DSM* to unveil important characteristics of the evaluated algorithms (e.g., TPC_{FN} and TPC_{FP}). As such, it is important to conduct further studies using these metrics. Furthermore, future work should aim at defining new *DSM* that take into account the important properties of the NILM problem. For example, since the ON and OFF events are not independent (e.g., for any appliance, the OFF should not come before the ON), it would be important to define metrics that take into consideration the ordering of the events. Likewise, it would be important to have metrics that take into account the complexity of the dataset (e.g., number of appliances, number of simultaneous or near-simultaneous events), and disaggregation information such as the number of missed cycles (i.e., when both ON and OFF are missed).
3. Lastly, it would also be relevant to introduce metrics that take into consideration concepts from cost-sensitive learning (Elkan, 2001). For instance, power events from small appliances may be less relevant than those from larger appliances. Hence they should have a smaller miss-detection/miss-classification cost. On the other hand, rarely used appliances should have a high miss-detection/miss-classification cost, since failing to correctly identify the few occurrences of such appliances will result in a significant underestimation of their consumption.

Finally, concerning the last research implications, this work also sheds some light with respect to performance metrics for the energy estimation problem.

1. Since there are several appliances that are only used for short periods (e.g., Kettles, Toasters), the energy estimation problem is also heavily affected by the dataset imbalance problem. To state more concretely, TNs (i.e., periods with zero consumption) will be much higher than *TP* (i.e., periods where consumption > 0). As such, when using metrics under the event detection category to evaluate the performance of energy estimation algorithms, most of the research implications of this paper should be taken into consideration.
2. Like with the Domain Specific Metric for event detection, energy estimation metrics (e.g., Relative Error (RE), Mean Absolute Error (MAE), and Energy Error (EE)) will be heavily affected by the underlying ground-truth data. For example, a metric like MAE will tend to report higher values in a dataset *A* with more power consumption than a dataset *B*, which does not necessarily mean that the algorithm is performing worst in *A* than in *B*. As such, even though there are several metrics already defined under the energy estimation category (Makonin & Popowich, 2017; Mayhorn et al., 2016; Pereira & Nunes, 2018), it is safe to say that proper cross-dataset benchmarks will require energy estimation metrics that take into consideration the power levels of the different datasets. As for now, the only alternative is to resort to normalized variations of these metrics. E.g., normalizing with respect to the power range, or the total energy.

5.2. Limitations and future work directions

As it was mentioned in 2.2, at the time of this research, only one of the publicly available datasets offered labeled data. As such, and despite it was possible to overcome this issue with two additional weeks of fully labeled data, this process introduces some limitations: First, the two datasets were labeled from ground-truth data that was collected every 6 s, and there was no ground-truth for all the appliances. Second, the same person labeled the two datasets; as such, the labels are subject to the interpretation of one single agent. Consequently, to better generalize the results from this work, future iterations should look at incorporating properly curated datasets, i.e., dataset whose labels have been previously validated by other researchers.

Future work should include different event detection algorithms, in particular, those under the matched filters category. Likewise, another important research direction would be to consider the possibility of finding which sets of metrics are more suitable for each class of loads. For example, such knowledge could be used to define ensembles of event detection algorithms, each one targeting a particular type of load.

More importantly, future iterations of this work should also consider performance metrics for energy estimation. While only a few authors have addressed the energy estimation step in event-based NILM (Giri & Bergés, 2015; He et al., 2018; Zhao et al., 2016), there are many eventless implementations (e.g., HMM and Deep ANNs) (Gomes & Pereira,

2020; Harell et al., 2019; Makonin, Popowich, et al., 2016; Murray et al., 2019) and public datasets (Pereira & Nunes, 2018) that can serve as a basis for this task. Furthermore, despite there is a category of metrics dedicated to energy estimation, it would be of crucial importance to also evaluate the potential of the metrics under the event detection category to assess the performance of energy estimation algorithms.

Given the scarcity of energy estimation algorithms for event-based NILM, another exciting research direction would be to understand if the performance of event detection and classification algorithms can be good predictors of energy estimation performance. For instance, if for each appliance we set all the samples between ON and OFF transitions to one (1) and the remaining to zero (0), it is possible to calculate approximations of metrics such as Precision and Recall.

Another critical research contribution for performance evaluation in NILM would be to study the interconnections between the different algorithms that comprise an end-to-end NILM system by introducing the concept of ceiling analysis (Roncancio, Hernandes, & Becker, 2013). This kind of analysis is associated with the ceiling effect, in which an independent variable no longer affects the dependent variable after reaching a certain level. For example, in the particular case of event-based NILM, this effect can be seen as the impossibility of the event detection step (independent variable) to increase the performance of the event classification step (dependent variable) because it has already reached the “ceiling”.

It is also relevant to mention the existence of other methods that could have been used to study the behavior of the performance metrics. These include the application of different rank correlation coefficients, like Kendall's tau and Spearman's *footrule* (Kendall, 1938; Kumar & Vassilvitskii, 2010; Spearman, 1906), and graphical representation methods such as multidimensional scaling (MDS) and non-linear mapping (NLM) (Borg & Groenen, 2005; Carroll & Arabie, 1980; Kumar & Leone, 1991).

Finally, it is also important to shed some light on the computational complexity of running such models and provide some suggestions that can save on computational resources while still obtaining equivalent results. Note that this discussion focuses on time complexity, which in this case is affected by the duration and number of events in the data set, and the number of tunable parameters in the event detection algorithms.

Concerning the testing data, assuming a constant rate of power events, the time complexity should be linear with the increase in the data set duration. I.e., the time required to run a detection model in two weeks of data would be twice the time to do it for one week. While in the present, computing power may not present a severe issue in many cases, it is possible to considerably reduce this time by avoiding computation across the entire data set. Instead, the computation can be performed in the non-steady state areas defined in advance based on the position of the power events in the ground-truth data. Nevertheless, besides reducing the number of computational operations, it should be noted that this would also considerably reduce the number of TN cases, which would have to be accounted for when calculating the performance metrics.

With respect to the number of tunable parameters of the detection algorithm that can quickly scale the number of possible models to unrealistic numbers, a possible solution would be to first run all the models in sub-set of the data (e.g., one day), and based on the obtained results select the ranges of values that will change in each parameter. This would considerably reduce the number of models, while still providing an excellent estimation of each algorithm's possible outcomes.

6. Conclusion

This paper studied the relationships between 23 performance metrics when used to assess the performance of event detection algorithms. To this end, 47,950 event detection models were executed across four (4) event detection scenarios taken from two public datasets. Linear and non-linear correlations and hierarchical clustering were then used to investigate the existing of any pairwise correlations as well as the formation of clusters of metrics.

Ultimately, this in-depth analysis of the pairwise correlations and the resulting clusters represents an advancement of the state-of-the-art towards defining a consistent set of metrics to evaluate event detection algorithms.

The key take-away messages are:

1. Balancing Precision and Recall with traditional F_β – Measure will be biased towards the former. SMMC and DPS are better metrics, although they still exhibit some bias towards Precision. Therefore, future work must address the lack of metrics to properly balance Precision and Recall in NILM, since the imbalance does not affect only event detection algorithms.
2. ROC based metrics (i.e., based on Recall and FPR) are of very little use in event detection since they are fully controlled by the former.
3. Domain-Specific Metrics are useful to gain institutions about the algorithm performance on a specific dataset. Yet, these should not be used for cross-dataset benchmarks since they are very dataset dependent.
4. It is of crucial importance to define new event detection and energy estimation metrics that take into consideration the imbalanced nature of the problem, as well as the differences in energy across the datasets.

Conflict of interest

The authors declare no conflict of interest.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgments

The authors would like to acknowledge Prof. Mario Bergés, for his valuable insights during the preparation of this work. We would also like to acknowledge Dr. Omid Jahromi, for the fruitful discussion in the sequence of the 4th International Workshop on Non-Intrusive Load Monitoring. This research was supported by the Portuguese Foundation for Science and Technology (FCT) under grants SFRH/DB/77856/2011, CEECIND/01179/2017, and UIDB/50009/2020.

Appendix A. Equations to calculate the performance metrics

Table A.1, Table A.2, Table A.3

Table A.1
Confusion matrix based performance metrics.

Metric	Equation
Accuracy	$A = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}$
Error-rate	$E = \frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \equiv 1 - A$
Precision	$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
Recall	$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
False Positive Rate	$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$
False Positive Perc.	$\mathrm{FPP} = \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FN}}$
F_β - Measure	$F_\beta = (1+\beta^2) \times \frac{P \times R}{(\beta^2 \times P) + R}$
Standardized MCC	$\mathrm{SMCC} = \frac{1+\mathrm{MCC}}{2}$
D. P. Score P-R	$DPS_{PR} = P^2 + R^2 - 2 \times (P+R) + 2$
D. P. Score TPR-FPR	$DPS_{Rate} = TPR^2 + FPR^2 - 2 \times TPR + 1$
D. P. Score TPP-FPP	$DPS_{Perc} = TPP^2 + FPP^2 - 2 \times TPP + 1$
Matthews CC	$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TN}+\mathrm{FN}) \times (\mathrm{TN}+\mathrm{FP}) \times (\mathrm{TP}+\mathrm{FN}) \times (\mathrm{TP}+\mathrm{FP})}}$

Table A.2
Rank based performance metrics.

Metric	Equation
Wilcoxon AUC Wicoxon AUC Balanced	$WAUCB$ $\mathrm{WAUC} = \frac{1}{2} \times (R + \mathrm{TNR})$ $= WAUC \times BF$
Gemetric Mean AUC	$\mathrm{GAUC} = \sqrt[2]{R \times \mathrm{TNR}}$
Biased AUC	$\mathrm{BAUC} = \frac{1}{2} \times (R \times \mathrm{TNR}) + B$
True Negative Rate	$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \equiv$ Specificity
Balancing Factor	$BF = 1 - \lvert R - TNR \rvert$
Bias	$B = \begin{cases} \frac{R}{2}, & \text{if target is the majority class} \\ \frac{\mathrm{TNR}}{2}, & \text{if target is the minority class} \end{cases}$

Table A.3
Domain specific performance metrics.

Metric	Equation
Total Power Change FP	$TPC_{FP} = \Sigma_{f \in FP} \lvert \Delta P_f \rvert$
Total Power Change FN	$TPC_{FN} = \Sigma_{m \in FN} \lvert \Delta P_m \rvert$
Avg. Power Change FP	$\mathrm{APC_{FP}} = \frac{1}{\lvert \mathrm{FP} \rvert} \times \mathrm{TPC_{FP}}$
Avg. Power Change FN	$\mathrm{APC_{FN}} = \frac{1}{\lvert \mathrm{FN} \rvert} \times \mathrm{TPC_{FN}}$

Appendix B. Contingency matrix calculations

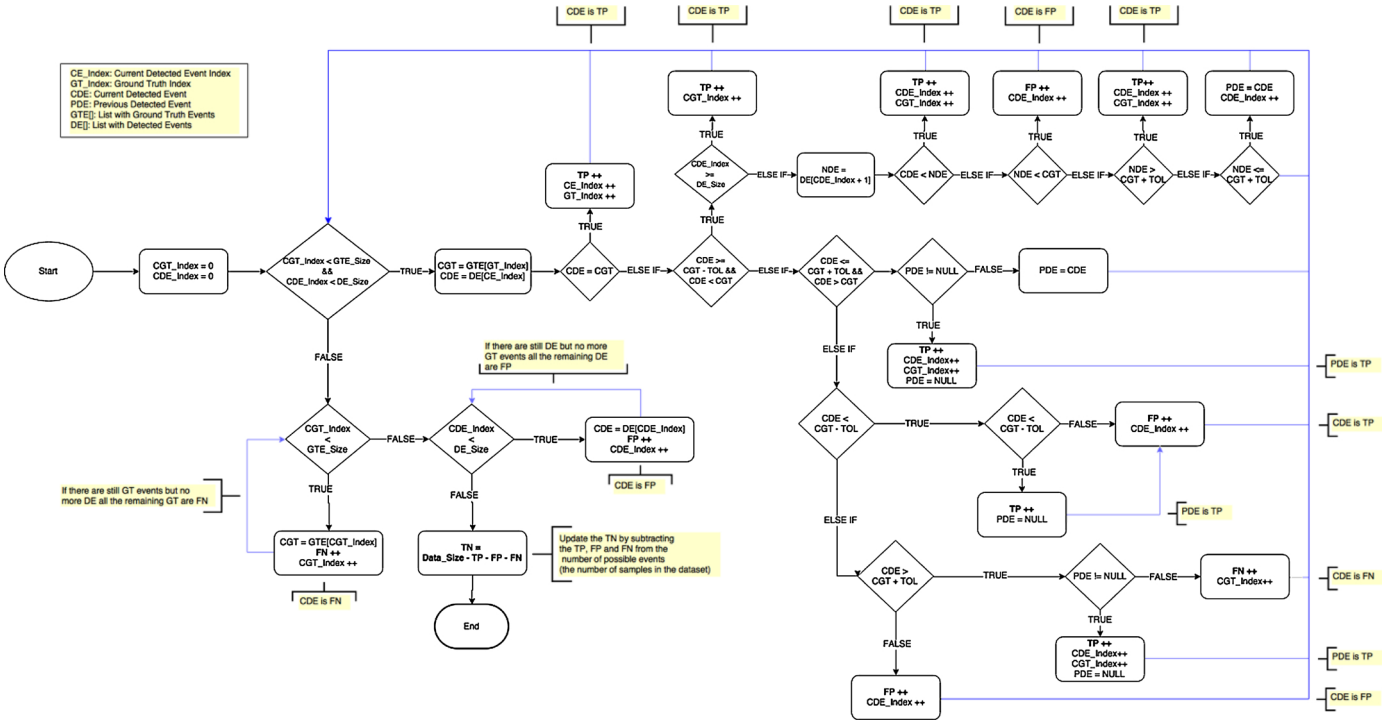

Fig. B.1. Flowchart of the algorithm used to create the contingency table of the event detection algorithms.

Appendix C. Standard deviation for cross-dataset correlation matrices

Fig. C.1, Fig. C.2

Linear

Non Linear	TP	FP	TN	FN	FPP	FPR	A	E	P	R	F05	F1	F2	SMCC	DPSpr	DPSrate	DPSperc	WAUC	WAUCB	GAUC	BAUC	TPCfn	TPCfp	DPStpc	APCfn	APCfp	DPSapc
TP		0,12	0,12	0,00	0,12	0,12	0,13	0,13	0,12	0,00	0,10	0,07	0,19	0,04	0,05	0,01	0,11	0,00	0,00	0,00	0,00	0,08	0,17	0,08	0,66	0,07	0,53
FP	0,10		0,00	0,12	0,00	0,00	0,00	0,00	0,02	0,12	0,02	0,07	0,25	0,09	0,10	0,12	0,01	0,12	0,13	0,12	0,12	0,11	0,13	0,12	0,48	0,07	0,38
TN	0,10	0,00		0,12	0,00	0,00	0,00	0,00	0,02	0,12	0,02	0,07	0,25	0,09	0,10	0,12	0,01	0,12	0,13	0,12	0,12	0,11	0,13	0,12	0,48	0,07	0,38
FN	0,00	0,10	0,10		0,12	0,12	0,13	0,13	0,12	0,00	0,10	0,07	0,19	0,04	0,05	0,01	0,11	0,00	0,00	0,00	0,00	0,08	0,17	0,08	0,66	0,07	0,53
FPP	0,10	0,00	0,00	0,10		0,00	0,00	0,00	0,02	0,12	0,02	0,07	0,25	0,09	0,10	0,12	0,01	0,12	0,13	0,12	0,12	0,11	0,13	0,12	0,48	0,07	0,38
FPR	0,10	0,00	0,00	0,10	0,00		0,00	0,00	0,02	0,12	0,02	0,07	0,25	0,09	0,10	0,12	0,01	0,12	0,13	0,12	0,12	0,11	0,13	0,12	0,48	0,07	0,38
A	0,14	0,06	0,06	0,14	0,06	0,06		0,00	0,02	0,13	0,03	0,05	0,23	0,08	0,09	0,12	0,01	0,13	0,13	0,12	0,13	0,11	0,12	0,12	0,43	0,06	0,34
E	0,14	0,06	0,06	0,14	0,06	0,06	0,00		0,02	0,13	0,03	0,05	0,23	0,08	0,09	0,12	0,01	0,13	0,13	0,12	0,13	0,11	0,12	0,12	0,43	0,06	0,34
P	0,06	0,03	0,03	0,06	0,03	0,03	0,03	0,03		0,12	0,01	0,07	0,26	0,08	0,07	0,12	0,03	0,12	0,12	0,12	0,12	0,13	0,12	0,09	0,44	0,06	0,36
R	0,00	0,10	0,10	0,00	0,10	0,10	0,14	0,14	0,06		0,10	0,07	0,19	0,04	0,05	0,01	0,11	0,00	0,00	0,00	0,00	0,08	0,17	0,08	0,66	0,07	0,53
F05	0,06	0,05	0,05	0,06	0,05	0,05	0,03	0,03	0,02	0,06		0,04	0,21	0,05	0,04	0,09	0,03	0,10	0,10	0,09	0,10	0,11	0,11	0,10	0,40	0,06	0,33
F1	0,14	0,19	0,19	0,14	0,19	0,19	0,10	0,10	0,13	0,14	0,08		0,10	0,01	0,01	0,06	0,03	0,07	0,07	0,07	0,07	0,09	0,07	0,10	0,30	0,06	0,24
F2	0,23	0,34	0,34	0,23	0,34	0,34	0,27	0,27	0,29	0,23	0,25	0,10		0,09	0,08	0,17	0,19	0,19	0,19	0,19	0,19	0,20	0,19	0,06	0,22	0,04	0,19
SMCC	0,12	0,19	0,19	0,12	0,19	0,19	0,12	0,12	0,13	0,12	0,08	0,01	0,10		0,00	0,04	0,05	0,04	0,04	0,05	0,04	0,10	0,07	0,08	0,27	0,06	0,20
DPSpr	0,15	0,20	0,20	0,15	0,20	0,20	0,12	0,12	0,15	0,15	0,09	0,01	0,08	0,00		0,04	0,08	0,05	0,05	0,05	0,05	0,09	0,06	0,08	0,28	0,08	0,21
DPSrate	0,00	0,11	0,11	0,00	0,11	0,11	0,15	0,15	0,07	0,00	0,06	0,14	0,23	0,12	0,15		0,11	0,01	0,01	0,01	0,01	0,08	0,16	0,10	0,61	0,06	0,48
DPSperc	0,15	0,09	0,09	0,15	0,09	0,09	0,01	0,01	0,05	0,15	0,04	0,08	0,22	0,11	0,12	0,15		0,11	0,11	0,11	0,11	0,09	0,13	0,13	0,39	0,07	0,31
WAUC	0,00	0,10	0,10	0,00	0,10	0,10	0,15	0,15	0,07	0,00	0,06	0,14	0,23	0,12	0,15	0,00	0,15		0,00	0,00	0,00	0,08	0,16	0,08	0,66	0,07	0,53
WAUCB	0,00	0,10	0,10	0,00	0,10	0,10	0,14	0,14	0,06	0,00	0,05	0,15	0,24	0,13	0,15	0,00	0,14	0,00		0,00	0,00	0,08	0,17	0,09	0,66	0,07	0,53
GAUC	0,00	0,10	0,10	0,00	0,10	0,10	0,14	0,14	0,07	0,00	0,06	0,14	0,23	0,12	0,15	0,00	0,15	0,00	0,00		0,00	0,08	0,16	0,08	0,65	0,07	0,52
BAUC	0,00	0,10	0,10	0,00	0,10	0,10	0,15	0,15	0,07	0,00	0,06	0,14	0,23	0,12	0,15	0,00	0,15	0,00	0,00	0,00		0,08	0,16	0,08	0,66	0,07	0,53
TPCfn	0,07	0,11	0,11	0,07	0,11	0,11	0,14	0,14	0,10	0,07	0,10	0,13	0,19	0,13	0,14	0,07	0,13	0,07	0,07	0,07	0,07		0,13	0,07	0,68	0,09	0,58
TPCfp	0,10	0,04	0,04	0,10	0,04	0,04	0,03	0,03	0,04	0,10	0,04	0,16	0,30	0,17	0,18	0,10	0,05	0,10	0,10	0,10	0,10	0,10		0,07	0,39	0,12	0,40
DPStpc	0,14	0,14	0,14	0,14	0,14	0,14	0,10	0,10	0,11	0,14	0,08	0,12	0,07	0,09	0,12	0,14	0,13	0,14	0,14	0,14	0,14	0,15	0,17		0,23	0,10	0,23
APCfn	0,65	0,56	0,56	0,65	0,56	0,56	0,41	0,41	0,45	0,65	0,37	0,25	0,23	0,23	0,24	0,64	0,43	0,64	0,65	0,64	0,64	0,69	0,51	0,12		0,17	0,14
APCfp	0,06	0,12	0,12	0,06	0,12	0,12	0,10	0,10	0,12	0,06	0,12	0,14	0,07	0,14	0,15	0,06	0,11	0,06	0,06	0,06	0,06	0,09	0,13	0,10	0,12		0,35
DPSapc	0,55	0,41	0,41	0,55	0,41	0,41	0,27	0,27	0,35	0,55	0,28	0,14	0,22	0,12	0,13	0,55	0,28	0,55	0,54	0,55	0,55	0,59	0,45	0,24	0,13	0,29	

Fig. C.1. Standard deviation values for the non linear (bottom-left triangle) and linear (top-right triangle) correlations, across all four event detection scenarios.

		TP	FP	TN	FN	FPP	FPR	A	E	P	R	F05	F1	F2	SMCC	DPSpr	DPSrate	DPSperc	WAUC	WAUCB	GAUC	BAUC	TPCfn	TPCfp	DPStpc	APCfn	APCfp	DPSapc
Linear	All	0,12	0,11	0,11	0,12	0,11	0,11	0,10	0,10	0,11	0,12	0,09	0,08	0,18	0,07	0,08	0,11	0,09	0,12	0,12	0,11	0,12	0,14	0,15	0,11	0,46	0,08	0,38
	No DSM	0,07	0,08	0,07	0,07	0,08	0,08	0,08	0,08	0,08	0,07	0,06	0,06	0,15	0,06	0,06	0,07	0,07	0,07	0,07	0,07	0,07						
	No Base					0,12	0,12	0,11	0,11	0,11	0,13	0,10	0,08	0,18	0,07	0,08	0,12	0,10	0,13	0,13	0,12	0,13	0,15	0,15	0,11	0,44	0,09	0,37
	No Base or DSM					0,08	0,12	0,08	0,08	0,08	0,07	0,07	0,06	0,19	0,05	0,06	0,09	0,07	0,07	0,08	0,07	0,07						
Non Linear	All	0,12	0,13	0,13	0,12	0,13	0,13	0,12	0,12	0,10	0,12	0,09	0,13	0,23	0,12	0,14	0,12	0,13	0,12	0,12	0,12	0,12	0,15	0,13	0,13	0,46	0,10	0,38
	No DSM	0,08	0,10	0,10	0,08	0,10	0,10	0,10	0,10	0,07	0,08	0,06	0,12	0,24	0,12	0,13	0,08	0,11	0,08	0,08	0,08	0,08						
	No Base					0,15	0,15	0,12	0,12	0,11	0,13	0,09	0,13	0,22	0,12	0,13	0,14	0,13	0,14	0,10	0,13	0,14	0,16	0,14	0,13	0,46	0,11	0,36
	No Base or DSM					0,11	0,11	0,10	0,10	0,08	0,09	0,07	0,11	0,11	0,11	0,12	0,09	0,11	0,09	0,09	0,09	0,09						

Fig. C.2. Standard deviation values of non linear and linear correlations, averaged by metric across the four event detection scenarios.

References

Alcalá, J., Parson, O., & Rogers, A. (2015). Detecting anomalies in activities of daily living of elderly residents via energy disaggregation and cox processes. *Proceedings of the 2nd ACM international conference on embedded systems for energy-efficient built environments,* 225–234.

Alcalá, J., Ureña, J., Hernández, A., & Gualda, D. (2017). Event-based energy disaggregation algorithm for activity monitoring from a single-point sensor. *IEEE Transactions on Instrumentation and Measurement,* 1–12.

Anderson, K., Ocneanu, A., Benitez, D., Carlson, D., Rowe, A., & Berges, M. (2012a). BLUED: A fully labeled public dataset for event-based non-intrusive load monitoring research. *Proceedings of the 2nd KDD workshop on data mining applications in sustainability (SustKDD)*.

Anderson, K. D., Bergés, M. E., Ocneanu, A., Benitez, D., & Moura, J. M. F. (2012b). Event detection for non intrusive load monitoring. *IECON 2012 – 38th annual conference on IEEE industrial electronics society,* 3312–3317.

Anderson, K. (2014). *Non-intrusive load monitoring: Disaggregation of energy by unsupervised power consumption clustering.* Pittsburgh, PA, USA: Carnegie Mellon University PhD.

Armel, K. C., Gupta, A., Shrimali, G., & Albert, A. (2013). Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy Policy, 52*, 213–234.

Attari, S. Z., DeKay, M. L., Davidson, C. I., & de Bruin, W. B. (2010). Public perceptions of energy consumption and savings. *Proceedings of the National Academy of Sciences United States of America*.

Barsim, K. S., & Yang, B. (2015). Toward a semi-supervised non-intrusive load monitoring system for event-based energy disaggregation. *2015 IEEE global conference on signal and information processing (GlobalSIP)* (pp. 58–62).

Batra, N., Kelly, J., Parson, O., Dutta, H., Knottenbelt, W., Rogers, A., Singh, A., & Srivastava, M. (2014). NILMTK: An open source toolkit for non-intrusive load monitoring. *Proceedings of the 5th international conference on future energy systems, e-Energy'14,* 265–276.

Batra, N., Kukunuri, R., Pandey, A., Malakar, R., Kumar, R., Krystalakos, O., Zhong, M., Meira, P., & Parson, O. (2019). *Towards reproducible state-of-the-art energy disaggregation. Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation, BuildSys'19*. New York, NY, USA: ACM193–202.

Beckel, C., Sadamori, L., Santini, S., & Staake, T. (2015). Automated customer segmentation based on smart meter data with temperature and daylight sensitivity. *IEEE international conference on smart grid communications (SmartGridComm),* 653–658.

Belley, C., Gaboury, S., Bouchard, B., & Bouzouane, A. (2013). *Activity recognition in smart homes based on electrical devices identification. Proceedings of the 6th international conference on pervasive technologies related to assistive environments, PETRA'13*. New York, N.Y., U.S.A: ACM7:1–7:8.

Berges, M. (2010). *A framework for enabling energy-aware facilities through minimally-intrusive approaches.* PhD CARNEGIE MELLON UNIVERSITY.

Bergés, M., & Kolter, Z. (2012). *Non-intrusive load monitoring: A review of the state of the art*.

Berges, M., Goldman, E., Matthews, H., Soibelman, L., & Anderson, K. (2011). User-centered nonintrusive electricity load monitoring for residential buildings. *Journal of Computing in Civil Engineering, 25*, 471–480.

Borg, I., & Groenen, P. (2005). *Modern multidimensional scaling – Theory and applicationsSpringer series in statistics* (2nd edition). New York, NY: Springer New York.

Carroll, J. D., & Arabie, P. (1980). Multidimensional scaling. *Annual Review of Psychology, 31*, 607–649.

Caruana, R., & Niculescu-Mizil, A. (2004). Data mining in metric space: An empirical analysis of supervised learning performance criteria. *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, KDD'04* (pp. 69–78). ACM.

Chisik, Y. (2011). *An image of electricity: Towards an understanding of how people perceive electricity. Proceedings of the 13th I.F.I.P. 13 international conference on human–computer interaction, INTERACT'11*. Lisbon Portugal: Springer-Verlag100–117.

Czarnek, N., Morton, K., Collins, L., Newell, R., & Bradbury, K. (2015). Performance comparison framework for energy disaggregation systems. *IEEE international conference on smart grid communications (SmartGridComm),* 446–452.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd international conference on machine learning, ICML'06* (pp. 233–240). ACM.

Egarter, D., Pöchacker, M., & Elmenreich, W. (2015). *Complexity of power draws for load disaggregation.* (cs) arXiv:1501.02954.

Elkan, C. (2001). *The foundations of cost-sensitive learning. Proceedings of the 17th international joint conference on artificial intelligence – Volume 2, IJCAI'01*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc973–978.

Esa, N. F., Abdullah, M. P., & Hassan, M. Y. (2016). A review disaggregation method in non-intrusive appliance load monitoring. *Renewable and Sustainable Energy Reviews, 66*, 163–173.

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters, 30*, 27–38.

Flach, P. A. (2003). *The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. Proceedings of the twentieth international conference on international conference on machine learning, ICML'03*. Washington DC, USA: AAAI Press194–201.

Gajowniczek, K., & Zkabkowski, T. (2017). Electricity forecasting on the individual household level enhanced based on activity patterns. *PLOS ONE, 12*, 26.

Giri, S., & Bergés, M. (2015). An energy estimation framework for event-based methods in non-intrusive load Monitoring. *Energy Conversion and Management, 90*, 488–498.

Glynn, E. F. (2005). *Correlation "Distances" and hierarchical clustering*.

Gomes, E., & Pereira, L. (2020). PB-NILM: Pinball guided deep non-intrusive load monitoring. *IEEE Access, 8*, 48386–48398.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29–36.

Harell, A., Makonin, S., Bajic, I. V., & Wavenilm:, A. (2019). Causal neural network for power disaggregation from the complex power signal. *ICASSP 2019 – 2019 IEEE international conference on acoustics speech and signal processing (ICASSP),* 8335–8339.

Hart, G. (1992). Nonintrusive appliance load monitoring. *Proceedings of the IEEE, 80*, 1870–1891.

Hart, G. (1985). *Prototype nonintrusive appliance load monitor.* Technical Report MIT Energy Laboratory Technical Report, and Electric Power Research Institute Technical Report.

He, K., Jakovetic, D., Stankovic, V., & Stankovic, L. (2018). Post-processing for event-based non-intrusive load monitoring. *4th international workshop on non-intrusive load monitoring,* 1–4 Conference date: 07-03-2018 Through 08-03-2018.

Iba, H., Hasegawa, Y., & Paul, T. K. (2009). *Applied genetic programming and machine learning* (1st edition). Boca Raton, FL, USA: CRC Press, Inc.

Kagolovsky, Y., & Moehr, J. R. (2003). Current status of the evaluation of information retrieval. *Journal of Medical Systems, 27*, 409–424.

Kavousian, A., Rajagopal, R., & Fischer, M. (2013). Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate building characteristics, appliance stock, and occupants' behavior. *Energy, 55*, 184–194.

Kelly, J., & Knottenbelt, W. (2015). The UK-DALE dataset domestic appliance-level electricity demand and whole-house demand from five UK homes. *Scientific Data, 2*.

Kelly, J., & Knottenbelt, W. (2016). *Does disaggregated electricity feedback reduce domestic electricity consumption? A systematic review of the literature.* (cs) arXiv:1605.00962.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*, 81–93.

Klemenjak, C., Reinhardt, A., Pereira, L., Makonin, S., Bergés, M., & Elmenreich, W. (2019). *Electricity consumption data sets: Pitfalls and opportunities. Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation, BuildSys'19*. New York, NY, USA: ACM159–162.

Klemenjak, C., Makonin, S., & Elmenreich, W. (2020). Towards comparability in non-intrusive load monitoring: On data and performance evaluation. *2020 IEEE power & energy society innovative smart grid technologies conference (ISGT)* (pp. 5). .

Kong, W., Xu, Y., Dong, Z. Y., Hill, D. J., Ma, J., & Lu, C. (2015). An extended prototypical smart meter architecture for demand side management. *2015 IEEE 13th international conference on industrial informatics (INDIN)* (pp. 1008–1013).

Kumar, V., & Leone, R. P. (1991). Nonlinear mapping: An alternative to multidimensional scaling for product positioning. *Journal of the Academy of Marketing Science, 19*, 165–176.

Kumar, R., & Vassilvitskii, S. (2010). Generalized distances between rankings. *Proceedings of the 19th international conference on world wide web, WWW'10* (pp. 571–580). ACM.

Liang, J., Ng, S. K. K., Kendall, G., & Cheng, J. W. M. (2010). Load signature study part I: Basic concept, Structure and Methodology. *IEEE Transactions on Power Delivery, 25*, 551–560.

Lucas, A., Jansen, L., Andreadou, N., Kotsakis, E., & Masera, M. (2019). Load flexibility forecast for DR using non-intrusive load monitoring in the residential sector. *Energies, 12*, 2725.

Luo, D., Norford, L. K., Leeb, S. B., & Shaw, S. R. (2002). Monitoring HVAC equipment electrical loads from a centralized location methods and field test results. *ASHRAE Transactions, 108*, 841–857.

Ma, G., Lin, J., & Li, N. (2018). Longitudinal assessment of the behavior-changing effect of app-based eco-feedback in residential buildings. *Energy and Buildings, 159*, 486–494.

Makonin, S., & Popowich, F. (2017). Nonintrusive load monitoring (NILM) performance evaluation. *Energy Efficiency,* 1–6.

Makonin, S., Ellert, B., Bajic, I. V., & Popowich, F. (2016a). Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014. *Scientific*

Data, 3.
Makonin, S., Popowich, F., Bajic, I. V., Gill, B., & Bartram, L. (2016b). Exploiting HMM sparsity to perform online real-time nonintrusive load monitoring. *IEEE Transactions on Smart Grid, 7*, 2575–2585.
Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (1st edition). New York: Cambridge University Press.
Mayhorn, E. T., Sullivan, G. P., Petersen, J. M., Butner, R. S., & Johnson, E. M. (2016). *Load disaggregation technologies: Real world and laboratory performance*Richland WA (US): Pacific Northwest National Laboratory (PNNL) Technical Report PNNL-SA-116560.
Mayhorn, E. T., Sullivan, G. P., Fu, T., & Petersen, J. M. (2017). *Non-intrusive load monitoring laboratory-based test protocols*Pacific Northwest National Laboratory (PNNL) Technical Report 26184.
Meehan, P., McArdle, C., & Daniels, S. (2014). An efficient, scalable time-frequency method for tracking energy usage of domestic appliances using a two-step classification algorithm. *Energies, 7*, 7041–7066.
Munir, S., Stankovic, J. A., & FailureSense: (2014). Detecting sensor failure using electrical appliances in the home. *IEEE 11th international conference on mobile Ad Hoc and sensor systems,* 73–81 2155-6814.
Murray, D., Stankovic, L., Stankovic, V., Lulic, S., & Sladojevic, S. (2019). Transferability of neural network approaches for low-rate energy disaggregation. *ICASSP 2019 – 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP),* 8330–8334.
Najafi, B., Moaveninejad, S., & Rinaldi, F. (2018). Chapter 17 – Data analytics for energy disaggregation: Methods and applications. In R. Arghandeh, & Y. Zhou (Eds.). *Big data application in power systems* (pp. 377–408). Elsevier.
Nalmpantis, C., & Vrakas, D. (2018). Machine learning approaches for non-intrusive load monitoring: From qualitative to quantitative comparation. *Artificial Intelligence Review,* 1–27.
Orji, U. A., Remscrim, Z., Laughman, C., Leeb, S. B., Wichakool, W., Schantz, C., Cox, R., Paris, J., Kirtley, J. L., & Norford, L. K. (2010). Fault detection and diagnostics for non-intrusive monitoring using motor harmonics. *2010 twenty-fifth annual IEEE applied power electronics conference and exposition (APEC),* 1547–1554.
Pereira, L. (2016). *Hardware and software platforms to deploy and evaluate non-intrusive load monitoring systems.* Funchal Portugal: Universidade da Madeira PhD.
Pereira, L. (2017a). Developing and evaluating a probabilistic event detector for non-intrusive load monitoring. *Proceedings of the fifth IFIP conference on sustainable internet and ICT for sustainability* (pp. 1–10).
Pereira, L. (2017b). *EMD-DF: A data model and file format for energy disaggregation datasets. Proceedings of the 4th ACM international conference on systems for energy-efficient built environments*. Delft, The Netherlands: ACM1–2.
Pereira, L., & Chisik, Y. (2017). A mouse over a hotspot survey: An exploration of perceptions of electricity consumption and patterns of indecision. *2017 sustainable internet and ICT for sustainability (SustainIT)* (pp. 1–4).
Pereira, L., & Nunes, N. J. (2015a). Semi-automatic labeling for public non-intrusive load monitoring datasets. *Proceedings of the fourth IFIP conference on sustainable internet and ICT for sustainability* (pp. 1–4).
Pereira, L., & Nunes, N. J. (2015b). Towards systematic performance evaluation of non-intrusive load monitoring algorithms and systems. *Sustainable internet and ICT for sustainability (SustainIT), 2015* (pp. 1–3).
Pereira, L., & Nunes, N. (2017). A comparison of performance metrics for event classification in non-intrusive load monitoring. *IEEE international conference on smart grid communications (SmartGridComm),* 159–164.
Pereira, L., & Nunes, N. (2018). Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools – A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* e1265.
Pereira, L., & Nunes, N. (2019). Understanding the practical issues of deploying energy monitoring and eco-feedback technology in the wild: Lesson learned from three long-term deployments. *Energy Reports, 6*, 94–106.
Pereira, L., Quintal, F., Barreto, M., & Nunes, N. J. (2013). Understanding the limitations of eco-feedback: A one-year long-term study. In A. Holzinger, & G. Pasi (Eds.). *Human–computer interaction and knowledge discovery in complex, unstructured, big data, lecture notes in computer science* (pp. 237–255). Maribor, Slovenia: Springer Berlin Heidelberg.
Pereira, L., Quintal, F., Gonc calves, R., & Nunes, N. J. (2014a). *Sustdata: A public dataset for ict4s electric energy research. International conference on ICT for sustainability (ICT4S'14)*. Stockholm, Sweden: Atlantis Press359–368.
Pereira, L., Nunes, N., & Bergés, M. (2014b). Surf and surf-pi: A file format and api for non-intrusive load monitoring public datasets. *Proceedings of the 5th international conference on future energy systems, e-Energy'14* (pp. 225–226). ACM.
Pereira, L., Ribeiro, M., & Nunes, N. (2017). Engineering and deploying a hardware and software platform to collect and label non-intrusive load monitoring datasets. *2017 sustainable internet and ICT for sustainability (SustainIT)* (pp. 1–9).
Pipattanasomporn, M., Kuzlu, M., Rahman, S., & Teklu, Y. (2014). Load profiles of selected major household appliances and their demand response opportunities. *IEEE Transactions on Smart Grid, 5*, 742–750.
Quintal, F., Pereira, L., Nunes, N., Nisi, V., & Barreto, M. (2013). WATTSBurning: Design and evaluation of an innovative eco-feedback system. In P. Kotz'e, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler (Eds.). *Human–computer interaction INTERACT 2013* (pp. 453–470). Berlin Heidelberg: Springer number 8117 Notes in Computer Science.
Rodney, A. M., & Poll, S. (2014). *Energy analysis of multi-function devices in an office environment*The Free Library Technical Report.
Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., & Bengio, Y. (2019). *Tackling climate change with machine learning.* (cs, stat) arXiv:1906.05433.
Roncancio, H., Hernandes, A., & Becker, M. (2013). Ceiling analysis of pedestrian recognition pipeline for an autonomous car application. *IEEE workshop on robot vision (WORV),* 215–220.
Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*.
Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*, 427–437.
Spearman, C. (1906). 'Footrule' for measuring correlation. *British Journal of Psychology 1904–1920, 2*, 89–108.
Stankovic, L., Stankovic, V., Liao, J., & Wilson, C. (2016). Measuring the energy intensity of domestic activities from smart meter data. *Applied Energy, 183*, 1565–1580.
Symeonidis, N., Nalmpantis, C., Vrakas, D., & Benchmark, A. (2019). Framework to evaluate energy disaggregation solutions. In J. Macintyre, L. Iliadis, I. Maglogiannis, & C. Jayne (Eds.). *Engineering applications of neural networks, communications in computer and information science* (pp. 19–30). Springer International Publishing, Cham.
Townson, B. (2016). *NILM: Vehicle or destination?* .
Yan, D., Jin, Y., Sun, H., Dong, B., Ye, Z., Li, Z., & Yuan, Y. (2019). Household appliance recognition through a Bayes classification model. *Sustainable Cities and Society, 46*, 101393.
Zeifman, M. (2012). Disaggregation of home energy display data using probabilistic approach. *IEEE Transactions on Consumer Electronics, 58*, 23–31.
Zeifman, M., & Roth, K. (2011). Nonintrusive appliance load monitoring: Review and outlook. *IEEE Transactions on Consumer Electronics, 57*, 76–84.
Zhai, S., Zhou, H., Wang, Z., & He, G. (2020). Analysis of dynamic appliance flexibility considering user behavior via non-intrusive load monitoring and deep user modeling. *CSEE Journal of Power and Energy Systems, 6*, 41–51.
Zhao, B., Stankovic, L., & Stankovic, V. (2016). On a training-less solution for non-intrusive appliance load monitoring using graph signal processing. *IEEE Access, 4*, 1784–1799.
Zoha, A., Gluhak, A., Imran, M. A., & Rajasegarar, S. (2012). Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey. *Sensors, 12*, 16838–16866.