Received: 5 January 2018 | Revised: 10 April 2018 | Accepted: 25 April 2018

DOI: 10.1002/widm.1265

WILEY WIREs DATA MINING AND KNOWLEDGE DISCOVERY

FOCUS ARTICLE

Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—A review

Lucas Pereira[1] | Nuno Nunes[2]

[1]M-ITI/LARSYS, Funchal, Portugal

[2]M-ITI/LARSYS, Técnico, University of Lisbon, Lisbon, Portugal

Correspondence
Lucas Pereira, M-ITI/LARSYS, Madeira Madeira Interactive Technologies Institute, Funchal, Portugal.
Email: lucas.pereira@m-iti.org

Funding information
Fundação para a Ciência e a Tecnologia Grant Numbers: SFRH/DB/77856/2011, UID/EEA/50009/2013

Non-intrusive load monitoring (also known as NILM or energy disaggregation) is the process of estimating the energy consumption of individual appliances from electric power measurements taken at a limited number of locations in the electric distribution of a building. This approach reduces sensing infrastructure costs by relying on machine learning techniques to monitor electric loads. However, the ability to evaluate and benchmark the proposed approaches across different datasets is key for enabling the generalization of research findings and consequently contributes to the large-scale adoption of this technology. Still, only recently researchers have focused on creating and standardizing the existing datasets in order to deliver a single interface to run NILM evaluations. Furthermore, there is still no consensus regarding, which performance metrics should be used to measure and report the performance of NILM systems and their underlying algorithms. This paper provides a review of the main datasets, metrics, and tools for evaluating the performance of NILM systems and technologies. Specifically, we review three main topics: (a) publicly available datasets, (b) performance metrics, and (c) frameworks and toolkits. The review suggests future research directions in NILM systems and technologies, including cross-datasets, performance metrics for evaluation and generalizable frameworks for benchmarking NILM technology.

This article is categorized under:
Application Areas > Science and Technology
Application Areas > Data Mining Software Tools
Technologies > Computational Intelligence
Technologies > Machine Learning

KEYWORDS
datasets, energy disaggregation, metrics, Non-Intrusive Load Monitoring, performance evaluation, smart-grids, tools

1 | INTRODUCTION

In the past three decades, a substantial body of research has been devoted to the development of non-intrusive load monitoring (NILM) approaches that are able to sense and disaggregate energy consumption from measurements taken at a limited number of locations in the electric distribution infrastructure. NILM technology, also known as energy disaggregation, is key to reduce the sensing infrastructure costs in buildings and even electrical grids. It contrasts approaches that rely on deploying and connecting multiple sensors in each appliance to monitor their consumption.

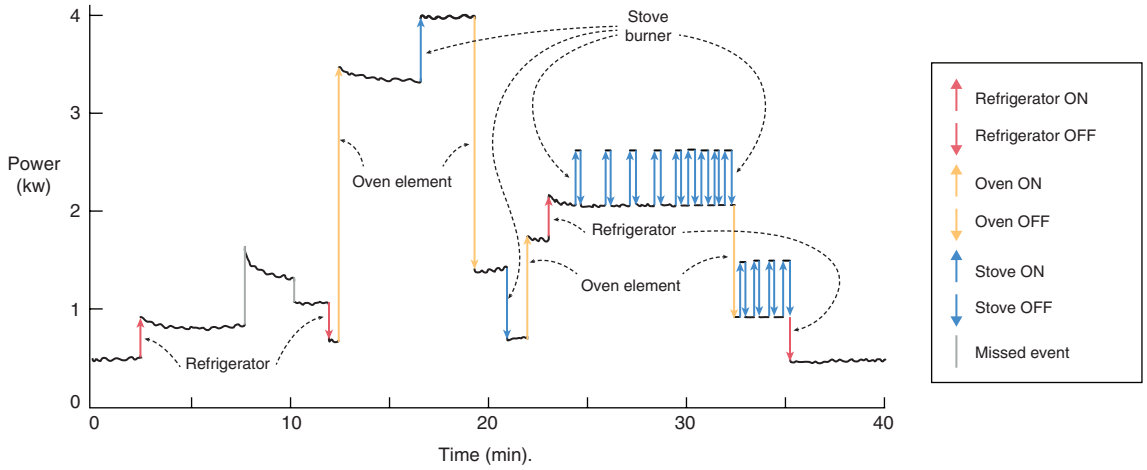

FIGURE 1 Example of event-based energy disaggregation (Reprinted with permission from Hart (1992). Copyright 1992 IEEE)

Early research on this topic dates back to 1985 when George Hart from the Massachusetts Institute of Technology (MIT) coined the term Non-intrusive (Appliance) Load Monitoring (NIALM) (Hart, 1985). In very simple terms NILM is defined as the set of signal processing and machine-learning techniques used to estimate the aggregate and individual appliance electricity consumption from electric power measurements taken at a limited number of locations in the electric distribution infrastructure of a house or building (optimally the mains, hence covering the demand of the entire space).

Still, only recently NILM gained renewed attention from the research community. This was motivated by the availability of advanced metering technologies (e.g., smart-grids) and also because of energy efficiency concerns motivated by the need to reduce the carbon footprint of buildings and households (Carrie Armel, Gupta, Shrimali, & Albert, 2013). Furthermore, leveraging on the advances in machine learning, NILM technology is expected to serve as the backbone technology that will enable the creation of innovative smart-grid services that go beyond helping individuals saving energy (Townson, 2016).

The combination of the need for technologies promoting low-carbon emissions and the advances in machine learning and statistical techniques is generating a substantial amount of energy disaggregation review papers such as Nalmpantis and Vrakas (2018), Esa, Abdullah, and Hassan (2016), Abubakar, Khalid, Mustafa, Shareef, and Mustapha (2016), Wong, Ahmet Sekercioglu, Drummond, and Wong (2013), Butner, Reid, Hoffman, Sullivan, and Blanchard (2013), Makonin (2012), Zoha, Gluhak, Imran, and Rajasegarar (2012), Jiang, Li, Luo, Jin, and West (2011), and Zeifman and Roth (2011). In general NILM approaches are grouped into two categories: (a) event-based approaches and (b) event-less approaches (Bergés & Kolter, 2012).

Event-based approaches are intrinsically related to the early days of NILM. They seek to disaggregate the total consumption by means of detecting and labeling every appliance transition in the aggregated signal (see Figure 1). Event-based approaches rely on previously trained supervised or semi-supervised learning algorithms to label the electric load power events. Consequently, approaches under this category require a data collection step where a number of transitions (i.e., power events) from the appliances of interest are collected, labeled, and stored, to be used later as training data.

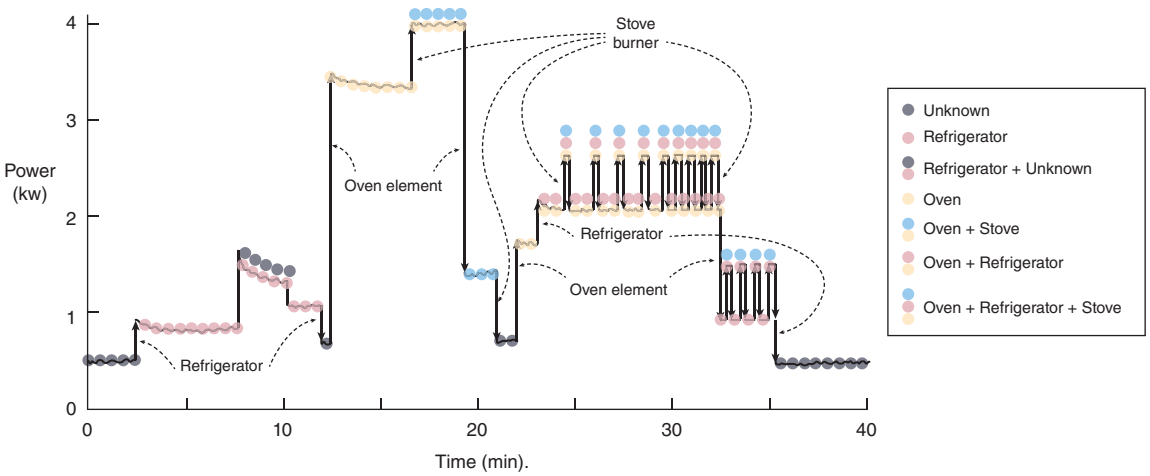

FIGURE 2 Example of event-less energy disaggregation (Reprinted with permission from Hart (1992). Copyright 1992 IEEE)

Event-less approaches, on the other hand, do not rely on event detection (ED) and classification. Instead, these approaches attempt to match each sample of the aggregated power with the consumption of one specific appliance or a combination of different appliances (see Figure 2), by means of statistical (e.g., Bayesian methods) and probabilistic (e.g., hidden Markov models) machine-learning methods. Therefore, the training data does not require any labeled transitions. Instead, only the aggregated consumption of the loads of interest is required, hence turning the process of collecting training data for event-less approaches more straightforward than for event-based approaches.

Despite the growing body of work in this field, there are still many challenges that need to be addressed before NILM technology becomes practical and reliable. One of such challenges is the many issues related with the replication and generalization of research findings (e.g., the lack of proper test and training data and the absence of a formal agreement on how to report the disaggregation results). This important research challenge only recently became the focus of a group of NILM researchers such as Butner et al. (2013), Batra, Kelly, et al. (2014), Batra, Parson, et al. (2014), Makonin and Popowich (2014), Mayhorn, Sullivan, Petersen, Butner, and Johnson (2016), and Pereira and Nunes (2017).

In this paper, we present the first review of the research efforts toward the performance evaluation of NILM algorithms and systems. These efforts are summarized in the next three sections. More specifically, first we present, describe, and compare the currently publicly available NILM datasets. Second, we review the performance metrics reported to assess the accurateness of the proposed NILM algorithms and systems. Third, we present a number of tools and frameworks developed to leverage the potential of NILM datasets and performance metrics.

2 | PUBLIC DATASETS

An energy disaggregation dataset is a collection of electrical energy measurements taken from real-world scenarios, without disrupting the everyday routines in the monitored space, that is, trying to keep the data as close to reality as possible.

These usually contain measurements from the aggregate consumption (taken from the mains) and of the individual loads (i.e., ground-truth data), which is obtained either by measuring each load at the plug-level or measuring the individual circuit to which the load is connected. In a real-world scenario, typically multiple loads are connected to the same circuit. Therefore, plug-level measuring does not always ensure the availability of individual consumption data for each load.

Similarly to the classification of NILM techniques, the currently available datasets can also be categorized as event-based or event-less datasets. The major difference between the two is that event-less approaches do not require the identification of individual power changes. Consequently, collecting datasets for event-less approaches is more straightforward and less time consuming. This partly explains the higher availability of event-less datasets, as we will see next.

Currently, the best of our knowledge, there are 26 publicly available NILM datasets. From these, the vast majority, 21, are suitable to evaluate event-less approaches and only 5 can be used to evaluate event-based approaches.

In order to facilitate comparisons between the existing solutions, in Table 1 we summarize the 26 datasets, sorted by year of initial release. From the 26 datasets, 23 are from individual households, 1 from a living lab household energy smart home lab (ESHL), and 2 from university buildings commercial building dataset (COMBED) and building-level office eNvironment dataset (BLOND).

The following characteristics are provided for each dataset: Year of release, country, number of monitored households, if the data is continuous or not (continuous—C or not continuous—NC), that is, if the data was collected in consecutive time periods. The approaches enabled by the dataset (event-based—EB or event-less—EL), types of smart-meters used in the collection (aggregate—A, individual circuit—IC or individual appliance—IA, and if a list of power event labels is available—LE). The available electric energy features (current—I, voltage—V, active power—P, reactive power—Q, apparent power—S, others—O), and the time resolution of the available data.

In the event-less category, there are 21 datasets. From these, 17 (REDD, Smart*, BLUED, Dataport, AMPds, iAWE, IHEPCDS, UK-DALE, ECO, REFIT, COMBED, DRED, SustDataED, EEUD, RAE, ESHL, and BLOND) contain aggregated and individual appliance/circuit consumption, thus making them suitable to be used simultaneously as training and testing data.

The remaining five datasets in this category (HES, Tracebase, ACS-Fx, GREEND, and RBSA), only provide individual appliance consumption information. Therefore, they can only serve as training data. One possibility to use them for training and testing would be to calculate the aggregate data by summing the power demand of each appliance. Still, this approach presents some caveats that would affect the final results. For example, this completely excludes the effects of the appliances that were not submetered, hence resulting in very simplistic and unrealistic datasets that may lead to overoptimistic disaggregation results.

In the event-based category, only BLUED and SustDataED contain aggregate consumption information and a list of appliance labels for the identified power changes. Therefore, these are the only two alternatives to evaluate event-based approaches. Still, since both datasets only contain, respectively, 7 and 10 days of data, they are not very suitable to evaluate classification

TABLE 1 Overview of publicly available energy monitoring and disaggregation datasets

Dataset	Country (Sites)	Dur.	Approach		Meters				Features						Resolution
			EB	EL	A	IC	IA	LE	I	V	P	Q	S	O[h]	
REDD (2011)	United States (6)	2–4 W (NC)	×	✓	✓	✓	✓	×	✓	✓	✓	×	×	—	I, V: 15 kHz P: 1 Hz IC, IA: 3–4 s
Smart* (2012)	United States (8[a])	3–4 M 3 Y (NC)	×	✓	✓	✓	×	×[b]	×	×	✓	×	✓	—	2012: 1 Hz 2017: 30 m
BLUED (2012)	United States (1)	1 W (C)	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	×	—	I, V: 12 kHz P, Q: 60 Hz IA: 1 Hz
HES (2012)	United Kingdom (251)	1–12 M (C)	×	✓[c]	×	×	✓	×	×	×	×	×	×	EE	2–5 m
Tracebase (2012)	Germany (N/A)	1883 d (N/C)	×	✓[c]	×	×	✓	×	×	×	✓	×	×	—	1–10 s
Dataport (2013)	United States (1400+)	4 Y[e] (C)	×	✓	✓	✓	✓	×	×	×	✓	×	✓	—	1 m
AMPds (2013)	Canada (1)	2 Y[e] (C)	×	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	F PF EE	1 m
iAWE (2013)	India (1)	74 D (C)	×	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	F PA EE	1 Hz
IHEPCDS (2013)	France (1)	4 Y (C)	×	✓	✓	✓	×	×	✓	✓	✓	✓	×	—	1 m
ACS-Fx (2013)	Switzerland (N/A)	N/A	×	✓[c]	×	×	✓	×	✓	✓	✓	✓	×	PA	10 s
UK-DALE (2014)	United Kingdom (5)	655 D (NC)	×	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	—	I, V: 16 kHz P, Q: 1 Hz A, IA: 6 s
ECO (2014)	Switzerland (6)	8 M (NC)	×	✓	✓	×	✓	×	✓	✓	✓	×	×	PA	1 Hz
REFIT (2014)	United Kingdom (20)	2 Y (C)	×	✓	✓	×	✓	×	×	×	✓	×	×	EP	6–8 s
GREEND (2014)	Austria, Italy (9)	3–6 M (C)	×	✓[c]	×	×	✓	×	×	×	✓	×	×	—	1 Hz
PLAID I and II (2014)	United States (55)	N/A	✓[d]	×	×	×	✓	×	✓	✓	×	×	×	—	30 kHz
RBSA (2014)	United States (101)	27 M	×	✓[c]	×	✓	✓	×	×	×	×	×	×	EE	15 m
COMBED (2014)[g]	India (1)	1 M	×	✓	✓	✓	×	×	✓	×	✓	×	×	EE	30 s
DRED (2015)	Holand (1)	6 M (C)	×	✓	✓	×	✓	×	×	×	✓	×	×	—	1 Hz, 1 m
HFED (2015)	India (N/A)	N/A	✓[d]	×	×	×	✓	×	×	×	×	×	×	EMI	10 kHz 5 MHz
WHITED (2016)	Germany, Austria, Indonesia (N/A)	N/A	✓[d]	×	×	×	✓	×	✓	✓	×	×	×	—	44.1 kHz
COOLL (2016)	France (N/A)	N/A	✓[d]	×	×	×	✓	×	✓	✓	×	×	×	—	100 kHz
SustDataED (2016)	Portugal (NC)	10 D	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	×	—	I, V: 12.8 kHz P, Q: 50 Hz IA: 0.5 Hz
EEUD (2017)	Canada (23)	1 Y	×	✓	✓	✓	✓	×	×	×	✓	×	×	—	1 m
ESHL (2017[e])	Germany (1[f])	4–5 Y (NC)	×	✓	✓	×	✓	×	✓	✓	✓	×	×	—	0.5–1 Hz
RAE (2018)	Canada (1)	72 D (C)	×	✓	✓	✓	×	×	✓	✓	✓	✓	✓	F PF	1 Hz
BLOND (2018)	Germany (1[g])	50–230 D (C)	×	✓	✓	×	✓	×	✓	✓	×	×	×	—	A: 50-250 kHz IA: 6.4-50 kHz

REDD (Kolter & Matthew, 2011)	Smart* (Barker, Mishra, Irwin, Cecchet, & Shenoy, 2012)
BLUED (Anderson et al., 2012)	HES (Household electricity survey a study of domestic electrical product usage, 2012)
Tracebase (Reinhardt et al., 2012)	Dataport (Holcomb, 2012)
AMPds (Makonin, Ellert, Bajić, & Popowich, 2016)	iAWE (Batra, Gulati, Singh, & Srivastava, 2013)
IHEPCDS (Bache & Lichman, 2013)	ACS-Fx (Gisler, Ridi, Zufferey, Khaled, & Hennebert, 2013; Ridi, Gisler, & Hennebert, 2014)
UK-DALE (Kelly & Knottenbelt, 2015)	REFIT (Murray et al., 2015)
GREEND (Monacchi, Egarter, Elmenreich, D'Alessandro, & Tonello, 2014)	PLAID I and II(Baets et al., 2017; Gao, Giri, Kara, & Bergés, 2014)
RBSA (Ecotope Inc, 2014)	COMBED (Batra, Parson, et al., 2014)
DRED (Uttama Nambi, Reyes Lua, & Prasad, 2015)	HFED (Gulati, Ram, & Singh, 2014)

TABLE 1 (Continued)

Dataset	Country (Sites)	Dur.	Approach		Meters				Features						Resolution
			EB	EL	A	IC	IA	LE	I	V	P	Q	S	O[h]	
WHITED (Kahl, Ui Haq, Kriechbaumer, & Hans-Arno, 2016)					COOLL (Picon et al., 2016)										
SustDataED (Ribeiro, Pereira, Quintal, & Nunes, 2016)					EEUD (Johnson & Beausoleil-Morrison, 2017)										
RAE (Makonin, Wang, & Tumpach, 2018)					ESHL (Kaibin Bao, 2016)										
BLOND (Kriechbaumer & Jacobsen, 2018)															

[a] In the original version (2012) there is data for three houses between 3 and 4 weeks, but only one of them contains submetered data. In the 2017 edition there is data for seven houses during 3 years.

[b] There is a list of light switch events that can be used to label the events of lighting appliances.

[c] These datasets can only be used as training data. Evaluation must happen in datasets where aggregate consumption data is available.

[d] Only for event classification using either cross-validation or power events from other datasets.

[e] Should have been released in 2017 according to the authors.

[f] The home is a living lab.

[g] University/office building.

[h] EE: electric energy; EMI: electromagnetic interference; F: frequency; P: energy price; PA: phase angle; PF: power factor.

and energy estimation (EE) algorithms. One possibility is using a resampling technique (e.g., bootstrapping or jackknifing) to generate new labels from the existing ones, that can later be used as training data for event classification.

Finally, the PLAID, HFED, WHITED, and COOLL datasets only contain data from the startup transients and spectral traces of several individual appliances. Consequently, they are only suitable to evaluate feature extraction and classification algorithms using cross-validation (Barsim, Mauch, & Yang, 2016; Gao, Kara, Giri, & Bergés, 2015). Likewise, it should also be possible to use PLAID, WHITED, and COOLL to classify power events from other datasets. Still, it should be noted that they only contain startup transients, therefore it will not be possible to classify OFF transitions.

3 | PERFORMANCE METRICS

As previously mentioned in the introduction of this paper, most of the early efforts in the NILM research were devoted to event-based approaches. Consequently, several performance metrics have been proposed to evaluate such systems. For example, in his seminal work, Hart (1985) used both the fraction of correctly classified power events and the fraction of total energy explained as accuracy metrics. The former evaluates the performance of the event classification step, whereas the latter evaluates the EE step.

Many other performance metrics have been defined in the following years. For instance, in Berges (2010) the author used different metrics for ED (e.g., failed detections—*FD*, and the detection error rate—*DER*), event classification (e.g., $F_1 - Score$), and EE (e.g., the energy identification rate—*EIR*). A similar approach, of considering the different steps in the NILM pipeline, was presented in Liang, Ng, Kendall, and Cheng (2010). In this work, the authors propose three different metrics to evaluate the event-based NILM pipeline, namely, detection accuracy—*DeA*, disaggregation accuracy—*DiA*, and overall accuracy—*OA*.

All these metrics take into consideration event detector, type I error (when no appliance changes its state, but an event is detected) and type II error (when an appliance is operated but no event is detected). However, they assume that all the events are equally important, which is far from plausible since not all appliances require the same energy. Therefore, in an attempt to quickly understand the interactions between ED errors and the actual energy needs, in Anderson, Ocneanu, et al. (2012) the authors proposed two new performance metrics: the total power change (*TPC*) and the average power change (*APC*), which are the sum (average in the second case) of the amount of power change for all the type I and type II errors.

Event-less approaches, on the other hand, rarely rely on a separate ED process. Instead event-less approaches attempt to disaggregate the total load in separate time slices. Consequently, event-less algorithms only require metrics to evaluate the final EE, which is the last stage in the case of event-based approaches.

The first event-less metrics formulations took into consideration the estimated and the ground-truth energy to capture the energy disaggregation error. For example, in Kolter and Matthew (2011) the authors propose a performance metric that captures the total error in the assigned energy (*TEAE*), normalized by the real energy consumption in each time slice averaged over all appliances. An equivalent metric was proposed in Kolter and Jaakkola (2012), but this time considering the individual appliance error (*IATEAE*) rather than the average between all the appliances, which reduces the chance of reporting large errors in certain time slices due to single appliances performing very poorly.

TABLE 2 Overview of performance metrics under the event detection category

Metric	Description	Equation
TP	A true positive (Batra, Kelly, et al., 2014; Beckel et al., 2014; Berges, 2010) is considered whenever the system detects/classifies something as being true and the actual output is true, for example, a power event is labeled as being triggered by appliance *A* and it actually was (event-based), or a time slice consumption is attributed to appliance *A* which is actually responsible for it (event-less).	—
TN	A true negative (Batra, Kelly, et al., 2014; Beckel et al., 2014; Berges, 2010) is considered whenever the system detects/classifies something as being false and the actual output is false, for example, no power event is detected in a given instant and actually no appliance changed its state in that instant (event-based), or for a given time slice no consumption is attributed to an appliance when that appliance is actually not consuming.	—
FP	A false positive (Batra, Kelly, et al., 2014; Beckel et al., 2014; Berges, 2010) is considered whenever the system detects/classifies something as being true and the actual output is false, for example, a power event is labeled as being triggered by appliance *A* when it was triggered by appliance *B* (event-based), or a time slice consumption being attributed to appliance > when that appliance is not working (event-less).	—
FN	A false negative (Batra, Kelly, et al., 2014; Beckel et al., 2014; Berges, 2010) is considered whenever the system detects/classifies something as being false and the actual output is true, for example, no power event is detected at a given instant but an appliance changed its state in that instant (event-based), or for a given time slice no consumption is attributed to an appliance when that appliance is actually consuming (event-less).	—
P	Precision (Anderson, Bergés, Ocneanu, Benitez, & Moura, 2012; Batra, Kelly, et al., 2014; Beckel et al., 2014; Berges, 2010) (also called positive predictive value—*PPV*), is the proportion of relevant instances that were reported as being relevant against all the instances that were reported as relevant.	$P = \frac{TP}{TP+FP}$
R	Recall (Anderson, Bergés, et al., 2012; Batra, Kelly, et al., 2014; Beckel et al., 2014; Berges, 2010) (also called sensitivity or true positive rate—*TPR*), is the proportion of relevant instances that were reported as being relevant against all the truly relevant instances.	$R = \frac{TP}{TP+FN}$
F_β	The F_β-measure (Anderson, Bergés, et al., 2012; Batra, Kelly, et al., 2014; Beckel et al., 2014; Berges, 2010) trades-off precision and recall. Mathematically, it is the harmonic mean between the two metrics. β is a weighing factor that is used to attach β times as much importance to recall as to precision. For example, if $\beta = 2$ (F_2-measure), recall is twice as important as precision, whereas if $\beta = 0.5$ ($F_{0.5}$-measure), precision is twice as important as recall. Finally, if $\beta = 1$ (F_1-measure), recall and precision have the same weight.	$F_\beta = \left(1+\beta^2\right) \times \frac{P \times R}{\left(\beta^2 \times P\right)+R}$
FPR	False positive rate (Anderson, Bergés, et al., 2012; Batra, Kelly, et al., 2014) is the proportion of false positives against the actual negative results.	$FPR = \frac{FP}{FP+TN}$
ROC − AUC	The receive operating characteristics - area under curve (Berges, 2010; Liang et al., 2010; Parson, Mark, & Rodgers, 2014) metric finds the algorithm / parameter configurations that have the best trade-off between *TPR* and *FPR*. The area under the ROC curve measures accuracy. An area of 1 represents a perfect test; an area of.5 represents a random (therefore worthless) test.	a
FD	Failed detections (Berges, 2010) is the sum of missed (*FN*) and wrongfully detected events (*FP*).	$FD = FP + FN$
DER	Detection error rate (Berges, 2010) is the ration between failed detections and the number of positive cases.	$DER = \frac{FD}{TP+FN}$
DeA	Detection accuracy (Liang et al., 2010) measures the accuracy of the algorithm, including the effects of the wrongfully detected events.	$DeA = \frac{N_{dis}}{N_{det}}$
DiA	Disaggregation accuracy (Liang et al., 2010) is the *DA* excluding the effects of false positives.	$DiA = \frac{N_{dis}}{N_{det}-N_{wro}}$
OA	Overall accuracy (Liang et al., 2010) is the disaggregation accuracy including the effects of false positives and false negatives.	$OA = \frac{N_{dis}}{N_{true}}$
TPP	True positive percentage (Anderson, Bergés, et al., 2012) is the percentage of the ratio between true positives and actual true results. *TPP* is equivalent to recall.	$TPP = \frac{FP}{TP+FN}$
FPP	False positive percentage (Anderson, Bergés, et al., 2012) is the percentage of the ratio between false positives and actual true results. Note that since the number of *FP* can be larger that the number of real events, *FPP* can be larger than 100%.	$FPP = \frac{FP}{TP+FN}$
TPC	Total power change (Anderson, Bergés, et al., 2012) is the sum of the deltas for all the false positives or false negatives. Event-based	$TPC_{FP} = \sum_{f \in FP} \mid \Delta P_f \mid TPC_{FN} = \sum_{m \in FN} \mid \Delta P_m \mid$
APC	Average power change (Anderson, Bergés, et al., 2012) is the average of the deltas for all the false positives or false negatives. Event-based	$APC_{FP} = \frac{1}{\mid FP \mid} \times TPC_{FP}\ APC_{FN} = \frac{1}{\mid FN \mid} \times TPC_{FN}$
HL	The hamming loss (Batra, Kelly, et al., 2014) measures the total information loss when appliances are incorrectly classified over the entire dataset.	$\frac{1}{T}\sum_t \frac{1}{A}\sum_a XOR\left(s_t^{(a)}, \hat{s}_t^{(a)}\right)$
MF − Score	The modified F-score (Kim, Marwah, Arlitt, Lyon, & Han, 2011) combines the performance of classification and energy estimation algorithms. It works by introducing the notion of accurate and inaccurate true positives (ATP and ITP). ATP are the correct classifications with low energy estimation errors, whereas ITP are the correct classifications with high estimation errors.[b].	$ATP = E_t > 0 \wedge \hat{E}_t > 0 \wedge \frac{\mid E_t - \hat{E}_t \mid}{E_t} \leq \rho$ $ITP = E_t > 0 \wedge \hat{E}_t > 0 \wedge \frac{\mid E_t - \hat{E}_t \mid}{E_t} > \rho$

TABLE 2 (Continued)

Metric	Description	Equation
$FSF-Score$	The finite-state F-score (Makonin & Popowich, 2014) is a discrete version of the F_β-score. A partial penalization (*inacc*) is applied to the correct classifications (TP), based on the distance between the estimated and the ground truth states.[c]	$inacc = \sum_t \frac{\lvert \hat{s}_t^{(a)} - s_t^{(a)} \rvert}{S^{(a)}}$
BM	Informedness (Powers, 2011; (Barsim & Yang, 2018) quantifies how informed a predictor is for the specified condition, and specifies the probability that a prediction is informed in relation to the condition (versus chance).	$BM = TPR + TNR - 1$
MK	Markedness (Barsim & Yang, 2018; Powers, 2011) quantifies how marked a condition is for the specified predictor and specifies the probability that a condition is marked by the predictor (versus chance).	$MK = PPV + NPV - 1$
MCC	The Mathews correlation coefficient (Barsim & Yang, 2018; Matthews, 1975) indicates the central tendency between Informedness and Markedness. Mathematically, it is the geometric mean between the two metrics.	$MCC = \sqrt{BM \times MK}$
N_{dis}	Number of events that is accurately recognized by a disaggregation algorithm, that is, *TP*	—
N_{det}	Total number of detected events	$(N_{det} = N_{true} + N_{wro} - N_{miss})$
N_{true}	True number of events that actually occurred, that is, $TP + FN$	—
N_{wro}	Number of wrongfully detected events, that is, *FP*	—
N_{miss}	Number of missed events, that is, *FN*	—
ΔP_m	Power change of a missed event, that is, *FN*	—
ΔP_f	Power change of a wrongfully detected event, that is, *FP*	—
T	Number of observations or events based on each time step	—
A	Number of appliances being considered	—
$s_t^{(a)}$	Metered state of appliance a at instant t	—
$\hat{s}_t^{(a)}$	Estimated state of appliance a at instant t	—
$E_t^{(a)}$	Metered energy of appliance a at instant t	—
$\hat{E}_t^{(a)}$	Estimated energy of appliance a at instant t	—
ρ	Threshold used to define *Accurate* and *Inaccurate* true positives	—
TNR	True negative rate (or specificity) is the inverse recall	$TNR = TN \div (TN + FP)$
NPV	Negative predictive value is the inverse precision	$NPV = TN \div (TN + FN)$

[a] Trapezoidal rule for scoring algorithms, the non-parametric Wilcoxon statistic for discrete algorithms. (Iba, Hasegawa, & Paul, 2009)
[b] $P = \frac{ATP}{ATP + ITP + FP}$, $R = \frac{TP - inacc}{TP + FN}$ [c] $P = \frac{TP - inacc}{TP + FP}$, $R = \frac{TP - inacc}{TP + FN}$

Furthermore, the authors working on event-less approaches "reinvented" the notions of false positives, false negatives, true positives, and true negatives in terms of time slice results. This enables common confusion-matrix based metrics like accuracy, precision, recall, sensitivity, F_1-score, and receiver operating characteristic (ROC) to be considered when evaluating event-less approaches (Batra, Kelly, et al., 2014; Batra, Parson, et al. 2014; Beckel, Kleiminger, Cicchetti, Staake, & Santini, 2014; Liang et al., 2010; Makonin & Popowich, 2014).

More recently, there was an effort to better understand how to report the performance of NILM algorithms (Butner et al., 2013; Makonin & Popowich, 2014; Mayhorm, Butner, Baechler, Sullivan, & Hao, 2015). These efforts culminated in a proposal to group the performance metrics in two main categories:(Mayhorn et al., 2016) (a) ED metrics, designed to evaluate the NILM's ability to track the consumption over time, and (b) EE metrics, designed to characterize and evaluate the NILM disaggregated data against the actual ground-truth.

ED performance includes metrics derived from the confusion matrix (i.e., true positives, false positives, true negative, and false negatives). EE performance includes metrics based on basic and advanced statistics (e.g., root mean squared error—*RMSE*, average error—*AE*, and standard deviation of error—*SDE*) (Holmes, 2014), as well as a number of metrics specifically designed for NILM, including the previously mentioned *EIR* (Berges, 2010), *TEAE* (Kolter & Matthew, 2011) and *IATEAE* (Kolter & Jaakkola, 2012).

In Tables 2 and 3, we summarize the event-detection and EE metrics, respectively. For each metric we present a brief description, and a list of papers where these metrics are referenced. Finally, if not otherwise mentioned, the metric can be used to evaluate both event-based and event-less approaches.

Regarding the ED metrics, we can observe that TPC and APC can only be used for ED and classification algorithms. This happens because these metrics rely on the amount of power change in the vicinity of the power events, which only makes sense when the ED problem is considered. As for the remaining metrics, they can be generalized for event-less by defining the confusion matrix for each individual appliance in terms of time slices.

It is also important to note that event classification and EE algorithms are multiclass problems. Consequently, when using metrics under this category, it is necessary to take into consideration the different strategies to calculate the confusion matrix and the metrics themselves.

Regarding event classification, the most well-known strategy is to transform the multiclass problem into separate binary problems either using the one-versus-all or one-versus-one strategy (Galar, Fernández, Barrenechea, Bustince, & Herrera, 2011). The one-versus-all (or one-vs.-rest, OvR or OvA, one-against-all, OAA) strategy involves evaluating the classifier as many times as the number of classes (i.e., N binary problems). For each class, the samples of that class are considered positives examples, and all the other samples negative examples. The one-versus-one approach (OvO) strategy involves evaluating all the possible pairs of classes (i.e., $N(N-1)/2$ binary problems) where one class is the positive and the other the negative example. In both cases, the final confusion matrix is given by summing up the individual binary matrices.

Regarding EE, multiclass metrics are calculated from the resulting OvA or OvO confusion matrices. They can be calculated over the entire class collection, which is called microaveraging, or by averaging the performance of each individual class, which is called macroaveraging (Van Asch, 2013). In microaveraging, each class counts the same for the average, as such larger classes dominate the measure. In macroaveraging, first the average for each class is determined, and only then each class counts the same for the final average. Finally, it is also common-practice to weight the individual class metrics by the respective number of instances, thus making the final average less sensitive to smaller classes. This is known as weighted macroaverage.

With respect to EE metrics, it is important to remark that in the cases of *TECA*, *ETEA*, *Dev*(iation), and *FTEAC*, individual appliance consumption data is required. Furthermore, despite the fact that these metrics give an overview of the performance with just one number, it should be noted that we lose track of which appliances are pushing the performance up or down. For example, poor estimation for high consuming appliances can lead to poor overall results, even if the EE of the remaining appliances is accurate.

As for the remaining metrics, they do not require individual appliance data, meaning that they can be used to evaluate EE algorithms in event-based datasets, even if the individual appliance consumption data is not available. Furthermore, when individual data is available, it is possible to evaluate the performance of each individual appliance, which will immediately give an idea of which appliances affect the performance the most.

4 | TOOLS AND FRAMEWORKS

There is a general consensus in the NILM research community on the importance of public datasets in furthering energy disaggregation research. Nevertheless, despite the tremendous efforts in releasing public data, there are many barriers toward making the datasets easily and efficiently available to the research community. In fact, the most common way of releasing publicly accessible NILM datasets is using text files that follow a certain structure that is then passed to the users in a disparate array of formats from CSV to plain-text. This is particularly inefficient given the requirements of NILM dataset which involve storing and recording information related to power signals, event labeling and properties of appliances and the household among many others (e.g., Lai, Trayer, Ramakrishna, & Li, 2012). Consequently, before any evaluation step, researchers have to understand the underlying structure of the datasets and produce code to interface with them, as well as to accommodate the different performance metrics.

Against this background, several efforts emerged to homogenize the datasets and provide a single interface to run evaluations. In this section we introduce some of these projects, namely the NILM Metadata framework (Kelly & Knottenbelt, 2014), the Energy Monitoring and Disaggregation Data Format (EMD-DF) (Pereira, 2016, 2017), the open-source NILM Toolkit (NILMTK) (Batra, Kelly, et al., 2014; Batra, Parson, et al. 2014; Kelly et al., 2014), and the NILM-Eval framework (Beckel et al., 2014; Cicchetti, 2013).

4.1 | NILM metadata

The NILM Metadata is a framework created with the goal of homogenizing the definition of the many elements that can be found on a NILM dataset (e.g., information about the monitored appliances, the monitoring hardware, and the buildings where the collection occurred). The proposed schema is divided in two main parts: (a) a central metadata with general information about how appliances are represented in NILM metadata; and (b) a schema describing the datasets.

The central metadata provides a base for all the appliances that will appear represented in any of the datasets. This includes, for example, (a) categories for each appliance type; (b) prior knowledge about the distribution of variables such as: ON power and ON duration; (c) appliance correlations (e.g., that the TV is usually ON if the games console is ON); and (d) additional properties for each appliance.

TABLE 3 Overview of performance metrics under the energy estimation category

Metric	Description	Equation
RE	The relative error (Mayhorn et al., 2016) gives an indication of how good the energy estimation is relative to the ground-truth data.	$\frac{\sum_{i=1}^{N} E_i - \sum_{i=1}^{N} \hat{E}_i}{\sum_{i=1}^{N} E_i}$
RMSE	The root mean square error (Chris Holmes, 2014;Batra, Kelly, et al., 2014 ; Mayhorn et al., 2016) is the standard deviation of the energy estimation errors. The *RMSE* reports based on how spread-out these errors are. In other words, it tells you how concentrated the estimations are around the true values. The *RMSE* reports on the same unit as the data, thus making it an intuitive metric.	$1 - \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(E_i - \hat{E}_i\right)^2}\overline{E}$
AE	The average error (Chris Holmes, 2014; Mayhorn et al., 2016) indicates if the estimated energy is on average overestimated or underestimated. A positive *AE* implies an overall higher proportion of overestimation, a negative *AE* implies a higher proportion of underestimation.	$\frac{1}{N}\sum_{i=1}^{N}\Delta E_i$
SDE	The standard deviation of error (Chris Holmes, 2014; Mayhorn et al., 2016) indicates the extent of spread of the differences around the *AE* estimation. A larger *SDE* implies a wider dispersion of the individual estimated values, a lower *SDE* implies tighter distributions.	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\Delta E_i - \overline{\Delta E}\right)^2}$
r^2	The *R*-squared (Mayhorm et al., 2015; Mayhorn et al., 2016; Mayhorn, Sullivan, Fu, & Petersen, 2017) is a statistical measure of how close the estimations are from the ground-truth data. The higher the coefficient, the higher percentage of estimations is inline with the ground-truth. Values of 1 or 0 indicate that the estimation represents all or none of the data, respectively.	$1 - \sum_{i=1}^{N} \frac{\left(E_i - \hat{E}_i\right)^2}{\left(E_i - \overline{E}_i\right)^2}$
%SDx	The percent (%) standard deviation eXplained (Mayhorm et al., 2015; Nau, 2017) is the percent by which the standard deviation of the errors is less than the standard deviation of the measured data. It is believed to be more intuitive than the r^2 since it reports in the same units as the actual data.	$1 - \sqrt{1 - r^2}$
EE	The energy error (Batra, Kelly, et al., 2014; Mayhorn et al., 2016) metric is the ratio of the absolute difference between estimated and true energy, and the total amount of true energy. In Mayhorn et al. (2016), the authors have shown that with relatively large errors this metric would result in values greater than 1, making it less intuitive and explainable.	$\frac{\sum_{i=1}^{N} \lvert\hat{E}_i - E_i\rvert}{\sum_{i=1}^{N} E_i}$
*EA*v1	The energy accuracy (Mayhorn et al., 2016) was proposed as an attempt to report the energy error between 0 and 1. It was reported in Mayhorn et al. (2016) as a stable metrics, however, it requires the tuning of the α parameter which may not be desirable.	$e^{-\alpha(EE)}$
MR	The match rate (Mayhorn et al., 2016, 2017) is a metric where the evaluation is based on the overlapping rate of true and estimated energy. It varies between 0 and 1. As the value tends to 1, the metric indicates a strong match between the estimated and the true energy. On the contrary, a value tending to 0 indicates a poor match. A value of zero is only possible if the true and estimated energy are zero. This was the metric that demonstrated best overall performance in Mayhorn et al. (2016).	$\frac{\sum_{i=1}^{N} min\{E_i, \hat{E}_i\}}{\sum_{i=1}^{N} max\{E_i, \hat{E}_i\}}$
SEM	The standard error of the mean (Kalla, 2017; Mayhorn et al., 2016) reports on how the mean varies with different experiments measuring the same quantity. In the case of energy estimation, if there are significant errors, the *SEM* will be higher. On the contrary, if there are few to no significant errors, the *SEM* will tend to zero.	$\frac{\sigma}{\sqrt{N}}$
FEE	The fraction of energy explained (Berges, 2010; Hart, 1985) is the ratio between the total estimated energy and actual energy used. In Berges (2010) this metric is referred to as energy identification rate (*EIR*).	$\frac{\sum_{i=1}^{N} \hat{E}_i}{\sum_{i=1}^{N} E_i}$
TECA	The Total energy correctly assigned (Johnson & Willsky, 2013; Kolter & Matthew, 2011; Makonin & Popowich, 2014) is the total error in assigned energy, normalized by the actual energy consumption in each time slice averaged over all appliances.	$\frac{1 - \sum_{i=1}^{N}\sum_{a=1}^{A} \lvert\hat{E}_i^{(a)} - E_i^{(a)}\rvert}{2\sum_{i=1}^{N} E_i}$
ETEA	The error in Total energy assigned (Batra, Kelly, et al., 2014) is the difference between the total energy assigned, and the actual energy consumed by a given appliance over the dataset.	$\left\lvert \sum_{i=1}^{N} \hat{E}_i^{(a)} - \sum_{i=1}^{N} E_i^{(a)} \right\rvert$
Dev	The deviation (Beckel et al., 2014) determines the deviation of the inferred electricity consumption from the actual electricity consumption over the dataset. It is the ratio between the *ETEA* and the actual energy consumed.	$\frac{ETEA}{\sum_{i=1}^{N} E_i^{(a)}}$
FTEAC	The fraction of Total energy assigned correctly (Batra, Kelly, et al., 2014) is the overlap between the fraction of energy assigned to each appliance and the actual fraction of energy consumed by each appliance over the dataset.	$\sum_a min\left(\frac{\sum_t E_t^{(a)}}{E}, \sum_t \hat{E}_t^{(a)} \hat{E}\right)$
T	Number of observations or events based on each time step	
A	Number of appliances being considered	
E_t	Metered energy at time interval t	
$E_t^{(a)}$	Metered energy for appliance a at interval t	
$\overline{E}$	Average metered energy over the dataset	
$\hat{E}_t$	Estimated energy at instant t	
$\hat{E}_t^{(a)}$	Estimated energy for appliance a at instant t	
ΔE_t	Error between the NILM and metered data at instant t	
$\overline{\Delta E}$	Error between the NILM and metered data over the entire dataset	
α	Mapping factor, defined to be 1.4	
σ	Standard deviation of the error over the dataset	

The dataset metadata schema is used to model the individual components that comprise the dataset. The modeled objects include: (a) electricity meters (whole-home and individual appliance meters); (b) domestic appliances; (c) mapping between meters and appliances; (d) buildings (e.g., appliances, number of rooms, and a description of the occupants); and (e) datasets (e.g., name, authors, and geographical location, etc.).

The NILM metadata framework is currently part of the NILMTK project, where it is used to provide structured information about the available datasets, and to define the rules to create datasets that are compatible with NILMTK. This is an open-source project, and the source code is available on Github.[1] As of this writing, the last update was in April 2017.

4.2 | Energy monitoring and disaggregation data format

The energy monitoring and disaggregation data formant (EMD-DF), is a common data model and file format to support the creation and manipulation of energy disaggregation datasets.

The EMD-DF data model defines three main data entities that should be present in a dataset for energy disaggregation: (a) *consumption data*, (b) *ground-truth data*, and (c) *data annotations.*

The *consumption data* entity represents all the data elements that refer to energy consumption. Consumption data can be of two different types: (a) *raw waveforms*, that is, current and voltage; or (b) *processed waveforms*, that is, different power metrics like real and reactive power.

The *ground-truth data* entity refers to all ground-truth elements and can be of four different types: (a) *individual appliance consumption*; (b) *individual circuit consumption*; (c) *appliance activity*; and (d) *user activity.* Individual appliance and individual circuit consumption data are mostly used in event-less approaches. Appliance activities provide information about power events, and are required in event-based approaches. Finally, user activities refer to actions that people perform involving the use of electric appliances, for example, doing the laundry (washer, dryer, and iron).

Lastly, there is the *data annotations* entity. These can be either metadata or general comments. EMD-DF defines three different types of metadata annotations, namely: (a) *local metadata*, which refers to specific samples in the consumption data; (b) *custom metadata* that are defined by the dataset creator and can serve multiple purposes (e.g., include NILM Metadata schemas); and (c) *RIFF Metadata*, which is composed by the metadata chunks defined by the resource interchange file format (RIFF[2]).

The current implementation of the EMD-DF is an extension of the well-known waveform audio file format (WAVE[3]) that was originally created to store audio data. WAVE is an application of the RIFF standard in which the file contents are grouped and stored in separate chunks, each of which following a predefined format. As an example of the application of EMD-DF, the authors have converted the BLUED dataset to their format. The original BLUED distribution is comprised of over 6,500 files, and takes about 320 GB of disk space. The converted version contains only 34 files and takes less than 70 GB, which represents a decease in storage space of about 80%.

The current version of EMD-DF is implemented in Java, and according to the author, the current version of the code is available on Gitlab.[4]

5 | NILMTK AND NILMTK V0.2

The NILMTK is an open source toolkit, released in April 2014, to enable the analysis of existing datasets and algorithms. It also provides a unified interface for the addition of new datasets and algorithms.

In order to represent the datasets in NILMTK, the authors defined a common data format (NILMTK-DF). The data is stored using the hierarchical data format (HDF5) and is loaded to the working memory in small batches. In addition to storing electricity data, NILMTK-DF also stores relevant metadata according to the NILM-metadata framework, as well as other sensor modalities such as gas, water, and temperature.

In terms of features, the toolkit is composed of several software components written in Python, including: (a) dataset converters and parsers; (b) function for dataset diagnosis (e.g., gap and dropout rate detection); (c) statistical analysis (e.g., proportion of submetered energy); and (d) data preprocessing (e.g., down-sampling, and voltage normalization). Additionally, the NILMTK implements two reference benchmark disaggregation algorithms (combinatorial optimization—CO, proposed by Hart (1985), and factorial hidden Markov models (Kim et al., 2011; Kolter & Matthew, 2011)), as well a number performance metrics (e.g., *EE*, *ETEA*, and *FTEAC*).

In order to assess the feasibility of their toolkit, the authors performed some evaluations, including several dataset statistical analysis and energy disaggregation benchmarks. Regarding the benchmarks, the two default benchmark algorithms were tested against six datasets (REDD, Smart*, PSRI - now DataPort, AMPds, iAWE, and the UK-Dale) at 1-minute resolution. The obtained results indicated that FHMM performance was superior to CO in three datasets (REDD, Smart*, and AMPds), while for the remaining three datasets CO and FHMM performed similarly.

The source code of NILMTK is available on Github.[5] As for this writing, the project includes contributions from 17 individuals, and the last update to source code was on March 2018. This suggests that the project is still attracting substantial attention, and there is a relative interest from the community around this project.

5.1 | NILM-Eval

The NILM-Eval is a Matlab-based open source framework for running comprehensive performance evaluations of NILM algorithms across multiple datasets. NILM-Eval is very similar in scope to the NILMTK in the sense that it allows evaluations across multiple datasets with common performance metrics. Yet it was designed to facilitate the design and execution of large experiments that consider several different parameter settings for the different algorithms in repeated experiments, therefore enabling the quick evaluation and benchmark of such algorithms under different settings.

The authors of NILM-Eval have also thoroughly tested their systems' ability to evaluate and benchmark disaggregation algorithms. To this end, they used their own dataset (ECO) to evaluate four different algorithms, two of them event-based (Baranski & Voss, 2004) and (Weiss, Helfenstein, Mattern, & Staake, 2012) and two event-less (Kolter & Jaakkola, 2012; Parson, Ghosh, Weal, & Rogers, 2012). These algorithms were tested under different parameter configurations, and the results reported using the system default performance metrics (*P*, *R*, *RMSE*, and *Dev*). An extensive discussion of the evaluation results is out of the scope of this paper, yet these have shown that the event-based approaches performed better than the event-less counterparts.

The source code of NILM-Eval is available on Github[6] but it naturally requires the commercial and expensive MATLAB package in order to be used. This may prevent the adoption of this framework by other researchers. In fact, this project has only one contributor, and the last update dates back to June 2015. This suggests that the project is no longer being supported by the author.

6 | LIMITATIONS AND CHALLENGES

So far in this paper, we reviewed the public datasets, performance metrics, and tools that are commonly used to assess the performance of NILM algorithms and systems. We now discuss the shortcomings and main challenges of the current state-of-the-art.

6.1 | Datasets

Regarding datasets, there are a number of limitations that make performance evaluation challenging: (a) missing data; (b) limited labeling; and (c) substantial differences in the available data.

By missing data, we do not refer only to the gaps that are very common in many of the existing datasets (Batra, Kelly, et al., 2014; Batra, Parson, et al. 2014), but also to the loads for which there are no submetered data. This happens due to many reasons, including the impracticality of installing plug-level meters in each and every load because they do not have a fixed plug (e.g., vacuum cleaners) and also the fact that some loads cannot be monitored using plugs (e.g., ceiling lights) (Pereira, Ribeiro, & Nunes, 2017). This challenge is particularly relevant when monitoring commercial/industrial spaces, where the number of individual loads can easily reach 100 or more (Jahromi, 2018).

One of the alternatives to increase the level of submetered data is by using circuit-level submetering as suggested in Jahromi (2018). This was already accomplished in datasets like AMPds (Makonin et al., 2016) and RAE (Makonin et al., 2018). However, since many loads can be attached to the same circuit (e.g., the kitchen circuit contains many different appliances), this solution alone will not guarantee that individual consumption is available for every load. Consequently, if the goal is to monitor individual loads, plug-level submetering is still required.

Alternatively, indirect sensing could be used to infer individual appliance activity that could then complement the circuit-level data. A similar approach was done for BLUED (Anderson, Bergés, et al., 2012), where environmental sensors were used to monitor ceiling lights. This solution, however, implies a considerable amount of postprocessing work in order to combine the different sensor streams, and subject to several challenges to ensure the correct synchronization of the internal clocks in each monitoring system (Anderson, 2014).

This brings us to the second limitation, the limited amount of labeled data. This limitation is particularly relevant in event-based approaches that require labeled transitions. In fact, the difficulty in labeling NILM datasets justifies the reduced number of event-based datasets. Producing fully labeled NILM datasets requires either the deployment of plug-level hardware and/or the lengthy, and error-prone manual inspection of the whole dataset. This process of labeling/annotating sensor data is transversal to many domains of machine learning, leading to a whole research community such as the International Workshop on Annotation of use R Data for UbiquitOUs Systems.[7]

In the context of NILM systems, on the one hand, the process cannot be fully automated because of the aforementioned issues with plug-level metering. On the other hand, a fully manual system would require an outstanding level of resources and still be prone to errors. Against this background, we believe that future work should look at ways of leveraging the potential of the existing data, for example by means of collaborative and semi-automatic annotation of datasets (Cao, Wijaya, Aberer, &

Nunes, 2015; Pereira et al., 2017; Pereira & Nunes, 2015), or even the creation of synthetic datasets by means of data synthesizing (Buneeva & Reinhardt, 2017; Henriet, Simsekli, Fuentes, & Richard, 2018).

A third limitation that ultimately prevents fair cross-dataset NILM benchmarks is the wide differences among the available datasets. These differences occur because of two main aspects: (a) the type and granularity of the available measurements; and (b) the different formats (e.g., plain-text, CSV, audio files, etc.) in which the data is made available to the research community.

While some research efforts provide common interfaces to access datasets, very little work has been carried out so far toward understanding the real implications of such differences. For example, it is not possible to quickly identify which approaches can be tested in a particular dataset, or which metrics can be calculated from the underlying data. Thus, in the near future it would be of crucial importance to find a scalable and easy to understand method to describe this dataset-algorithm-metric nexus. We argue that one way of doing so is by extending the NILM metadata project with meta-information about the algorithms and metrics that are supported by each dataset. Something is those lines is already done in the UCI machine learning repository (Dheeru & Karra Taniskidou, 2017), where it is possible to select datasets for a particular task (i.e., classification, regression, clustering, and other).

Finally, the wide differences in the existing datasets give rise to one of the most interesting challenges of NILM performance evaluation—the need to define the complexity of the NILM problem in each dataset (and respective subsets). For example, in Egarter, Pöchacker, and Elmenreich (2015), the authors propose two complexity measures, *Appliance Set Complexity*, and *Time Series Disaggregation Complexity*. Early results show considerable differences between datasets, which makes direct comparisons very difficult even when using the same performance metric (Egarter et al., 2015; Nalmpantis & Vrakas, 2018).

Consequently, while defining additional complexity metrics may be a more appealing research direction, we argue that in the present the research community could greatly benefit from having the ability to generate set of data with comparable complexities using the measures proposed in Egarter et al. (2015). For example, since the proposed measures are independent of the disaggregation algorithms (Egarter et al., 2015), this could be easily achieved by subsetting existing datasets until the same complexity scores are achieved. Still, while such an approach may be feasible, it should be taken into consideration that the resulting datasets should always be representative of the real-world conditions. Otherwise, the disaggregation results will be directly comparable, but not meaningful.

6.2 | Performance metrics

In terms of performance metrics, this review highlighted that the research community is still far from a consensus regarding which metrics should be used to assess and report the performance of NILM technology. We propose that in order to establish a consistent set of performance metrics, two main limitations must be addressed by the NILM research community: (a) the lack of a deep understanding on the behavior of traditional metrics when applied to NILM; and (b) the almost complete absence of domain/application specific metrics.

Regarding the former, most of the currently existing metrics have been inherited or extended from other application domains in machine-learning. For example, the precision and recall metrics that are widely used in ED problems have their origins in the information retrieval domain (Kagolovsky & Moehr, 2003). Consequently, future research should aim at fully understanding the behavior of the existing performance metrics when applied to the energy disaggregation problem. For example, in Pereira and Nunes (2017, 2018), the authors analyzed experimentally the behavior of several performance metrics when applied to classification and ED algorithms. Their results show very high correlations between the metrics in the classification problem, contrasting the much lower correlation values in the ED problem.

While the high correlations in classification problems appears to be in line with the relevant literature (e.g., Ferri, Hernández-Orallo, & Modroiu, 2009), ED proves to be a very distinct problem in part due to the highly unbalanced nature of the underlying data toward true negatives.

As such, future research should look at understanding how the characteristics of NILM data affect the mathematical properties of the different performance metrics (e.g., Sokolova & Lapalme, 2009). This knowledge combined with empirical evidence will be useful to fully understand the relationships between performance metrics, and ascertain to what extent the results and conclusions obtained using one particular metric can be extended to others.

Likewise, future work should also look at other metrics, in particular metrics that balance precision and recall (e.g., mean average precision, and break-even point; Manning, Raghavan, & Schütze, 2008), handle multiclass classification with unbalanced data (e.g., multiclass performance score (Kautz, Eskofier, & Pasluosta, 2017), and Mathews correlation coefficient (Matthews, 1975; Powers, 2011)), and the Jaccard index (Labatut & Cherifi, 2012; Powers, 2011; Real & Vargas, 1996), which can be used as a metric for ED/classification and EE.

Regarding the second limitation, the work on Pereira and Nunes (2018) briefly touched the topic of domain specific metrics, highlighting their potential to unveil important characteristics of the algorithms and the underlying datasets.

Consequently, while future work should aim at further understanding these metrics, there is also an opportunity to define new metrics. This is particularly relevant in the case of metrics that take into consideration concepts from cost-sensitive learning (Elkan, 2001), since it is well known that not all appliances contribute the same for the final EE (Anderson, Ocneanu, et al., 2012). For instance, missing many power events from a smaller appliance may be less costly than missing just one from a larger appliance. Thus it should have a smaller cost associated. On the contrary, an appliance that is rarely used should have a higher cost associated, since failing to correctly identify the few occurrences of that appliance will result in significant underestimations of its consumption.

Nevertheless, while computing the value of cost-sensitive metrics can be easily achieved by means of techniques such as ceiling analysis (e.g., Roncancio, Hernandes, & Becker, 2013), defining the correct values for the cost is not straightforward since it is highly dependent on the application domain. For example, if NILM is used to detect malfunctioning loads by studying the periodicity of the power events, there should be a high penalty for missed and erroneously detected events. Consequently, in order to define cost-sensitive metrics (and any other meaningful domain specific metrics), the different stakeholders must first clearly identify their NILM use-cases. Only then, the metrics that make sense in each case can be identified, as well as the performance thresholds that should be achieved in order to validate each use-case.

6.3 | Tools and frameworks

With respect to frameworks and toolkits, we argue that important research breakthroughs can only be achieved if the research community manages to continuously and incrementally integrate contributions in a common framework. For instance, integrating EMD-DF in NILMTK would add to the latter the support to evaluate the performance of high-frequency NILM approaches. Likewise, integrating EMD-DF or NILMTK-DF into the existing dataset simulators (Buneeva & Reinhardt, 2017; Henriet et al., 2018), would ensure that the generated datasets are immediately ready to integrate with NILMTK or other platform that interfaces with these data formats.

Additionally, as the smart-grid becomes a reality, one should expect more datasets to emerge, some of which with a data granularity in the order of several kHz. As such, framework and toolkit developers should look at new techniques to compress, store and represent NILM data. For example, in Reinhardt (2017), the author proposes the creation of hybrid load signatures by representing data at different sampling rates (e.g., 1 Hz for steady-state, and several kHz for transients). Although this was originally proposed to reduce the bandwidth in communication channels, if integrated with the existing data formats, it can work as custom data compression tool for NILM datasets.

Lastly, NILM research can greatly benefit from the development of an online platform for performance evaluation. Among other benefits, this would guarantee that the proposed approaches are evaluated under the same conditions and enable the community to keep track of research progress. The idea of an online NILM platform was briefly mentioned as future work in Kelly (2016), and in the present emuNILM (Makonin, 2018) is the most recent effort toward creating such a platform. The main goal of emuNILM is to be able to emulate smart-meters and communication infrastructures so that NILM algorithms can be tested in real-world like situations. For example, by emulating the consumption of broken appliances it will be possible to understand how a particular algorithm responds in the presence of malfunctioning loads.

7 | CONCLUSION

Energy disaggregation is a key cost-effective technology to monitor energy consumption and contribute to the many challenges faced by the transition to a reliable, sustainable, and competitive energy system. NILM is a maturing field showing promising results in measuring appliance-specific energy consumption, while keeping installation cost and hardware complexity low. Even if IoT technologies increase the availability of disaggregated data, legacy infrastructures will take decades to catch up and NILM seems the only viable alternative for practical and cost-effective energy disaggregation. Consequently, performance evaluation will continue to be a fundamental research direction in NILM, as this is the only mean to validate the proposed solutions and leverage the potential adoption of this technology by utilities worldwide at building and city scale.

In this paper, we reviewed the main challenges and advances in this are of NILM research: datasets, performance metrics, and frameworks/toolkits. Our goals of this paper are to: (1) motivate the importance of proper performance evaluation; (2) provide a solid background and comprehensive knowledge of the current state-of-the-art; and (3) outline the main challenges and future research opportunities.

To conclude, we should mention that in this paper we have only looked at performance evaluation from an algorithmic perspective. As such, it is important to point out the need to conduct research toward evaluating the value proposition of NILM technology, as this is only seldom reported in the literature. For example, it was until only recently that we saw the first publications regarding the assessment of the value proposition of disaggregated data as a tool to reduce energy consumption (Batra, Singh, & Whitehouse, 2015; Kelly & Knottenbelt, 2016), or trying to educate the research community about the practical issues of deploying such systems in real-world scenarios (Kosonen & Kim, 2016; Pereira, 2016). This, we believe, is of crucial importance to the large-scale adoption of NILM technology in years to come, and presents the perfect opportunity to conduct multidisciplinary research in the field of NILM.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

RELATED WIREs ARTICLES

Research with disaggregated electricity end-use data in households: review and recommendations

NOTES

[1]NILM Metadata Source Code, https://github.com/nilmtk/nilm_metadata

[2]RIFF file format, http://fileformats.archiveteam.org/wiki/RIFF

[3]WAVE file format: http://fileformats.archiveteam.org/wiki/WAV

[4]EMD-DF Source Code, https://gitlab.com/alspereira/EMD-DF

[5]NILMTK Source Code, https://github.com/nilmtk/nilmtk

[6]NILM-Eval Source Code, https://github.com/beckel/nilm-eval

[7]ARDUOUS Workshop, https://text2hbm.org/arduous

REFERENCES

Abubakar, I., Khalid, S. N., Mustafa, M. W., Shareef, H., & Mustapha, M. (2016). Recent approaches and applications of non-intrusive load monitoring. *ARPN Journal of Engineering and Applied Sciences*, *11*(7), 4609–4618. Retrieved from https://pure.utm.my/en/publications/recent-approaches-and-applications-of-non-intrusive-load-monitori

Anderson, K. (2014). *Non-intrusive load monitoring: Disaggregation of energy by unsuper-vised power consumption clustering* (PhD thesis). Carnegie Mellon University, Pittsburgh, PA. Retrieved from http://repository.cmu.edu/dissertations/507

Anderson, K., Ocneanu, A., Benitez, D., Carlson, D., Rowe, A., & Berges, M. (2012). BLUED: A fully labeled public dataset for event-based non-intrusive load monitoring research. In *2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD). Beijing, China.*

Anderson, K. D., Bergés, M. E., Ocneanu, A., Benitez, D., & Moura, J. M. F. (2012). Event detection for non intrusive load monitoring. In *IECON 2012—38th Annual Conference on IEEE Industrial Electronics Society* (pp. 3312–3317). https://doi.org/10.1109/IECON.2012.6389367

Bache, K., & Lichman, M. (2013). *Individual household electric power consumption dataset*. Irvine, CA: University of California, School of Information and Computer Science. Retrieved from http://archive.ics.uci.edu/ml

Baets, L. D., Develder, C., Dhaene, T., Deschrijver, D., Gao, J., & Berges, M. (2017). Handling imbalance in extended PLAID dataset. In *Proceedings of the Fifth IFIP Conference on Sustainable Internet and ICT for Sustainability*. Funchal, Portugal: IEEE/IFIP.

Baranski, M., & Voss, J. (2004). Genetic algorithm for pattern detection in NIALM systems. In *2004 I.E. International Conference on Systems, Man and Cybernetics* (Vol. 4, pp. 3462–3468). IEEE. https://doi.org/10.1109/ICSMC.2004.1400878

Barker, S., Mishra, A., Irwin, D., Cecchet, E., & Shenoy, P. (2012). Smart*: An open data set and tools for enabling research in sustainable homes. In *Data Mining Applications in Sustainability (SustKDD)*.

Barsim, K., Mauch, L., & Yang, B. (2016). *Neural network ensembles to real-time identification of plug-level appliance measurements.* Vancouver, BC, Canada. Retrieved from https://www.researchgate.net/publication/301771233_Neural_Network_Ensembles_to_Real-time_Identification_of_Plug-level_Appliance_Measurements

Barsim, K., & Yang, B. (2018). On the feasibility of generic deep disaggregation for single-load extraction. In *4th International NILM Workshop.*

Batra, N., Gulati, M., Singh, A., & Srivastava, M. B. (2013). It's different: Insights into home energy consumption in India. In *Proceedings of the 5th ACM Workshop on Embedded Systems for Energy-Efficient Buildings* (pp. 3:1–3:8). New York, NY: ACM. Retrieved from http://doi.acm.org/10.1145/2528282.2528293

Batra, N., Kelly, J., Parson, O., Dutta, H., Knottenbelt, W., Rogers, A., … Sri-vastava, M. (2014). NILMTK: An open source toolkit for non-intrusive load monitoring. In *Proceedings of the 5th International Conference on Future Energy Systems* (pp. 265–276). New York, NY: ACM. Retrieved from http://doi.acm.org/10.1145/2602044.2602051

Batra, N., Parson, O., Berges, M., Singh, A., & Rogers, A. (2014). *A comparison of non-intrusive load monitoring methods for commercial and residential buildings.* arXiv: 1408.6595 [cs]. Retrieved from http://arxiv.org/abs/1408.6595

Batra, N., Singh, A., & Whitehouse, K. (2015). If you measure it, can you improve it? Exploring the value of energy disaggregation. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments* (pp. 191–200). New York, NY: ACM. Retrieved from http://doi.acm.org/10.1145/2821650.2821660

Beckel, C., Kleiminger, W., Cicchetti, R., Staake, T., & Santini, S. (2014). The ECO data set and the performance of non-intrusive load monitoring algorithms. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-efficient Buildings* (pp. 80–89). New York, NY: ACM. Retrieved from http://doi.acm.org/10.1145/2674061.2674064

Berges, M. (2010). *A framework for enabling energy-aware facilities through minimally-intrusive approaches* (Ph.D. thesis), Carnegie Mellon University. Retrieved from http://gradworks.umi.com/34/38/3438459.html

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

Buneeva, N., & Reinhardt, A. (2017). AMBAL: Realistic load signature generation for load disaggregation performance evaluation. In *Proceedings of 2017 I.E. International Conference on Smart Grid Communications*. Dresden, Germaby: IEEE.

Butner, R. S., Reid, D. J., Hoffman, M. G., Sullivan, G., & Blanchard, J. (2013). *Non-intrusive load monitoring assessment: Literature review and laboratory protocol* (Tech. Rep. No. PNNL-22635). Richland, WA: Pacific Northwest National Laboratory (PNNL). Retrieved from http://www.osti.gov/scitech/biblio/1095438

Cao, H. A., Wijaya, T. K., Aberer, K., & Nunes, N. (2015). A collaborative framework for annotating energy datasets. In *2015 I.E. International Conference on Big Data (Big Data)* (pp. 2716–2725). IEEE. https://doi.org/10.1109/BigData.2015.7364072

Carrie Armel, K., Gupta, A., Shrimali, G., & Albert, A. (2013). Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy Policy*, *52*, 213–234. https://doi.org/10.1016/j.enpol.2012.08.062

Cicchetti, R. (2013). *NILM-Eval: Disaggregation of real-world electricity consumption data* (Master's thesis). ETH Zurich, Zurich, Switzerland. Retrieved from http://www.vs.inf.ethz.ch/res/project/eco-data-files/masters thesis chicchetti.pdf

Dheeru, D., & Karra Taniskidou, E. (2017). *UCI machine learning repository.* Retrieved from http://archive.ics.uci.edu/ml

Ecotope Inc. (2014). *Residential building stock assessment: Metering study* (Tech. Rep. No. #E14-283). Seatle, WA: Northwest Energy Efficiency Alliance.

Egarter, D., Pöchacker, M., & Elmenreich, W. (2015). *Complexity of power draws for load disaggregation.* arXiv: 1501.02954 [cs]. Retrieved from http://arxiv.org/abs/1501.02954

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence* (Vol. 2, pp. 973–978). San Francisco, CA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=1642194.1642224

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. https://doi.org/10.1016/j.rser.2016.07.009

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38. https://doi.org/10.1016/j.patrec.2008.08.010

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). An overview of ensemble methods for binary classiérs in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, *44*(8), 1761–1776. https://doi.org/10.1016/j.patcog.2011.01.017

Gao, J., Giri, S., Kara, E. C., & Bergés, M. (2014). PLAID: A public dataset of high-resoultion electrical appliance measurements for load identification research: Demo abstract. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings* (pp. 198–199). New York, NY: ACM. https://doi.org/10.1145/2674061.2675032

Gao, J., Kara, E. C., Giri, S., & Bergés, M. (2015). A feasibility study of automated plug-load identification from high-frequency measurements. In *2015 I.E. Global Conference on Signal and Information Processing (GlobalSIP)* (pp. 220–224). IEEE. https://doi.org/10.1109/GlobalSIP.2015.7418189

Gisler, C., Ridi, A., Zufferey, D., Khaled, O. A., & Hennebert, J. (2013). Appliance consumption signature database and recognition test protocols. In *2013 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA)* (pp. 336–341). https://doi.org/10.1109/WoSSPA.2013.6602387

Gulati, M., Ram, S. S., & Singh, A. (2014). An in depth study into using EMI signatures for appliance identification. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-efficient Buildings* (pp. 70–79). New York, NY: ACM. https://doi.org/10.1145/2674061.2674070

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

Hart, G. (1992). Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, *80*(12), 1870–1891. https://doi.org/10.1109/5.192069

Henriet, S., Simsekli, U., Fuentes, B., & Richard, G. (2018). *A generative model for non-intrusive load monitoring in commercial buildings.* arXiv:1803.00515 [cs]. Retrieved from http://arxiv.org/abs/1803.00515

Holcomb, C. (2012). *Pecan Street Inc.: A Test-bed for NILM* (Invited Talk). Pittsburgh, PA. Retrieved from http://www.ices.cmu.edu/psii/nilm/

Holmes, C. (2014). *Non-intrusive load monitoring (NILMS) research activity: End-use energy efficiency & demand response*. Austin, TX: EPRI.

Household electricity survey a study of domestic electrical product usage (Tech. Rep. No. R66141). (2012). Leatherhead, England: Intertek Testing & Certification Ltd.

Iba, H., Hasegawa, Y., & Paul, T. K. (2009). *Applied genetic programming an0064 machine learning* (1st ed.). Boca Raton, FL: CRC Press.

Jahromi, O. (2018). Real-time itemized electricity consumption intelligence for military bases. In *4th International NILM Workshop. Austin, TX*.

Jiang, L., Li, J., Luo, S., Jin, J., & West, S. (2011). Literature review of power disaggregation. In *Proceedings of 2011 International Conference on Modelling, Identification and Control (ICMIC)* (pp. 38–42). IEEE. https://doi.org/10.1109/ICMIC.2011.5973672

Johnson, G., & Beausoleil-Morrison, I. (2017). Electrical-end-use data from 23 houses sampled each minute for simulating micro-generation systems. *Applied Thermal Engineering*, *114*, 1449–1456. https://doi.org/10.1016/j.applthermaleng.2016.07.133

Johnson, M. J., & Willsky, A. S. (2013). Bayesian nonparametric hidden semi-Markov models. *Journal of Machine Learning Research*, *14*(1), 673–701. Retrieved from http://dl.acm.org/citation.cfm?id=2502581.2502602

Kagolovsky, Y., & Moehr, J. R. (2003). Current status of the evaluation of information retrieval. *Journal of Medical Systems*, *27*(5), 409–424. https://doi.org/10.1023/A:1025603704680

Kahl, M., Ui Haq, A., Kriechbaumer, T., & Hans-Arno, J. (2016). WHITED—a worldwide household and industry transient energy data set. In *3rd International NILM Workshop. Vancouver, BC, Canada*.

Kaibin Bao. (2016). *Preparations to publish the energy smart home lab dataset* (Workshop presentation). London, England.

Kalla S.. (2017). *Standard error of the mean—An estimate of the standard deviation* (Research notes). Retrieved from https://explorable.com/standard-error-of-the-mean

Kautz, T., Eskofier, B. M., & Pasluosta, C. F. (2017). Generic performance measure for multiclass-classifiers. *Pattern Recognition*, *68*, 111–125. https://doi.org/10.1016/j.patcog.2017.03.008

Kelly, D. (2016). *Disaggregation of domestic smart meter energy data* (Ph.D. thesis). London, England: Imperial College London. Retrieved from http://spiral.imperial.ac.uk/handle/10044/1/49452

Kelly, J., Batra, N., Parson, O., Dutta, H., Knottenbelt, W., Rogers, A., … Srivastava, M. (2014). NILMTK V0.2: A non-intrusive load monitoring toolkit for large scale data sets: Demo abstract. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-efficient Buildings* (pp. 182–183). New York, NY: ACM. https://doi.org/10.1145/2674061.2675024

Kelly, J., & Knottenbelt, W. (2014). *Metadata for energy disaggregation.* arXiv:1403.5946 [cs]. Retrieved from http://arxiv.org/abs/1403.5946

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*, 150007. https://doi.org/10.1038/sdata.2015.7

Kelly, J., & Knottenbelt, W. (2016). *Does disaggregated electricity feedback reduce domestic electricity consumption? A systematic review of the literature.* arXiv: 1605.00962 [cs]. Retrieved from http://arxiv.org/abs/1605.00962

Kim, H., Marwah, M., Arlitt, M., Lyon, G., & Han, J. (2011). Un-supervised disaggregation of low frequency power measurements. In *Proceedings of the 2011 SIAM International Conference on Data Mining* (pp. 747–758). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972818.64

Kolter, Z., & Jaakkola, T. (2012). Approximate inference in additive factorial HMMs with application to energy disaggregation. In *JMLR: W&CP 22* (Vol. 22, pp. 1472–1482). La Palma, Canary Islands. Retrieved from http://people.csail.mit.edu/kolter/lib/exe/fetch.php?media=pubs:kolter-aistats12.pdf

Kolter, Z., & Matthew, J. (2011). REDD: A public data set for energy disaggregation research. In *Data Mining Applications in Sustainability (SustKDD)*.

Kosonen, H., & Kim, A. (2016). Quantifying plug load energy use in a LEED gold building? Lessons learned in the installation phase. *Construction Research Congress*, *2016*, 180048. Retrieved from https://ascelibrary.org/doi/abs/10.1061/9780784479827.124

Kriechbaumer, T., & Jacobsen, H.-A. (2018). BLOND, a building-level office environment dataset of typical electrical appliances. *Scientific Data*, *5*, 180048. https://doi.org/10.1038/sdata.2018.48

Labatut, V., & Cherifi, H. (2012). *Accuracy measures for the comparison of classifiers.* arXiv:1207.3790 [cs]. Retrieved from http://arxiv.org/abs/1207.3790

Lai, P.-H., Trayer, M., Ramakrishna, S., & Li, Y. (2012). Database establishment for machine learning in NILM. In *1st International NILM Workshop*.

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*(2), 551–560. https://doi.org/10.1109/TPWRD.2009.2033799

Makonin, S. (2012). *Approaches to non-intrusive load monitoring (NILM) in the Home* (Tech. Rep.). Retrieved from http://summit.sfu.ca/item/14475

Makonin, S. (2018). An emulator for NILM and smart home research. In *4th International NILM Workshop*.

Makonin, S., Ellert, B., Bajić, I. V., & Popowich, F. (2016). Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014. *Scientific Data*, *3*, 160037. https://doi.org/10.1038/sdata.2016.37

Makonin, S., & Popowich, F. (2014). Nonintrusive load monitoring (NILM) performance evaluation. *Energy Efficiency*, *1–6*. https://doi.org/10.1007/s12053-014-9306-2

Makonin, S., Wang, Z. J., & Tumpach, C. (2018). RAE: The rainforest automation energy dataset for smart grid meter data analysis. *Data*, *3*(1), 8. https://doi.org/10.3390/data3010008

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (1st ed.). New York, NY: Cambridge University Press. Retrieved from https://nlp.stanford.edu/IR-book/

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)—Protein Structure*, *405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

Mayhorm, E., Butner, R., Baechler, M., Sullivan, G., & Hao, H. (2015). *Characteristics and performance of existing load disaggregation technologies* (Tech. Rep. No. PNNL-24230). Richland, WS: Pacific Northwest National Laboratory.

Mayhorn, E. T., Sullivan, G. P., Fu, T., & Petersen, J. M. (2017). *Non-intrusive load monitoring laboratory-based test protocols* (Tech. Rep. No. 26184). Richland, WS: Pacific Northwest National Laboratory (PNNL).

Mayhorn, E. T., Sullivan, G. P., Petersen, J. M., Butner, R. S., & Johnson, E. M. (2016). *Load disaggregation technologies: real world and laboratory performance* (Tech. Rep. No. PNNL-SA-116560). Richland, WA: Pacific Northwest National Laboratory (PNNL). Reetrieved from https://www.osti.gov/scitech/biblio/1334878

Monacchi, A., Egarter, D., Elmenreich, W., D'Alessandro, S., & Tonello, A. M. (2014). GREEND: An energy consumption dataset of households in Italy and Austria. In *2014 I.E. International Conference on Smart Grid Communications (SmartGridComm)*. IEEE. https://doi.org/10.1109/SmartGridComm.2014.7007698

Murray, D., Liao, J., Stankovic, L., Stankovic, V., Hauxwell-Baldwin, R., Wilson, C., … Firth, S. (2015). *A data management platform for personalised real-time energy feedback.*

Nalmpantis, C., & Vrakas, D. (2018). Machine learning approaches for non-intrusive load monitoring: From qualitative to quantitative comparation. *Artificial Intelligence Review*, 1–27. https://doi.org/10.1007/s10462-018-9613-7

Nau, F. (2017). *Percent of standard deviation explained* (Course material). Retrieved from http://people.duke.edu/~rnau/rsquared.htm#percentexplained

Parson, O., Ghosh, S., Weal, M., & Rogers, A. (2012). Non-intrusive load monitoring using prior models of general appliance types. In *Proceedings of theTwenty-Sixth Conference on Artificial Intelligence (AAAI-12) Proceedings of theTwenty-Sixth Conference on Artificial Intelligence (AAAI-12)* (pp. 356–362). Retrieved from http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/4809

Parson, O., Mark, W., & Rodgers, A. (2014). *A scalable non-intrusive load monitoring system for fridge-freezer energy efficiency estimation.*

Pereira, L. (2016). *Hardware and software platforms to deploy and evaluate non-intrusive load monitoring systems* (PhD theis). Universidade da Madeira, Funchal, Portugal.

Pereira, L. (2017). EMD-DF: A data model and file format for energy disaggregation datasets. In *Proceedings of the 4th ACM International Conference on Systems for Energy-efficient Built Environments*. Delft, The Netherlands: ACM.

Pereira, L., & Nunes, N. (2018). An experimental comparison of performance metrics for event detection algorithms in NILM. In *4th International NILM Workshop*.

Pereira, L., & Nunes, N. J. (2015). Semi-automatic labeling for public non-intrusive load monitoring datasets. In *2015 Sustainable Internet and ICT for Sustainability (SustainIT)* (pp. 1–4). https://doi.org/10.1109/SustainIT.2015.7101378

Pereira, L., & Nunes, N. J. (2017). A comparison of performance metrics for event classification in non-intrusive load monitoring. In *Proceedings of the 2017 I.E. International Conference on Smart Grid Communications*. Dresden, Germany: IEEE.

Pereira, L., Ribeiro, M., & Nunes, N. J. (2017). Engineering and deploying a hardware and software platform to collect and label non-intrusive load monitoring datasets. In *Proceedings of the Fifth IFIP Conference on Sustainable Internet and ICT for Sustainability*. Funchal, Portugal: IEEE/IFIP.

Picon, T., Meziane, M. N., Ravier, P., Lamarque, G., Novello, C., Bunetel, J.-C. L., & Raingeaud, Y. (2016). *COOLL: Controlled on/off loads library, a public dataset of high-sampled electrical signals for appliance identification.* arXiv:1611.05803 [cs]. Retrieved from http://arxiv.org/abs/1611.05803

Powers, D. (2011). Evaluation: From precision, recall and f-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Real, R., & Vargas, J. M. (1996). The probabilistic basis of Jaccard's index of similarity. *Systematic Biology*, *45*(3), 380–385. https://doi.org/10.2307/2413572

Reinhardt, A. (2017). Adaptive load signature coding for electrical appliance monitoring over low-bandwidth communication channels. In *Proceedings of the Fifth IFIP Conference on Sustainable Internet and ICT for Sustainability*. Funchal, Portugal: IEEE/IFIP.

Reinhardt, A., Baumann, P., Burgstahler, D., Hollick, M., Chonov, H., Werner, M., & Steinmetz, R. (2012). On the accuracy of appliance identification based on distributed load metering data. In *2012 Sustainable Internet and ICT for Sustainability (SustainIT)* (pp. 1–9).

Ribeiro, M., Pereira, L., Quintal, F., & Nunes, N. (2016). SustDataED: A public dataset for electric energy disaggregation research. In *Proceedings of ICT for Sustainability 2016* (pp. 244–245). Amsterdam, The Netherlands: Atlantis Press. https://doi.org/10.2991/ict4s-16.2016.36

Ridi, A., Gisler, C., & Hennebert, J. (2014). ACS-F2: A new database of appliance consumption signatures. In *2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR)* (pp. 145–150). https://doi.org/10.1109/SOCPAR.2014.7007996

Roncancio, H., Hernandes, A., & Becker, M. (2013). Ceiling analysis of pedestrian recognition pipeline for an autonomous car application. In *2013 I.E. Workshop on Robot Vision (WORV)* (pp. 215–220). IEEE. https://doi.org/10.1109/WORV.2013.6521941

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Townson, B. (2016). *NILM: Vehicle or destination?* (Pre-sentatin). Vancouver, BC, Canada. Retrieved from http://nilmworkshop.org/2016/slides/Ecotagious.pdf

Uttama Nambi, A. S., Reyes Lua, A., & Prasad, V. R. (2015). LocED: Location-aware energy disaggregation framework. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-efficient Built Environments* (pp. 45–54). New York, NY: ACM. https://doi.org/10.1145/2821650.2821659

Van Asch, V. (2013). *Macro- and micro-averaged evaluation measures [BASIC DRAFT]—Semantic Scholar* (Short paper).

Weiss, M., Helfenstein, A., Mattern, F., & Staake, T. (2012). Leveraging smart meter data to recognize home appliances. In *2012 I.E. International Conference on Pervasive Computing and Communications* (pp. 190–197). IEEE. https://doi.org/10.1109/PerCom.2012.6199866

Wong, Y. F., Ahmet Sekercioglu, Y., Drummond, T., & Wong, V. S. (2013). Recent approaches to non-intrusive load monitoring techniques in residential settings. In *2013 I.E. Symposium on Computational Intelligence Applications in Smart Grid (CIASG)* (pp. 73–79). IEEE. https://doi.org/10.1109/CIASG.2013.6611501

Zeifman, M., & Roth, K. (2011). Nonintrusive appliance load monitoring: Review and outlook. *IEEE Transactions on Consumer Electronics*, *57*(1), 76–84. https://doi.org/10.1109/TCE.2011.5735484

Zoha, A., Gluhak, A., Imran, M. A., & Rajasegarar, S. (2012). Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey. *Sensors*, *12*(12), 16838–16866. https://doi.org/10.3390/s121216838

How to cite this article: Pereira L, Nunes N. Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—A review. *WIREs Data Mining Knowl Discov*. 2018;8:e1265. https://doi.org/10.1002/widm.1265