Artif Intell Rev (2019) 52:217–243
https://doi.org/10.1007/s10462-018-9613-7

Machine learning approaches for non-intrusive load monitoring: from qualitative to quantitative comparation

Christoforos Nalmpantis[1] · Dimitris Vrakas[1]

Published online: 13 January 2018

Abstract Non-intrusive load monitoring (NILM) is the prevailing method used to monitor the energy profile of a domestic building and disaggregate the total power consumption into consumption signals by appliance. Whilst the most popular disaggregation algorithms are based on Hidden Markov Model solutions based on deep neural networks have attracted interest from researchers. The objective of this paper is to provide a comprehensive overview of the NILM method and present a comparative review of modern approaches. In this effort, many obstacles are identified. The plethora of metrics, the variety of datasets and the diversity of methodologies make an objective comparison almost impossible. An extensive analysis is made in order to scrutinize these problems. Possible solutions and improvements are suggested, while future research directions are discussed.

Keywords Non-intrusive load monitoring (NILM) · Power disaggregation algorithms · Hidden Markov Model · Deep learning

1 Introduction

Some of the biggest world challenges such as global warming, acid rains, air and water pollution, depletion, disruption of our natural environment are undoubtedly related to energy. Global energy demand has increased by 16 times in the twentieth century, whereas the population has increased fourfold (Kamat 2007). The governments of many countries have realized the importance of climate change and have made provisions to limit the annual carbon emissions and urban waste by 2050 ("Climate Change Act 2008"). In this direction, traditional fossil fuels expected to be replaced by non-conventional energy sources, whereas current electricity infrastructure will be transformed to a smart grid.

✉ Christoforos Nalmpantis
christofn@csd.auth.gr

Dimitris Vrakas
dvrakas@csd.auth.gr

[1] School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

According to estimates, 40% of the carbon dioxide emissions in the USA relates to energy consumption that comes from electric power (Eia 2017). Residential and commercial buildings account for about 40% of total energy consumption, 20% of which can be saved by applying efficiency improvements (Armel et al. 2013). The significance of these savings is highlighted by the fact that approximately 10–15% of the electricity consumed by US homes requires 200 billion kWh per annum. The equivalent energy production of 16 nuclear power plants or 81.3 million tons of coal (Froehlich et al. 2011).

It is therefore obvious that reducing electricity consumption in buildings can significantly decrease energy wastage. Studies have shown that energy monitoring and direct feedback, such as personalized recommendations or real-time consumption on appliance level, are extremely valuable. They could reduce electricity bills and make residents aware of their house energy profile (Darby 2006).

Power disaggregation is the key to an efficient and accurate energy monitoring in a domestic building. However, the benefits of power disaggregation are not restricted only to residents. Valuable appliance data can empower research and development, help redesigning household appliances and improve building operational efficiency. Utilities and policies will be improved by providing more accurate energy consumption forecasting, more efficient economic models, reformed funding allocations, better energy building evaluation, smart grid optimization etc. (Armel et al. 2013).

The problem of breaking down the total power signal to several appliance level signals, using non-intrusive methods, was firstly introduced by Hart (1992) as non-intrusive load monitoring (NILM). Since then, NILM has been the preferred method for power disaggregation in contrast to other pervasive methods, where appliance level data has to be collected. The reasons that researchers and engineers prefer the NILM method are both economic and practical. This means that NILM research is not only focused on theoretical models, but also on the deployment of these systems in the real world. A large-scale deployment favors NILM over ILM, because it offers lower costs, there is no need for multiple sensor configuration and installation is much simpler. The only advantage of intrusive methods is the high accuracy in measuring energy consumption of specific appliances.

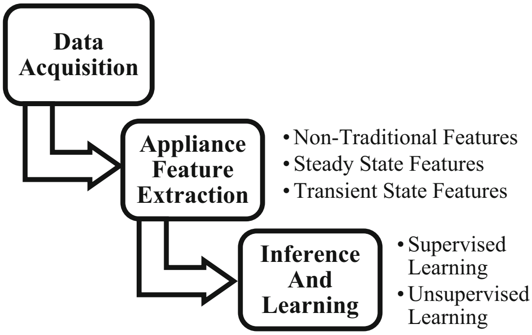
This paper reviews the current status of the NILM research. The Sect. 1 describes the problem of power disaggregation, presenting some formal definitions of the theoretical problem and its complexity. It subsequently describes the NILM framework, referring to relative literature and previous work of researchers. The same section also describes the requirements for a NILM system. The Sect. 3 describes the current state of research in NILM, presenting and evaluating the most recent algorithms. Thereafter, a quantitative analysis of the proposed solutions is presented. In the end, conclusions are drawn.

2 Power disaggregation

The purpose of power disaggregation is to break the total power drawn down into its components. In a domestic building, the resultant power is the outcome of the power consumption of each electrical device. Thus, the problem is to identify how much power each appliance consumes. The superimposition of the power of N devices in a time period T, can be defined as:

$$P(t) = P_{noise}(t) + \sum_{i=1}^{N} p_i(t), t\epsilon \{1, T\} \tag{1}$$

Fig. 1 NILM framework



where p_i is the power of each appliance and P_{noise} is the power of unwanted signal, defined as noise.

Many different approaches have been formulated for solving the problem of power disaggregation. The most common one is to estimate the p_i for $i = 1, 2, \ldots, N$, given only the $P(t)$. From a different perspective, a scenario oriented approach would be: “estimate the cost of electricity, consumed by the oven in the timeframe of a month” (Dong et al. 2013). The former approach can be characterized as the principal goal. Thus, the solution of the latter can be extrapolated from the solution of the first one.

According to the above, a formal definition of power disaggregation is described as follows (Batra et al. 2014b; Parson et al. 2011, 2012):

> Given a discrete sequence of observed aggregate power readings $x = x_1, \ldots, x_T$ determine the sequence of appliance power demands $w^{(n)} = w_1^{(n)}, \ldots, w_T^{(n)}$ where n is one of N appliances. Alternatively, this problem can be represented as the determination of appliance states $z^{(n)} = z_1^{(n)}, \ldots, z_T^{(n)}$, if a mapping between states and power demands is known. Each appliance state corresponds to an operation of approximately constant power draw (e.g. ‘on’, ‘off’ or ‘standby’) and t represents one of T discrete time measurements.

Another similar definition is (Batra et al. 2014a):

> The aim of energy disaggregation is to provide estimates, $\hat{y}_t^{(n)}$, of the actual power demand, $y_t^{(n)}$, of each appliance n at time t, from household aggregate power readings, $\bar{y}_t$. Most NILM algorithms model appliances using a set of discrete states such as off, on, intermediate, etc. We use $x_t^{(n)} \epsilon \mathbb{Z} > 0$ to represent the ground truth state, and $\hat{x}_t^{(n)}$ to represent the appliance state estimated by a disaggregation algorithm.

2.1 NILM framework

A nonintrusive load monitoring (NILM) system collects energy consumption data from the central meter reading of a residence. It subsequently can infer the consumption of each appliance, present in the residence. The NILM framework as described by Carrie Armel et al. (2013) and later by Burbano (2015), is made of three basic steps (Fig. 1).

The data acquisition step refers to the way the energy data is collected and is mainly based on hardware solutions. Smart meters are playing a significant role, because of the growing adoption rate, the low cost and the minimum effort to install them. Considering the way data is collected, there are three characteristics affecting the performance of the NILM system.

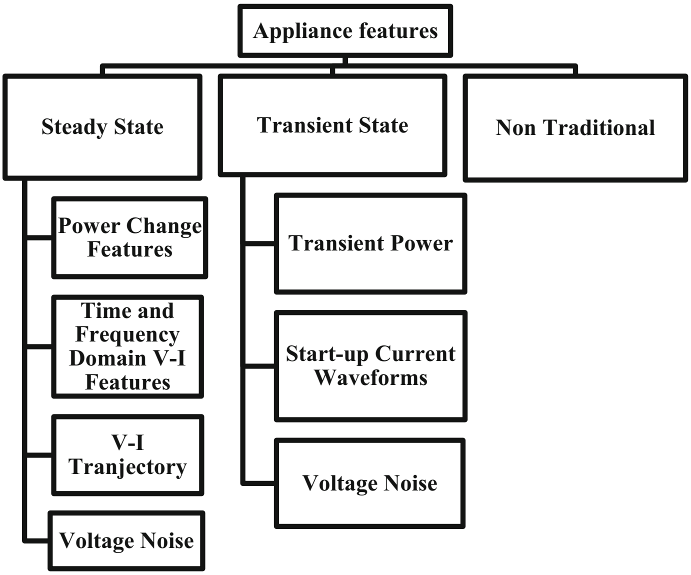


Fig. 2 Hierarchical taxonomy of appliance features

The first one is the type of power, which can be real or reactive. The second characteristic is the power level resolution, which is basically the smallest quantity of power that can be detected by the system. The final characteristic is the data sampling frequency, which is split into two categories: low and high frequency. A typical low frequency is considered up to 1 kHz and high frequencies range from 10 to 100 MHz (Armel et al. 2013).

From a different perspective, the data acquisition step will also define the limitations of a NILM system with major impact on privacy of users. The fundamental limits of a NILM system are explored in depth by Dong et al. (2013). In this paper, an upper bound on the probability of distinguishing scenarios, for an arbitrary NILM system, is derived and depends on the data characteristics. This is the first theoretical proof that NILM system requirements can be predefined, in order to protect user's privacy.

The appliance feature extraction step, can be divided into the hierarchical taxonomy (Zoha et al. 2012) depicted in (Fig. 2).

There is an extensive analysis of the appliance features by Zoha et al. (2012). In a nutshell, this analysis states that the group of steady state features require low sampling frequency, while being a lower cost solution. On the other hand, a solution using transient state features is more expensive, because of the need for sophisticated hardware. Performance can be improved, when appliances with overlapping steady-state features are involved. In practice, a solution using steady state features is more feasible, considering the existing hardware in the market.

Furthermore, there is a third category of features, called Non-Traditional. This group includes all the features, which do not belong to any of the two traditional categories. They can be several traditional features combined, contextual based, behavioral or indicators of the electric devices. Kim et al. (2011) proposed a conditional factorial hidden semi-Markov model, which integrates non-traditional features. The model is based on the observation that usage of appliance should have temporal patterns, while considering the time of day and day of week. A contextual supervised source separation was proposed by Wytock and Kolter (2013), in which paper the contextual features are radial basis functions over temperature and hour of day. It proves that these features enhance the final performance of disaggregation. In a similar way Aiad and Lee (2016) proposed an unsupervised approach with devices interactions. Low power quality issues can lead to disturbance of harmonics and interference currents, due to the operation of other devices. Embedding these interactions in a factorial Hidden Markov Model, shows improved disaggregation accuracy for the devices involved in mutual interactions. Finally, Lange and Bergés (2016) used binary subcomponents as features. A neural decoder is used to extract the subcomponents which are then combined to disaggregate the power consumption of the appliances.

The final step of NILM algorithm, called inference and learning, is the mathematical model that disaggregates the total power signal into appliance level signals. There are two main approaches to solve the problem of inference in a NILM system: optimization and machine learning. Researchers have proposed many solutions based on optimization methods (Baranski and Voss 2003; Liang et al. 2010; Inagaki et al. 2011). This approach proved to be inefficient in terms of computational complexity. The ability to distinguish similar signature profiles is limited, especially when tested in a very complicated environment with a large number of devices.

Given the difficulties to improve optimization based algorithms, modern approaches focus on machine learning methods, including both supervised and unsupervised learning. For the former type, researchers have worked using techniques such as neural networks (Ruzzelli et al. 2010; Srinivasan et al. 2006), support vector machine SVM (Kato et al. 2009; Lin et al. 2010), Hidden Markov Model (Zia et al. 2011), Bayesian approaches (Marchiori et al. 2011; Zhong et al. 2015), clustering (Hart 1992; Laughman et al. 2003) and combinations of different techniques (Chang et al. 2011; Lai et al. 2013; Liang et al. 2010). The unsupervised type of machine learning methods outperforms supervised ones in the sense that training of the NILM system can be avoided and therefore require minimum effort from the user. Zeifman and Roth (2011) reviewed these methods in detail.

Most modern approaches are based on hidden Markov models. Scientists focus on unsupervised solutions, which require minimum effort from the user. Since 2012, the unprecedented performance of deep neural networks in the field of visual recognition (Krizhevsky et al. 2012) and other domains, caught NILM researchers' attention as well. Despite the difficult process of training, deep neural networks perform efficiently, when compared to state of the art versions of HMM. In this paper, the latest HMM based algorithms are compared against recent approaches, which are based on neural networks.

2.2 Complexity

When power disaggregation was firstly introduced by G. W. Hart, it was described as a combinatorial optimization problem, based on the assumption that the number of the devices is known a priori. Consequently, it is an NP-complete problem, which means that it is computationally intractable for large number of devices. Additionally, the complete set of p_i is never known, because the number of devices in a residence can change from time to time. Restricting the problem to a specific number of devices in order to make it solvable, is not a viable solution (Hart 1992).

The abovementioned approach is based on the theory of computational complexity making a fundamental assumption, that the complexity depends on the number of appliances. On the other hand, it has been practically proved that the complexity is increasing when (Egarter et al. 2015):

- The number of appliances increases.
- The switching frequency of device states is higher.

- The devices have many operation states.
- There are several devices with similar power consumption.
- There is additional noise above the average or devices that are not considered.

It is evident that a different complexity measure is necessary, so that a comparison amongst different power disaggregation systems is practicable. Shannon's Entropy has been considered, but fails in the sense that it doesn't solve the problem of the similarity of multiple states. Furthermore, Kolmogorov's complexity is a promising theoretical concept, but lacks in practicability, because there is no typical method to estimate it. Egarter et al. (2015a) introduced a novel complexity measure, fulfilling the following requirements:

- Quantifies the complexity of the load disaggregation problem without depending on the disaggregation approach.
- Takes into consideration the number of appliances, the number of states and the similarities between states and appliances.
- Considers the appliance usage in time and is applicable to time series.
- Is simple and comprehensive.
- Its scope is limited to power disaggregation problems and makes such systems comparable, without considering the disaggregation approach.

The introduced complexity measure uses the similarity factor of power states amongst the devices of an appliance set. It is called defined by using the overlapping coefficient as follows (Egarter et al. 2015a):

$$C_k = \sum_{j=1}^{M} OVL\left(f_{P_k}, f_{P_j}\right), \quad k \in [1, M] \tag{2}$$

or

$$C_k = \sum_{j=1}^{M} \int_0^{P_M} min\left(f_{P_k}(p), f_{P_j}(p)\right) dp, \quad k \in [1, M] \tag{3}$$

where C_k is the disaggregation complexity for power state k, P_k is the reference power value, k is the reference power state and M is the number of combinations of power states.

Equation 3 gives the complexity of a specific power value k against a set of M power state combinations. Accordingly, the disaggregation complexity in a time series will be the average complexity in a specific time frame. Given a time series set of T power samples, we calculate the complexity C_t for each sample t against all the M different appliance state combinations. Then we calculate the average complexity for all C_t. The formula of the disaggregation complexity in a time series is:

$$C_{total} = \frac{1}{T} \sum_{t=1}^{T} C_t = \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{M} OVL\left(f_{P_t}, f_{P_k}\right), \quad k \in [1, M] \tag{4}$$

where T is the observation timeframe, C_t is the disaggregation complexity for the power state in time t.

2.3 System requirements

NILM framework describes the three basic parts, that are necessary to solve the problem of power disaggregation. However, it is not adequate to set the criteria that will make a solution robust and applicable in the real world. For this reason, Zeifman (2012) proposed six requirements for NILM systems to be applicable, according to the already established technology of smart meters:

- Feature selection. The majority of smart meters and the lower cost solution demand a sampling rate of 1 Hz. This requirement refers to the data acquisition and feature extraction parts of NILM framework, because the frequency of data collection affects the features that can be extracted.
- Accuracy. There are indications that an acceptable user experience would require a minimum accuracy of 80–90%.
- No training. For a seamless user experience, the user should put minimum effort configuring and training the system. Then, the system should be able to detect new devices, unseen devices and discard unused devices.
- Near real-time capabilities. The system should work online and should give immediate and accurate feedback regarding the current energy status of the house.
- Scalability. The system should be robust and efficient even for more complex environments e.g. more than 20 devices.
- Various appliance types. The system should be able to recognize four different types of electrical devices: a) on/off, b) finite-state, c) variable power and d) permanent consumer.

Zeifman's requirements are widely accepted by researchers and are extensively used to evaluate NILM systems. In this paper the original naming convention is used, although there might be some confusion. For example, the term "feature selection" usually refers to the process of extracting and selecting features from the collected data. In NILM, it also refers to the features of the hardware which collects the data. These two cases are related, because the sample rate of the data affects which features can be extracted or not. However, sampling rate is equally important for methodologies that do not require a manual process of feature extraction (e.g. neural networks). As far as "accuracy" is concerned, no metric is specified. The term "accuracy" is quite abstract and the 80–90% is an indication. Finding the right metric of accuracy for NILM systems is a subject of research and researchers haven't agreed on which one is the most suitable. "No training" should not be confused with the meaning of "training" in machine learning, although it could be related to that. According to Zeifman it refers to the configuration of the system and how the user is involved in this configuration e.g. select which appliances will be used in the house or configure each appliance separately by switching it on and off.

According to Kelly and Knottenbelt (2015a), another equally important requirement of a NILM system is the ability to generalize to unseen houses. In the real world it is very unlikely that ground truth appliance data will be available for each house. Generalization is a property very popular in machine learning but older approaches haven't taken this aspect into account. In this paper NILM approaches will be compared considering "generalization" as well. Consequently, a metric and an acceptance threshold should be proposed to add "generalization" as an additional requirement. Generalization should not be confused by "no training". The fact that an algorithm works well on an unseen environment doesn't necessarily affect the user's effort to configure the system.

Additionally, a NILM system should respect user's privacy. The limitations of a NILM system are analyzed by Dong et al. (2013), emphasizing the importance to know what information can be disclosed through power disaggregation. Indeed Greveler et al. (2012) prove that power signal from a house can provide useful information about the residents e.g. audiovisual content showed on TV. The risk of information leakage, without user's permission, is conclusive to add privacy as a requirement for a NILM system in real life. Privacy should not be confused with security. Data transmission can be secured, in the same way any data is secured in a network. However, in the past no user thought that the energy data of the house they live could reveal aspects of their private life. This can be prevented by limiting the capabilities of power disaggregation, which most of the time is against the first requirement of high accuracy. To set an example, a user would accept a NILM system that reveals if the TV is on or off, but not one that also reveals which program he or she watches on TV.

The above criteria constitute a qualitative way to evaluate a NILM system. There are also many quantitative metrics, but they are not mapped to all qualitative requirements. There is an inconsistency between quality and quantity criteria.

2.4 Performance evaluation criteria

A disputable challenge in NILM is the existence of a solid benchmarking tool. Researchers have been evaluating their solutions on different datasets, with different criteria and using different metrics. Therefore, a direct comparison between different methods is ambiguous.

Liang et al. (2010) inquire into this challenge. Firstly, accuracy measures, such as detection, disaggregation and overall accuracy, are defined. Secondly, appliance-based accuracy is deduced. Similarity and complementary ratio are also defined, in order to quantify the divergence between two electrical signatures and the correlation between two solutions, respectively.

The aforementioned assessment tools are inefficacious, when a NILM solution doesn't include the step of event detection and doesn't take into account the complexity of the environment. Kim et al. (2011) address the problems of using accuracy as a metric for evaluating NILM systems, based on the fact that the event of a specific appliance is negligible in a reasonable period of evaluation. In this case, a model predicting that an appliance is never working, will have high accuracy. To counteract this problem, they propose F-measure as a more suitable metric. The F-measure is the harmonic mean of precision and recall. These two measures are redefined in Kim et al. (2011), considering accurate true positive and inaccurate true positive results, depending on whether the predictions exceed a threshold of distance from ground truth.

Another detection and classification metric, which is popular among pattern recognition researchers, is the receiver operating characteristic (ROC) curve. ROC curve has been proposed as an evaluation method for NILM systems by Zeifman and Roth (2011), but there is no experiment verifying the suitability of the method, when applied on power disaggregation.

The most common performance evaluation criteria among NILM researchers, as described by Bonfigli et al. (2015), are split into two categories. The first category is based on the comparison between the observed aggregate power signal and the reconstructed signal after disaggregation. These metrics include: root-mean-square error (RMSE), mean average error (MAE) and disaggregation percentage (D). The respective equations are:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left|\bar{y}_t - \hat{\bar{y}}_t\right|^2} \tag{5}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|\bar{y}_t - \hat{\bar{y}}_t\right| \tag{6}$$

$$D = \frac{\sum_{i=1}^{k} E_i}{E_{tot}} \tag{7}$$

where $\bar{y}_t$ is the observed aggregate signal, $\hat{\bar{y}}_t$ is the reconstructed signal after disaggregation, E_{tot} is the total energy of the observed aggregate signal, K is the number of appliances signals and T is the number of samples.

The second category describes how effectively the disaggregated signal signatures are assigned to appliance signatures and include: total energy correctly assigned (TECA), disaggregation error (DE), precision (P), recall (R), accuracy (Acc) and F-measure (f1). They are defined as follows:

$$TECA = 1 - \frac{\sum_{t=1}^{T} \sum_{i=1}^{K} |\hat{y}_t^{(i)} - y_t^{(i)}|}{2 \sum_{t=1}^{T} \bar{y}_t} \tag{8}$$

$$DE = \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} |\hat{y}_t^{(i)} - y_t^{(i)}|^2 \tag{9}$$

where $\hat{y}_t^{(i)}$ is the separated appliance signal, $y_t^{(i)}$ is the original appliance signal, $\bar{y}_t$ is the observed aggregate signal, K is the number of appliances signals and T is the number of samples.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

$$f1 = \frac{2 * P * R}{P + R} \tag{13}$$

where TP is the true positive that the appliances was working, FP is the false positive that the appliance was working, TN is the true negative and FN is the false negative.

Bonfigli et al. (2015) focus on unsupervised methods. The same metrics are used for supervised methods as well. Despite the plethora of evaluation metrics, they are not adequate to perform a fair comparison of the most promising solutions. It is very important though to keep a consistency and new experimental results should include the same metrics with older ones. Otherwise researchers should replicate older experiments and use the most suitable metric.

In order to facilitate an objective and reproducible way of comparison between energy disaggregation algorithms, Batra et al. (2014a) designed an open source toolkit called Non-intrusive Load Monitoring Toolkit (NILMTK). The main features of NILMTK are: easy dataset parsing, efficient loading of data to RAM, preprocessing methods, statistical characteristics of the datasets, plotting, a couple of disaggregation algorithms and common accuracy metrics.

Beckel et al. (2014) tackled the performance evaluation challenge, by introducing a new dataset called ECO as well as an evaluation framework, called NILM-Eval. ECO dataset provides quality and quantity of both aggregate and appliance specific power data, including real and reactive power. NILM-Eval framework helps running benchmarking experiments and compare different NILM algorithms by tuning multiple parameters. These include a variety of datasets, diversified houses or time periods, particular data features and distinctive algorithm configurations. The metrics, that are used by NILM-Eval, are RMSE, F1 score, precision and recall.

Finally, a new version of F1 score was proposed by (Makonin and Popowich 2015), called finite-state-f-score (FS-fscore). FS-fscore is calculated using precision and recall, which are redefined including a partial penalization measure. This is called inaccurate portion of true-positives (inacc). The inacc is given by the following equation:

$$inacc = \sum_{t=1}^{T} \frac{\left| \hat{x}_t^{(m)} - x_t^{(m)} \right|}{K^{(m)}} \tag{14}$$

where $\hat{x}_t^{(m)}$ is the estimated state from appliance m at time t, $x_t^{(m)}$ is the ground truth state and $K^{(m)}$ is the number of states for appliance m. Precision, recall and FS-fscore are given by:

$$precision = \frac{tp - inacc}{tp + fp} \tag{15}$$

$$recall = \frac{tp - inacc}{tp + fn} \tag{16}$$

$$FS - fscore = 2 * \frac{precision * recall}{precision + recall} \tag{17}$$

where tp is true-positives, fp is false-positives and fn false-negatives.

3 Modern approaches in NILM

The problem of power disaggregation is time dependent by nature. NILM researchers have always thought models of sequential data and time as possible solutions. Hidden Markov Models have been gaining much attention, not only in the domain of NILM systems, but also in speech recognition and other sequential models. Several approaches are using different versions of HMMs and have demonstrated impressive results. However, the main limitations of traditional Markov models haven't been overcome. They are restricted in relatively small discrete state space, the algorithmic complexity for inference is intractable and the state space can easily grow exponentially. This exponential space complexity is worse, when extending the model in context window (Zachary 2015). In this survey, the most recent HMM models in NILM will be reviewed, showing the aforementioned problems. On the other side, deep neural networks have shown remarkable results in sequential models, such as speech recognition and translation. Some NILM researchers have shown strong interest in these models and the results justify this approach. The state of the art of traditional NILM solutions will be compared to deep learning solutions, in order to find out if deep learning solutions can be considered as state of the art.

3.1 Particle filter-based load disaggregation (PALDi)

Egarter et al. (2015) proposed a method, named Particle Filter-Based Load Disaggregation (PALDi). The devices that were used belong to Type-I (On/Off), Type-II (Finite State Machines) and combinations of them. Appliances' load signatures and superimposition of them were modeled with HMM and factorial HMM, respectively. Inference was estimated by Particle Filtering (PF). For the evaluation, both synthetic and real data were used and the preferred metrics were: normalized root-mean-square error (RMSE) and accuracy of classification, as presented in Eq. (10).

Five different scenarios were implemented on synthetic data and three on real data from REDD dataset (Kolter and Johnson 2011). The three scenarios on real data were three variations of PALDi: with noise adaptation, without noise adaptation and resetting posterior estimation. The number of particles was the same for all three cases, Np = 100. The variety of scenarios enhance the impressive results of total accuracy of 90%. However, the fact that the proposed algorithm was not compared directly with other approaches and was not tested on different databases, might indicate a case of overfitting.

3.2 FHMM exploiting context-based features.

Paradiso et al. (2016) proposed a new power disaggregation method, which uses Factorial Hidden Markov Models (FHMM) and exploits context-based features. The context information consists of the user presence and the power consumption patterns of appliances. The experiments used real home data from Tracebase dataset, sampled at low frequency. The proposed solution is cost effective and easily applicable, because sampling requirements are already met from existing smart meters. Any extra sensors, to gather the contextual features, are widely available and can have low cost.

Data from Tracebase was preprocessed and the devices that were tested are: coffee maker, LCD TV, microwave oven, pc, refrigerator and washing machine. Each appliance behavior was approximated by a few states and power states were extracted by using clustering analysis. The adopted clustering method was Gaussian Mixture Model (GMM). Next, extensive experiments were run, testing each context-based scenario and comparing the results against the Additive Factorial Approximate MAP (AFAMAP) (Kolter and Jaakkola 2012). Finally, both types of contextual information were combined and evaluated.

AFAMAP is based on FHMM, has the same scalability issues and is susceptible to local optima. The main advantage of the model is that the result fits very well with the total aggregate signal. Kolter and Jaakkola (2012) combined this algorithm with a model called difference FHMM. The final model is computationally tractable, is robust to noise and overcomes problems of local optima. Paradiso et al. compared their model against the basic additive model and the results are very encouraging. It would be interesting to see a comparison against the combined model as well.

The basic scenarios of the experiments were: user presence single interval conditioning, user presence double interval conditioning, usage statistics conditioning and a combination of the last two. the metrics that were used for the evaluation are precision and F-Measure. On average the scenario with combined context data showed the best results, whereas each scenario performed better than the basic AFAMAP algorithm. The suggested approach is not directly compared with other solutions different from AFAMAP, but the results prove that context information can be very valuable in a NILM system.

3.3 FHMM with device interactions

Aiad and Lee (2016) suggested an unsupervised disaggregation model, taking into account the interactions between devices. The device interactions were modeled using Factorial Hidden Markov Model and inference was performed by the Viterbi algorithm. The model was tested on the well-known REDD dataset (Kolter and Johnson 2011) and was compared to the standard FHMM. The results were significantly improved in the case of device interactions, whereas no substantial improvement was observed in the absence of interactions.

According to this methodology, a state of an appliance was described by three values: the initial, average and final power consumption. In order to simulate any pulsations, a state was also represented as a random variable with normal distribution. Only a small fraction of the power consumption was used to model the states and estimate any mutual device interactions. Then, power disaggregation took place using the Viterbi algorithm. Finally, the results were compared to real data. The total energy, which is correctly assigned, was estimated by the metric of accuracy. The data of the experiments come from house 2 in REDD dataset.

A direct comparison was shown against the standard FHMM where device interactions were ignored. The results are encouraging as the proposed solution showed the same or better performance for all devices. However, still the average accuracy, over 7 devices, was less than 70%. These devices include: microwave, outlet 1, outlet 2, refrigerator, washer dryer, lights and stove.

3.4 Sparse Viterbi algorithm

Stephen Makonin et al. (2015) proposed a new algorithm, tackling the efficiency problem of Viterbi algorithm. This approach was based on super-state Hidden Markov Model and a different version of Viterbi algorithm. A super-state was defined as an HMM that describes the overall power state of a set of appliances. Each appliance could be ON or OFF and during operation could have an operation state. Each combination of the devices' states represented a unique state of the house.

The main advantage was that exact inference was feasible in computationally efficient time, by calculating sparse matrices with large number of super-states. Disaggregation could also run in real time, even on an embedded processor with limited capabilities.

The methodology started with separation of data in two categories, priors and testing. For the priors, it was necessary to know the sub-meter readings of the devices that would be disaggregated. Then a Model Builder was used to create a super state HMM. Firstly, each load was mapped to a probability mass function (PMF) by using the priors. Secondly, each PMF was quantized finding each load's states and the peak value of the state. Thirdly, the load's states were combined and constituted the super-state HMM. The super-state and the smart meter data were introduced to the Sparse Viterbi Algorithm, finding the state with the maximum probability. The key of the algorithm was that it ignored zero-probability terms. Finally load consumption was estimated by decoding the quantized state and finding the peak of the respective PMF.

Regarding the data during the experiments, REDD and AMPds (Makonin et al. 2013) were selected, because they support low frequency data. The evaluation criteria were accuracy, FS-fscore and consumption estimation accuracy (Kolter and Johnson 2011). For more robust results, tenfold cross-validation process was applied.

It is worth mentioning that this algorithm was directly compared to related work, Kolter and Johnson (2011) and Johnson and Willsky (2013), where factorial Hidden Markov Model (FHMM) and Hierarchical Dirichlet Process Hidden semiMarkov Model (HDP-HSMM) were used respectively. The former one is the standard FHMM and it has been the basis when comparing NILM systems. The latter one is a method which overcomes two restrictions of Hidden Markov Models. The first restriction is that HMMs operate using discrete states and the distributions of state duration consist a graph that doesn't describe adequately real-world data. The second one is that hidden states are set a priori defining model's complexity in a non-Bayesian way. HDP-HSMM is based on Hierarchical Dirichlet Process HMM (HDP-HMM) and takes advantage of semi-Markovian properties.

Johnson and Willsky claim that their method is unsupervised. HDP-HSMM is a Bayesian approach and the graphical model encodes prior appliance information. Therefore, there is no need for manual labelling. On the other hand, other researchers (Makonin et al. 2015;

Parson et al. 2014) argue that the method cannot be considered unsupervised in the real world because it requires prior knowledge for each appliance. The model must know the number and the type of the appliances in the house. Finally, this Bayesian approach has not been tested in complex environments with more than 15 appliances.

To conclude, both the basic FHMM and the HDP-HSMM are more complex models than the proposed sparse Viterbi algorithm. The comparison showed that HMM sparsity not only reduced complexity, but also presented major improvement in terms of accuracy, 13.3% better than HDP-HMM and 48.3% better than FHMM on average.

3.5 The neural energy decoder

Lange and Bergés (2016) examined a method using a neural decoder, in order to extract features from high frequency data. In summary, the aggregated power signal was broken into sub-components in an unsupervised way. Then, they were combined to shape appliance profiles. This was the first attempt of unsupervised feature extraction, by using a deep neural decoder.

The architecture of the deep neural network had a linear output layer and a binary activation function in last layer. The total number of layers was six and the framework that was used is the python package Keras. The inputs of the network were the real and imaginary parts of the Discrete Fourier Transform of the current signal and the target outputs were active and reactive power. The selected optimizer was stochastic gradient descent. The data came from Phase B of the BLUED (Anderson et al. 2012) dataset and the network was trained on all data. After training, the last layer of the decoder was removed, and the network could infer the binary subcomponents. Then, the subcomponents were used as features to predict which devices were on or off. For this purpose, two different algorithms were tested: greedy search and logistic regression. A naïve energy estimation was also examined. According to the results, logistic regression performed better F1 score than greedy search, with values more than 0.90 for the majority of the devices.

The main goal was to face three basic drawbacks of existing disaggregation solutions: computational efficiency, data transmission limitations and prior knowledge of electrical devices. The proposed algorithm overcomes these pitfalls, but also a direct comparison against other algorithms is needed.

3.6 Deep neural networks applied to energy disaggregation

Kelly and Knottenbelt (2015a) focused on solving the problem of power disaggregation by means of deep neural networks. In this scope, the authors proposed three different approaches: a) a solution using a specific type of recurrent neural network, named "long short-term memory" (LSTM), b) a solution reducing noise with denoising autoencoders and c) a regression algorithm forecasting the start time, end time and average power demand for each device.

The source of the data was UK-DALE dataset (Kelly and Knottenbelt 2015b). Both real and synthetic energy data were used for training. This approach of training on a mixture of real and synthetic data showed better generalization, when neural nets were tested against unseen data. The step of validation and testing included only real energy data, for more realistic results. NILMTK was used to preprocess data. The algorithms were written in Python using Pandas, Numpy and Lasagne. The target devices included washing machine, kettle, fridge, dish washer and microwave. The metrics that were used are seven: F1 score, precision, recall, accuracy, relative error in total energy, proportion of total energy correctly assigned and mean absolute error.

Regarding training, the first step was to find all the activations of the target device. Then, the algorithm decided randomly to include one of the activations in the timeframe that will be the target of the net during training. Below, the three architectures that were tested, are presented.

The first, of the three neural networks, had six layers in total. An input layer with length which depended on the duration of the device. The second layer was a 1D convolution layer that played the role of extracting features from the signal. The third and fourth layers were bidirectional LSTM. The last two were fully connected with the last being the output of the network. The recurrent neural network was trained with the method of backpropagation through time.

The second neural network was a denoising autoencoder and was used to reduce noise. The network was trained having a noisy signal as input and the clean signal as target, learning in that way to remove noise. The input was the total aggregate power signal and the target was the appliance's load. The architecture of this network consisted of an input layer, a 1D convolutional layer, three fully connected and one convolutional as the output.

The third architecture aimed to predict the start time, the end time and mean power demand of a target device. The network consisted of the input with length depending on the duration of the appliance, two 1D convolutional layers and five fully connected layers. The last one had three outputs, one for each of the target values.

The proposed algorithms were compared with combinatorial optimization and factorial Hidden Markov Model, using implementations of the tool NILMTK (Batra et al. 2014a). In general, denoising autoencoder and regression neural network showed better results than the two algorithms from the benchmark, except for some specific cases. For seen houses, CO and FHMM outperformed Denoising Autoencoder on the metric of relative error in total energy. Regression neural network was outperformed by CO and FHMM only when disaggregating microwave. It is worth mentioning that the two neural networks performed strong generalization capabilities, when data came from a new unseen house. The LSTM network outperformed CO and FHMM only on two-state devices. On more complex devices, such as dish washer and washing machine, LSTM was worse than the other two algorithms.

Further research is recommended with emphasis on optimization of neural networks, understanding LSTMs inferior performance and direct comparison against state of the art algorithms.

3.7 Deep recurrent LSTM network

Mauch and Yang (2015) presented another architecture of deep recurrent LSTM network, in order to test if this type of network is possible to overcome the known problems of previous NILM approaches. Such problems include disaggregation of various appliance types, automatic feature extraction from low frequency data, generalization of a solution to other buildings and unseen devices, extensibility of the approach to continuous time and computational tractability.

According to the suggested methodology, the experiments used synthetic power signals, by summing up sub metered data. The data source was the well-known REDD open dataset. The house, that was used both for training and testing, was house 1, whereas house 2 was used to test the generalization of the algorithm. The target devices included fridge, microwave and dishwasher. The first two appliances were considered as ON/OFF and the third one as multistate. The authors also had the intention to test this algorithm on variable load devices, but unfortunately the selected database doesn't include such devices. Another worth mentioning characteristic of the group of the chosen devices is that the fridge had a periodicity, in contrast to the other two.

The proposed solution needed one network for each device in a house. As a result, in these experiments three networks were used, one for each of the three target appliances. Each network had the following architecture: one input layer with ten units, two bidirectional recurrent layers with 140 units each and one output layer. Each network was trained until the validation error didn't decrease any more or for a maximum number of 100 epochs. The evaluation metrics that were used were estimated energy, consumed energy, NRMS for active periods, precision, recall and F1 score.

The proposed algorithm was not compared directly with other solutions, used only synthetic data and depended on sub metered signals. Nevertheless, the experiments showed encouraging results and lead to the following conclusions: supervised disaggregation was feasible with the proposed architecture, the LSTM architecture worked well for appliances showing some periodicity, the system worked with low frequency data, there was no need for event detection and feature extraction and the trained networks of the fridge and dishwasher could generalize efficiently when tested on house 2.

3.8 Sequence-to-point learning with neural networks

Zhang et al. (2017) proposed a deep learning solution for the problem of single-channel blind source separation with application in NILM. The method is called sequence-to-point (seq2point) learning, because it uses a window as input and a single point as target. The proposed solution is a deep convolutional neural network (CNN), which also learns the signature of the appliances in a house.

The seq2point solution differs from a sequence-to-sequence (seq2seq) solution regarding the target output. Instead of predicting a window of appliance power consumption, it predicts the midpoint of that window. The intuition behind the midpoint prediction is that this point is related not only to past values but also to future ones.

The data sources of the experiments come from UK-DALE and REDD. The devices that were tested are: kettle, microwave, fridge, dish washer and washing machine. Kettle wasn't tested on REDD database because it is not included. Regarding UK-DALE, houses 1, 3, 4, and 5 were used for training and house 2 was used for testing. In the case of REDD, houses 2 to 6 were used for training, and house 1 for testing.

The implementation of the model was done in Python, using Tensorflow framework. The architecture of the neural network has an input layer with length 599. Then there are 5 convolution layers with activation function ReLU. Next is a dense layer with 1024 units and activation function ReLU. The final layer is the output of the neural network. There are two versions of this layer, one representing the seq2seq solution and one representing the seq2point solution. For the former case, the layer has 599 units, whereas for the latter one it has only one unit. The activation is Linear for both cases.

Both seq2seq and seq2point versions of the proposed architecture are compared to AFHMM (Kolter and Jaakkola 2012) and the seq2seq autoencoder architecture (Kelly and Knottenbelt 2015a). The autoencoder is referred as seq2seq(Kelly), to differentiate it from the proposed seq2seq solution. The evaluation metrics that were used are: mean absolute error (MAE) and normalized signal aggregator error (SAE). MAE is more suitable for the error in power at each time step, whereas SAE is more suitable for the total error in a timeframe. The four models were tested using UK-DALE dataset. Overall the seq2point model gave the best results. It is very impressive that, compared to seq2seq(Kelly), it reduces MAE by 84% and SAE by 92%. Both versions of the CNN achieved much better results than the other two models. Next, only the two versions of the CNN were compared, by using REDD dataset. The results show that the seq2point is superior to seq2seq in most devices with substantial difference. It is worth noting that the tests, where seq2seq had lower error, the difference was negligible (less than 1 watt).

The deep convolution neural network did not only show state of the art performance, but also seems to understand the data. It extracts meaningful features, regardless of the version (seq2seq or seq2point). The authors show systematically that the model learns features that previously were being extracted manually. Such features are: change points, typical usage durations and power levels of appliances. They can be illustrated by plotting the last convolution layer of the proposed neural network.

4 Comparative analysis

Comparing disaggregation algorithms is an obscured process and in most cases unfeasible. This is probably the obstacle to find a practical solution, which would be ready for production. In this chapter both, a qualitative and a quantitative analysis, are presented. The challenges of each analysis are highlighted, while solutions are suggested. Finally, there is an effort to explain the relation between the two different methods and how NILM systems could be easier evaluated in future.

4.1 Qualitative analysis

Ten different approaches are compared, taking into consideration the NILM system requirements. These include not only Zeifman's requirements, but also "generalization" and "privacy". The former one, a known and of major importance aspect of artificial intelligence, is also identified as an indispensable property of a NILM. So far there is no indication for an acceptable value of a specific metric. For this qualitative evaluation, a system that meets the requirement of "accuracy" in both the seen and unseen environments, will be considered that it meets the requirement of "generalization". A more accurate metric will be defined in chapter 5. The "privacy", concerns sensitive data and private information that can be extracted by a NILM system. Table 1 presents which criteria each algorithm meets.

It is obvious that no conclusion can be made for the requirement of "privacy", which means that further research has to be made to integrate this feature into a NILM system. Also, regarding "generalization", deep learning techniques seem to be more suitable, although sometimes, no conclusion can be made for HMM based solutions, due to lack of experimental results. Another feature, where deep learning seems to be superior to HMM, is "scalability". This is one of the most difficult problems in HMM and only FHMM exploiting context-based features overcomes it, because it is based on (Kolter and Jaakkola 2012). "Feature selection" and "real-time" capabilities are the requirements the majority of the suggested systems meet successfully. As far as "accuracy" is concerned, the outcomes are: it is not met by all algorithms, context based features have a positive impact and yet sometimes it is difficult to conclude, because of the large variety of metrics. Finally, "no training" as well as "appliance types" are very challenging. Researchers have proposed unsupervised methods to meet the requirement of "no training", but this is not an easy task. Also, it is difficult to make a conclusion for the "appliance types", since there aren't enough data for variable power appliances. In fact, this requirement is partially met, e.g. for on/off and multistate appliances.

Particle Filter-Based Load Disaggregation (PALDi) Given the fact that a direct numerical comparison against other NILM systems is very complicated, the authors have evaluated

Table 1 (√) Met, (×) not met, (–) no conclusion

	Feature selection	Accuracy	No training	Real-time capabilities	Scalability	Appliance types	Generalization	Privacy
PALDi	√	√	×	√	×	–	–	–
FHMM exploiting context-based features	√	–	√	√	√	–	–	–
FHMM with device interactions	√	×	√	√	×	×	–	–
Sparse Viterbi algorithm	√	√	×	√	×	–	–	–
The neural energy decoder	×	–	×	–	√	–	–	–
Denoising autoencoders	√	√	×	√	√	–	√	–
Deep convolutional neural network for regression	√	√	×	√	√	–	√	–
RNN LSTM with convolution input layer	√	×	×	√	√	–	×	–
RNN LSTM	√	–	×	√	√	–	√	–
Seq2point CNN	√	√	×	–	√	–	√	–

PALDi algorithm against Zeifman's requirements. Three of them are fully met. The active power is collected with 1 sec, resolution reported accuracy is around 90% and operational running time is computationally efficient. On the other hand, "no training" requirement is partially met. Although there is no need for training during operation, the algorithm requires to know the used devices in the house and cannot recognize new or unseen devices. The other two requirements, "scalability" and "various appliance types", are not met. Regarding the former, the algorithm's complexity depends on two parameters, the number of particles and the number of appliances. HMM models are, in general, computationally impractical for more than 18 devices and according to the tests, particle filtering shows proportionally better results in the case of more particles. Consequently, PALDi is not scalable. For the appliance types, there is no evidence that the algorithm will present equal results for types such as variable power and permanent consumer devices. Finally, no conclusion can be made for "generalization" and "privacy".

FHMM exploiting context-based features This NILM system fulfills four requirements. "Feature selection", because active power is the basic feature sampled with low frequency. "No training", because it identifies the power profile states of a specific appliance in an unsupervised way, using Gaussian Mixture Model. "Real-time capabilities", because inference can be done in real-time. "Scalability", because it is based on the work of Kolter and Jaakkola (2012), where the proposed method scales almost linearly in the space of HMM states. On the contrary, there is no reported accuracy, thus no conclusion can be made. Only precision and F-Measure results are compared against a basic AFAMAP algorithm with 13 and 14% improvement respectively. Additionally, the proposed method was mainly tested using finite state machines such as coffee maker, LCD TV, microwave oven, refrigerator and washing machine. Therefore, the variety of devices used in the test is not sufficient to meet this requirement and further experiments need to be run. Concerning "generalization" and "privacy" no conclusion can be drawn.

FHMM with device interactions The proposed approach meets successfully three of the qualitative criteria. First of all, it uses low frequency data. Secondly, it is unsupervised, although manual labeling is required at the end of the process. Thirdly, inference can be done in real-time. The requirements that are not met are "accuracy", "scalability" and "appliance types". Regarding "generalization" and "privacy", unfortunately no conclusion can be made.

Sparse Viterbi algorithm The authors of this paper state that the proposed algorithm fulfils four requirements. The algorithm was successfully tested on low frequency data, average accuracy exceeds the minimum threshold of 80% and inference can be completed in under one millisecond. The fourth requirement is "various appliances types". According to the authors, it is met considering a more relaxed version of the requirement. In this case and for the purpose of approximate estimation, continuously variable devices can be represented by one more general state. However, someone could argue that mainly finite state machine devices were used during evaluation and eventually the algorithm doesn't fulfill this requirement. Additionally, there is no proof that all variable state devices can be represented by one general state. More experiments with more appliances could possibly solve this ambiguity. For the criteria of "no training" and "scalability", they are not met. The requirement of "no training" is not met because of the process of selecting priors. Regarding "scalability", the system demonstrated the ability to disaggregate efficiently up to 18 loads. This is a notable improvement, because previous methods were restricted to 6–9 loads. However, the problem of exponential time and space complexity remains unsolvable, when the target devices are more than 18. As far as “generalization” and “privacy” are concerned, the system cannot be evaluated, because of lack of further information and experiments.

The neural energy decoder This system meets only one requirement, “scalability”. There are two reasons explaining why the system is scalable. Firstly, after the neural decoder is trained, inference is computationally very cheap. Secondly, high frequency data will be transformed in binary subcomponents, which require much less space. This transformation of high frequency data makes it possible to exploit more information, but at the same time fails the requirement of “feature selection”. The other requirement that is not met is “no training”. Although the binary subcomponents can be extracted in an unsupervised way, the inference process is supervised. The system will need extra training to identify unseen devices. Regarding “accuracy” no conclusion can be made, because the metric of disaggregation performance is F1 score. In the same way, there is no evaluation outcome concerning “near real-time capabilities”. The characteristics of the system are promising, because once the decoder is trained, both feature extraction and inference can be done very efficiently. Nevertheless, the system is tested off line and the need for temporal randomization of the dataset can be an obstacle during online testing. Finally, there are not enough experiments or information to evaluate the algorithm against “various appliance types”, “generalization” and “privacy”.

Deep neural networks applied to energy disaggregation In this paper, there are three different architectures of deep neural networks: denoising autoencoder, deep convolutional neural network for regression and RNN LSTM with convolution input layer. The first two architectures, according to the experimental results, show similar behavior. They fulfill five of the eight qualitative system requirements. On the other hand, the third architecture doesn’t perform equally and meets only three requirements. In more details, the “feature requirement” is met by all three architectures, due to the low frequency data. According to the results, all but LSTM based architectures show average accuracy above 80%, outperforming the basic algorithms of CO and FHMM. Training is supervised for all the suggested networks and requires knowledge of the device signature profile. Therefore, they all fail to meet the requirement of “no training”. In contrast, they are all considered capable of performing in real time, because after training, inference is computationally efficient. For the same reason, neural networks are suitable for scalability. Regarding “generalization” the denoising autoencoder and the deep convolutional network have shown impressive results. Finally, there is no conclusion about “privacy” and “appliance types”. The devices were considered to have only two states ON or OFF. Consequently, further research is needed to examine if the proposed methods are suitable for other appliance types.

Deep recurrent LSTM network The proposed recurrent neural network meets four requirements: “feature selection”, “real-time capabilities”, “scalability” and “generalization”. Indeed, the experiments used low frequency data, neural networks, theoretically, can perform inference computationally efficient and the results showed satisfactory generalization. Regarding “no training”, it is supervised and requires knowledge of the devices in the house. Unfortunately, no conclusion can be made for “accuracy”. It is not used as a metric, but the other metrics showed encouraging results. Similarly, there is no conclusion for “various appliance types” and “privacy”. Only two device types are tested and more complicated experiments are required.

Deep convolution neural network (seq2point) Two versions of this architecture are presented in the paper. The seq2point version will be analyzed here because the results are by far the best. Also since the basic architecture is the same, equivalent results should be expected for the seq2seq version. The requirements that are clearly met are: “feature selection", "scalability" and "generalization". The reasons are that the selected databases have low frequency data, inference is computationally efficient and the model was successfully tested on unseen houses. "Accuracy" is not clear if it is met because of the different metrics that are used. This is also a deficit of the system requirements, where minimum acceptance values of other metrics should also be considered. However, the authors directly compare their proposed solution to Kelly's autoencoder. The results are impressive with a reduction of 84% in MAE and 92% in SAE. The model achieved better results than the other systems, for all devices. Considering the impressive results and their consistency, leads to the outcome that accuracy criterion should be met. Regarding the "real-time capabilities", there is an argument of how much of the future data is used as input, when there is seq2point learning. Considering that there are 599 inputs, the future data are 299. Also, the readings are recorded every 1–6 seconds. As a result, inference takes place with a lag of 299–1794 s or 5–30 min. A real-time system should respond to any changes in a few seconds. On the other hand, for the case of seq2seq the predictions are clearly real-time. Consequently, deciding if the seq2point can give real-time results, depends on the sampling rate and how much of the future data is needed. To answer such questions more experiments must be done with exact specification of real-time requirements e.g. how many seconds of lag is accepted for a NILM system. The requirement of "no training" is not met, as the proposed solution is supervised. Finally regarding "privacy" and "appliance types" no direct conclusion can be made.

4.2 Quantitative analysis

Table 2 shows the results from the suggested algorithms, the databases that were used and the number of target devices. For comparison reasons, the table presents the three most common performance metrics. These are: Acc (Eq. 10), F1 score (Eq. 13) and TECA (Eq. 8). To avoid any confusion it is worth mentioning that in bibliography TECA is also called Acc (Aiad and Lee 2016; Kolter and Johnson 2011) or Est Acc (Estimated Accuracy) (Makonin et al. 2015). For each metric the mean and the standard deviation are calculated. The calculations are based on results taken from the respective papers. For PALDi system the Table VII was used from the respective paper of Egarter et al. (2015), for FHMM with device interactions Table 5 was taken from the paper of Aiad and Lee (2016), for Sparse Viterbi algorithm Table V of Makonin's et al. work (2015), for The Neural Energy Decoder Table I from Lange's and Bergés' paper (2016), for the systems described by Kelly et al. (2015a) Figure 3 and Figure 4, and finally for RNN LSTM the source of the data were Table 1 and Table 2 from the paper of Mauch and Yang (2015). The results from the seq2point CNN are not presented in Table 2 because different metrics are used. Moreover, Zhang et al. (2017) presented a direct comparison to other models as discussed in chapter 3.8.

At the same table, appliance set complexity is also presented, to make comparison more objective. The calculation of complexity is based on Eq. (3) and is implemented in Python, using the NILMTK for parsing the databases[1]. The complexity of the appliance set, that Neural Energy Decoder used, is not calculated, because BLUED dataset is not integrated with NILMTK toolkit.

As it is noticed, the complexity of each NILM environment differs a lot and makes a direct comparison very difficult. For example, the RNN LSTM system described by Mauch and Yang, has been tested on an environment with maximum complexity 1.84 and mean complexity 1.2 for house 1. The mean f1 score is 0.78. On the other hand, the Deep Convolutional

[1] The authors would like to thank Odyssefs Krystalakos for his contribution in developing the software and running the experiments for the appliance set complexity. The source code is available at: https://github.com/christofernal/power-disaggregation-complexity.

Table 2 Numerical results, where: N = number of target devices, M = max, m = mean, avg = average, std = standard deviation

Systems		Datasource	Complexity	Avg and std of Acc	TECA	Avg and std of F1	N
PALDi	Noise-adapt	REDD House 1	M = 23.34 m = 4.39	m = 0.73 std = 0.24	–	–	6
	No noise-adapt			m = 0.72 std = 0.24	–		
	Resetting			m = 0.90 std = 0.10	–		
FHMM exploiting context-based features		Tracebase	M = 10.83 m = 6.17	–	–	Not detailed results	6
FHMM with device interactions		REDD house 2	M = 32.37 m = 19.46	–	m = 0.68 std = 0.11	–	7
Sparse Viterbi algorithm		REDD house 1	M = 6.09 m = 2.83	–	m = 0.95	–	5
		REDD house 2	M = 4.04 m = 1.97	–	m = 0.94		
		REDD house 3	M = 5.49 m = 2.10	–	m = 0.90		
		REDD house 6	M = 1.90 m = 1.19	–	m = 0.98		
		REDD all houses	–	–	avg = 0.94		
The neural energy decoder		BLUED phase B	–	–	–	m = 0.92 std = 0.08	13
Denoising autoencoders		UK-DALE seen house	M = 10.10 m = 4.75	m = 0.90 std = 0.08	–	m = 0.55 std = 0.18	6
		UK-DALE unseen house		m = 0.92 std = 0.06	–	m = 0.52 std = 0.32	
Deep convolutional neural network for regression		UK-DALE seen house	M = 10.10 m = 4.75	m = 0.94 std = 0.07	–	m = 0.58 std = 0.15	6
		UK-DALE unseen house		m = 0.96 std = 0.04	–	m = 0.54 std = 0.25	
RNN LSTM with convolution input layer		UK-DALE seen house	M = 10.10 m = 4.75	m = 0.68 std = 0.29	–	m = 0.39 std = 0.28	6
		UK-DALE unseen house		m = 0.66 std = 0.33	–	m = 0.38 std = 0.37	
RNN LSTM		REDD house 1	M = 1.84 m = 1.20	–	–	m = 0.78 std = 0.12	3
		REDD house 2	M = 1.84 m = 1.06		–	m = 0.56 std = 0.43	

Neural Network is tested on environments with mean and maximum complexities around 4.7 and 10, whereas the mean f1 score is around 0.55. Based on these numbers, it is obvious that a direct comparison is unfair. Even if two systems are compared by using results from the same house e.g. Redd House 1, the complexity can be different, because each experiment might use different appliance sets. The more the appliances and the states of each appliance, the higher the disaggregation complexity.

The appliance set complexity doesn't depend on the disaggregation algorithm. It characterizes how difficult it is, to solve the problem of disaggregation for a given appliance set. This means that if the results were based on datasets with similar complexity, a comparison among all the approaches would be easier. As an example, let's compare the results of the PALDi system and the Deep Convolution Neural Network. According to the metric of Acc the latter one seems to outperform the former. Both have similar mean complexity, which is not a surprise, because they both recognize the activity of the same number of devices.

Another difficulty that can emerge from comparing these NILM systems is that they are using different metrics. To set an example, regarding the three common metrics Acc, TECA and f1 score, PALDi uses only Acc, whereas The Neural Energy Decoder uses only f1 score. A different metric could also be the result of how each solution is solving the problem of power disaggregation. TECA, for example, is used when a system predicts the power consumption of a device, whereas Acc and f1 score are used to decide if an appliance is ON or OFF or to classify its state. In addition, researchers do not always disclose all the details of the experimental results. As it can be noticed from Table 2, some useful information is missed. The standard deviation, for example, could give an overview of the performance of a system for the various devices.

To conclude, a quantitative analysis of various NILM systems is not feasible. The variety of methodologies, datasets, metrics and ways of presenting results make this procedure difficult.

5 Mapping quality requirements to quantity metrics

Power disaggregation is an old and yet unsolved problem. Comparing different NILM systems is essential to solve it, but still not possible. Each of the qualitative and quantitative evaluations, has proved to be inadequate to distinguish the best system. This means that maybe there should be a connection between the two methodologies.

Table 3 shows that researchers' efforts focused on measuring the performance of algorithms in terms of disaggregation accuracy. Now, there are several metrics, making a comparison even more confusing. On the other hand, there is plenty of room for improvement on how to quantify the rest of the qualitative requirements. On these grounds, some simple evaluation measurements are proposed, while further studies and experiments are encouraged.

A new metric called NTR (no training) is proposed for the requirement of "no training", which ensures the minimum user involvement. Ideally, the algorithm should be able to identify any new devices with no further user interaction. In practice, there are three distinct categories regarding user interaction, thus NTR values can be the following: a) no user interaction (NUI) where the system can recognize unseen devices, b) light user interaction (LUI), in which case the system doesn't need training, but knowing the used devices in the house is necessary, c) heavy user interaction (HUI), when the system requires training for each new device and consequently the user has to reconfigure the system. From another perspective, the three states also represent how much the algorithm knows about its environment: unknown environment for NUI, partial known environment for LUI and known environment for HUI.

Table 3 Mapping of qualitative requirements and quantitative metrics

Qualitative requirements	Quantitative metrics
Feature selection	Sampling rate
Accuracy	Root-mean-square error (RMSE), mean average error (MAE), disaggregation percentage (D), total energy correctly assigned (TECA), disaggregation error (DE), precision (P), recall (R), accuracy (Acc), F-measure (f1), finite-state-f-score (FS-fscore)
No training	Three-state variable for no training (NTR)
Real-time capabilities	Algorithm computational complexity
Scalability	Algorithm computational complexity
Appliance types	Four appliance types standard deviation (FATσ)
Generalization	Generalization over unseen houses (GoUHσ)
Privacy	Upper bound on the probability of distinguishing scenarios (UBPDS)

$$NTR = \begin{cases} NUI, \textit{no user interaction} \\ LUI, \textit{light user interaction} \\ HUI, \textit{high user interaction} \end{cases} \tag{18}$$

The second metric which is proposed, measures the performance of power disaggregation on four different appliance types: on/off, finite-state, variable power and permanent consumer. A naïve approach, to quantify this requirement, is to count how many of the four types can be recognized e.g. 1/4, 3/4 etc. A more sophisticated approach is to calculate the standard deviation of the accuracies of each of the four appliance types. The new metric is called $FAT\sigma$ (four appliance types standard deviation) and is described by the following formula:

$$FAT\sigma = \sqrt{\frac{1}{4}\sum_{i=1}^{4}(Acc_i - FAT\mu)^2} \tag{19}$$

where $FAT\mu = \frac{1}{4}\sum_{i=1}^{4} Acc_i$ and Acc_i the accuracy (Eq. 10) of each appliance type. Instead of Acc other metrics of "accuracy" can be used such as F1 score.

As far as the requirement of generalization is concerned, a metric similar to $FAT\sigma$ is proposed, representing how well the algorithm generalizes on unseen houses. It is defined as the standard deviation of the total disaggregation accuracy of the system for various houses. The smaller the value, the better the algorithm can generalize. The following formula defines the metric of $GoUH$ (generalization over unseen houses):

$$GoUH = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(Acc_h - \mu)^2} \tag{20}$$

where H is the number of different houses, Acc_h is the average disaggregation accuracy on house h, and μ is the mean accuracy over all houses. Instead of Acc as defined in Eq. 10, other metrics such as F1 score are also accepted. Based on Ziefman's requirement for a minimum accuracy of 80%, a maximum threshold could also be defined regarding $GoUH$ with value $GoUH_{max} = Acc - 0.8$.

Finally, it is equally important to quantify privacy. Ignoring the limitations of power disaggregation, could lead to leakage of private information without user's agreement. In order to protect privacy of occupants, Sankar et al. (2013) propose a theoretical framework. Another approach would be to use a distance metric (Katos et al. 2011) finding the correlation of scenarios, given the features and power data that will be used by the algorithm. A deep analysis of private prediction problems, mainly based on hidden Markov models, is presented in related work by Polat et al. (2010). Privacy is a multilateral problem, requiring knowledge of laws and definition of policies. Theoretical models, will only be useful after setting clear rules and policies on what violates an individual's privacy. Next an example of a metric which probably best fits NILM privacy is presented and is based on an upper bound on the probability of distinguishing private scenarios (Dong et al. 2013). Assume a NILM system which can identify N scenarios. Also assume that m out of N of these scenarios are considered private information. According to Dong et al, there is an upper bound probability of distinguishing successfully m private scenarios. The lower this upper bound is, the better privacy this NILM system provides. This metric can be calculated as follows (Dong et al. (2013)):

$$UBPDS = \sum_{i=1}^{m} P(\hat{u}_{MAP}(y) = v_i | u = v_i) p(u = v_i) \tag{21}$$

where MAP is given by:

$$\hat{u}_{MAP}(y) = arg\, max_{v \in V} P(G(u, \cdot) = y | u = v) p(u = v) \tag{22}$$

and V is a finite set of inputs representing scenarios, $\hat{u}$ is an estimator, G is the distribution of the power consumption, u is the input representing the scenario that will be distinguished, y is the observed signal.

6 Conclusions

In this review, the problem of power disaggregation was discussed extensively. Ten different algorithms were presented followed by a comparative analysis. Through this analysis many challenges have been discovered, leading to valuable conclusions and suggestions for future research.

The comparison of various NILM systems is still cumbersome. However, there has been a lot of improvement and there are novel approaches and mathematical tools that haven't been used extensively yet. A strong example is Zeifman's requirements. They consist a qualitative evaluation method, which has been adopted by many NILM researchers. These requirements can be complemented by two new ones. One regarding "generalization" and one concerning "privacy". Another example is disaggregation complexity. It measures the complexity of a NILM environment and sets a common basis of comparison. It would be unfair to compare two solutions on two environments with different complexity.

Another conclusion is that there is lack of metrics for the majority of the requirements. For this reason, some simple metrics are suggested for the following requirements: "no training", "appliance types", "generalization" and "privacy". Further research is encouraged for the discovery of more suitable metrics.

Furthermore, it is worth noting that, although there are several metrics for the requirement of disaggregation accuracy, only accuracy (Acc) has been specified with a minimum acceptable value of 80–90%. For a fair and objective evaluation, other metrics should also be assigned a threshold, in order to justify if the requirement is fulfilled or not. An accurate way to calculate these values would be to find mathematical correlations between metrics. Unfortunately, this is not possible because they are not directly related. An estimation approach would be to setup a series of tests of a known NILM system and get feedback from users. Another proposal would be to create a large dataset with results including all the metrics from different NILM systems and use computational methods to find out if there is any correlation among these metrics.

Finally, neural networks show evidence that they can meet the requirement of “generalization” quite confidently. Only two requirements are not met along all deep learning solutions: ‘appliance types” and “no training”. The first one is because of lack of data from variable power appliances, thus there are not enough tests. The second one is due to the supervised nature of these solutions. For the future work, it is suggested to run various experiments using different deep learning architectures and unsupervised training.

References

Aiad M, Lee PH (2016) Unsupervised approach for load disaggregation with devices interactions. Energy Build 116:96–103. https://doi.org/10.1016/j.enbuild.2015.12.043

Anderson K, Ocneanu AF, Benitez D, Carlson D, Rowe A, Bergés M (2012) BLUED? A fully labeled public dataset for event-based non-intrusive load monitoring research. In: Proceedings of the 2nd KDD workshop on data mining applications in sustainability (SustKDD), Oct 2011, pp 1–5

Armel CK, 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

Baranski M, Voss J (2003) Non-intrusive appliance load monitoring based on an optical sensor. In: 2003 IEEE Bologna power tech conference proceedings, vol 4, pp 267–274. https://doi.org/10.1109/PTC.2003.1304732

Batra N, Kelly J, Parson O, Dutta H, Knottenbelt W, Rogers A, Srivastava M (2014) NILMTK? An open source toolkit for non-intrusive load monitoring categories and subject descriptors. In: International conference on future energy systems (ACM E-Energy), pp 1–4. https://doi.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. Retrieved from arXiv:1408.6595 [Cs]

Beckel C, Kleiminger W, 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. https://doi.org/10.1145/2674061.2674064

Bonfigli R, Squartini S, Fagiani M, Piazza F (2015) Unsupervised algorithms for non-intrusive load monitoring: an up-to-date overview. In: 2015 IEEE 15th international conference on environment and electrical engineering, EEEIC 2015—conference proceedings, pp 1175–1180. https://doi.org/10.1109/EEEIC.2015.7165334

Burbano D (2015). Intrusive and non-intrusive load monitoring (a survey). Latin Am J Comput LAJC 2(1):45–53. Retrieved from https://mail-attachment.googleusercontent.com/attachment/u/0/?ui=2&ik=a8b4c7ac5b&view=att&th=14e3165cca971edb&attid=0.2&disp=inline&realattid=f_ibe10fe21&safe=1&zw&saddbat=ANGjdJ_AVSj69eWsgIDKUm0W9eHeDXeCzMksC_qs-I4333QRnHUGyg5RlEFzoxH62VE5QrorrWywBPR9t4B

Chang HH, Chien PC, Lin LS, Chen N (2011) Feature extraction of non-intrusive load-monitoring system using genetic algorithm in smart meters. In: Proceedings—2011 8th IEEE international conference on e-business engineering, ICEBE 2011, pp 299–304. https://doi.org/10.1109/ICEBE.2011.48

Department of Energy and Climate Change (2008) Climate change act, Tech rep, UK Retrieved from http://www.legislation.gov.uk/ukpga/2008/27/pdfs/ukpga_20080027_en.pdf

Darby S (2006) The effectiveness of feedback on energy consumption a review for defra of the literature on metering, billing and direct displays. Retrieved from http://www.eci.ox.ac.uk/research/energy/downloads/smart-metering-report.pdf

Dong R, Ratliff L, Ohlsson H, Sastry SS (2013) Fundamental limits of nonintrusive load monitoring. In: 3rd ACM international conference on high confidence networked systems (HiCoNS), pp 11–18. https://doi.org/10.1145/2566468.2576849

Egarter D, Bhuvana VP, Elmenreich W (2015) PALDi: online load disaggregation via particle filtering. IEEE Trans Instrum Meas 64(2):467–477. https://doi.org/10.1109/TIM.2014.2344373

Egarter D, Pöchacker M, Elmenreich W (2015) Complexity of power draws for load disaggregation 1–26. https://arxiv.org/abs/1501.02954

Eia (2017) Annual Energy Outlook 2017 with projections to 2050. Retrieved from https://www.eia.gov/outlooks/aeo/pdf/0383(2017).pdf

Froehlich J, Larson E, Gupta S, Cohn G, Reynolds M, Patel S (2011) Disaggregated end-use energy sensing for the smart grid. IEEE Perv Comput 10(1):28–39. https://doi.org/10.1109/MPRV.2010.74

Greveler U, Justus B, Loehr D (2012) Forensic content detection through power consumption. In: IEEE international conference on communications, pp 6759–6763. https://doi.org/10.1109/ICC.2012.6364822

Hart GW (1992) Nonintrusive appliance load monitoring. Proc IEEE 80(12):1870–1891

Inagaki S, Egami T, Suzuki T, Nakamura H, Ito K (2011) Nonintrusive appliance load monitoring based on integer programming. Electr Eng Jpn (English Translation of Denki Gakkai Ronbunshi) 174(2):1386–1392. https://doi.org/10.1002/eej.21040

Johnson MJ, Willsky AS (2013) Bayesian nonparametric hidden semi-Markov models. 14:673–701. arXiv Preprint Retrieved fromarxiv:1203.1365

Kamat PV (2007) Meeting the clean energy demand: nanostructure architectures for solar energy conversion. https://doi.org/10.1021/jp066952u

Kato T, Cho HS, Lee D, Toyomura T, Yamazaki T (2009) Appliance recognition from electric current signals for information-energy integrated network in home environments. In: Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 5597, LNCS, pp 150–157. https://doi.org/10.1007/978-3-642-02868-7_19

Katos V, Vrakas D, Katsaros P (2011) A framework for access control with inference constraints. In: Proceedings—international computer software and applications conference, pp 289–297. https://doi.org/10.1109/COMPSAC.2011.45

Kelly J, Knottenbelt W (2015a) Neural NILM: deep neural networks applied to energy disaggregation. In: Proceedings of the 2nd ACM international conference on embedded systems for energy-efficient built environments. ACM, pp 55–64

Kelly J, Knottenbelt W (2015b) The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. Sci Data 2:150007. https://doi.org/10.1038/sdata.2015.7

Kim H, Marwah M, Arlitt MF, Lyon G, Han J (2011) Unsupervised disaggregation of low frequency power measurements. In: Proceedings of the 11th SIAM international conference on data mining, pp 747–758. https://doi.org/10.1137/1.9781611972818.64

Kolter JZ, Johnson MJ (2011) REDD? A public data set for energy disaggregation research. SustKDD Workshop, no (1), pp 1–6. Retrieved from http://users.cis.fiu.edu/~lzhen001/activities/KDD2011Program/workshops/WKS10/doc/SustKDD3.pdf

Kolter Z, Jaakkola T (2012) Approximate inference in additive factorial HMMs with application to energy disaggregation. In: Proceedings of the international conference on artificial intelligence and statistics, vol XX, pp 1472–1482. Retrieved from http://people.csail.mit.edu/kolter/lib/exe/fetch.php?media=pubs:kolter-aistats12.pdf

Krizhevsky A, Sutskever I, Geoffrey EH (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems 25 (NIPS2012), pp 1–9. https://doi.org/10.1109/5.726791

Lai YX, Lai CF, Huang YM, Chao HC (2013) Multi-appliance recognition system with hybrid SVM/GMM classifier in ubiquitous smart home. Inf Sci 230:39–55. https://doi.org/10.1016/j.ins.2012.10.002

Lange H, Bergés M (2016) The neural energy decoder: energy disaggregation by combining binary subcomponents. In: NILM2016 3rd international workshop on non-intrusive load monitoring. Retrieved from nilmworkshop.org

Laughman C, Lee K, Cox R, Shaw S, Leeb S, Norford L, Armstrong P (2003) Power signature analysis. IEEE Power Energy Mag 1(2):56–63. https://doi.org/10.1109/MPAE.2003.1192027

Liang J, Ng SKK, Kendall G, Cheng JWM (2010) Load signature studypart I: basic concept, structure, and methodology. IEEE Trans Power Deliv 25(2):551–560. https://doi.org/10.1109/TPWRD.2009.2033799

Lin GY, Lee SC, Hsu JYJ, Jih WR (2010) Applying power meters for appliance recognition on the electric panel. In: Proceedings of the 2010 5th IEEE conference on industrial electronics and applications, ICIEA 2010, pp 2254–2259. https://doi.org/10.1109/ICIEA.2010.5515385

Makonin S, Popowich F (2015) Nonintrusive load monitoring (NILM) performance evaluation: a unified approach for accuracy reporting. Energy Effic 8(4):809–814. https://doi.org/10.1007/s12053-014-9306-2

Makonin S, Popowich F, Bajic IV, Gill B, Bartram L (2015) Exploiting HMM sparsity to perform online real-time nonintrusive load monitoring. IEEE Trans Smart Grid. https://doi.org/10.1109/TSG.2015.2494592

Makonin S, Popowich F, Bartram L, Gill B, Bajic IV (2013) AMPds: a public dataset for load disaggregation and eco-feedback research. In: 2013 IEEE electrical power and energy conference, EPEC 2013. https://doi.org/10.1109/EPEC.2013.6802949

Marchiori A, Hakkarinen D, Han Q, Earle L (2011) Circuit-level load monitoring for household energy management. IEEE Perv Comput 10(1):40–48. https://doi.org/10.1109/MPRV.2010.72

Mauch L, Yang B (2015) A new approach for supervised power disaggregation by using a deep recurrent LSTM network. In proceedings of the 3 rd IEEE global conference on signal and information processing (GlobalSIP), pp 63–67

Paradiso F, Paganelli F, Giuli D, Capobianco S (2016) Context-based energy disaggregation in smart homes. Future Internet 8(1):4. Retrieved from http://www.mdpi.com/1999-5903/8/1/4

Parson O, Ghosh S, Weal M, Rogers A (2011) Using hidden Markov models for iterative non-intrusive appliance monitoring. Electronics and Computer Science, University of Southampton, Hampshire, UK, pp 1–4. Retrieved from http://eprints.soton.ac.uk/272990/

Parson O, Ghosh S, Weal M, Rogers A (2012) Non-intrusive load monitoring using prior models of general appliance types. In: Proceedings of the 26th AAAI conference on artificial intelligence, pp 356–362

Parson O, Ghosh S, Weal M, Rogers A (2014) An unsupervised training method for non-intrusive appliance load monitoring. Artif Intell 217:1–19. https://doi.org/10.1016/j.artint.2014.07.010

Polat H, Du W, Renckes S, Oysal Y, Polat H, Renckes S, Du W (2010) Private predictions on hidden Markov models. Artif Intell Rev 34:53–72. https://doi.org/10.1007/s10462-010-9161-2

Ruzzelli AG, Nicolas C, Schoofs A, O'Hare GMP (2010) Real-time recognition and profiling of appliances through a single electricity sensor. In: SECON 2010—2010 7th Annual IEEE communications society conference on sensor, mesh and ad hoc communications and networks. https://doi.org/10.1109/SECON.2010.5508244

Sankar L, Raj Rajagopalan S, Mohajer S, Vincent Poor H (2013) Smart meter privacy: a theoretical framework. IEEE Trans Smart Grid 4(2):837–846. https://doi.org/10.1109/TSG.2012.2211046

Srinivasan D, Ng WS, Liew AC (2006) Neural-network-based signature recognition for harmonic source identification. IEEE Trans Power Deliv 21(1):398–405

Wytock M, Kolter JZ (2013) Contextually supervised source separation with application to energy disaggregation. In: Twenty-eighth AAAI conference on artificial intelligence, pp 1–10. Retrieved from arXiv:1312.5023

Zachary CL (2015) A critical review of recurrent neural networks for sequence learning. arXiv Preprint, pp 1–35. https://doi.org/10.1145/2647868.2654889

Zeifman M (2012) Disaggregation of home energy display data using probabilistic approach. IEEE Trans Consum Electron 58(1):23–31. https://doi.org/10.1109/TCE.2012.6170051

Zeifman M, Roth K (2011) Nonintrusive appliance load monitoring: review and outlook. IEEE Trans Consum Electron 57(1):76–84. https://doi.org/10.1109/TCE.2011.5735484

Zhang C, Zhong M, Wang Z, Goddard N, Sutton C (2017) Sequence-to-point learning with neural networks for non-intrusive load monitoring. Retrieved from https://pdfs.semanticscholar.org/b519/ffec16d6256b872c6c108023a64d02646293.pdf

Zhong M, Goddard N, Sutton C (2015) Latent Bayesian melding for integrating individual and population models. In: Advances in neural information processing systems, pp 3618–3626. http://papers.nips.cc/paper/5756-latent-bayesianmelding-for-integrating-individual-and-population-models.pdf

Zia T, Bruckner D, Zaidi A (2011) A hidden Markov model based procedure for identifying household electric loads. In: IECON proceedings (industrial electronics conference), pp 3218–3223. https://doi.org/10.1109/IECON.2011.6119826

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