An IEEE Societies and Technical Councils Publication
IEEE Data Descriptions

Received 1 October 2024; revised 7 October 2024; accepted 7 October 2024; Date of publication 10 October, 2024; date of current version 7 November, 2024.

Digital Object Identifier 10.1109/IEEEDATA.2024.3478184

Meta: *Defining the Disaggregated Component Assignation Error Metric for Complex Time-Series Signal Data*

STEPHEN MAKONIN [1] (SENIOR MEMBER IEEE) AND ALEJANDRO RODRIGUEZ-SILVA [2]

[1]School of Engineering Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
[2]MDA Space, Richmond, BC V6V 2J3, Canada

CORRESPONDING AUTHOR: Stephen Makonin (e-mail: smakonin@sfu.ca).

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2018-06192.

ABSTRACT Researchers designing disaggregation algorithms have constant debates as to what accuracy and error metrics to use to evaluate/measure performance. What is the best measure of the classification accuracy of the disaggregated components? What is the best way to measure the error in the magnitude of each signal component? In some cases, metrics that measure regression (for example, power consumption estimation) tend to report better than actual performance. We propose a novel metric based on quadratic programming that we coined the disaggregated component assignation error (DCAE). DCAE (pronounced like the word *decay*) is suitable for blind source separation problems such as unsupervised disaggregation because it is robust under a set of fundamental test cases for disaggregation. The main motivation for this metric is to detect poor unsupervised disaggregation performance in cases where traditional classification or estimation metrics cannot. DCAE is tested using time-series power data with the classical disaggregation problem of nonintrusive load monitoring (NILM). DCAE demonstrates automatically matching unsupervised disaggregated appliance power readings to their corresponding ground-truth components.

IEEE SOCIETY/COUNCIL Power and Energy Society (PES), Signal Processing Society (SPS)

DATA TYPE/LOCATION Time-Series, Signals; n/a

DATA DOI/PID n/a

INDEX TERMS Blind source separation, disaggregation, error measurement, nonintrusive load monitoring (NILM), unsupervised learning.

BACKGROUND

Disaggregation the ill-defined problem of an inverse summation algorithm. Similar to the blind source separation problem. However, different in that disaggregation attempts to blindly separate signals that are nonstationary. Mathematically, disaggregation is as follows:

$$y_t = \sum_{m=1}^{M} y_t^{(m)} + \epsilon \tag{1}$$

where $y_t^{(m)}$ is the magnitude at time t associated with signal m out of a total of M signals that constitute the total magnitude y_t, and ϵ is some measurement noise.

A typical disaggregation system [1], [2], [3], [4], [5], [6], [7], [8] is depicted in Fig. 1. Here, the aggregate appliance power reading y_t is composed of M ground-truth appliance power readings $y_t^{(m)}$.

The aggregate appliance power reading is then passed through a filtering process (resulting in $\hat{y}_t$) to remove noise and transient behavior that often interferes with proper appliance classification/tracking. Once power values are more *square*—in nature—to help identify steady-states, a source separation algorithm (i.e., disaggregator) extracts the edges (i.e., events) from $\hat{y}_t$ and separates it into N disaggregated appliance power readings $\hat{y}_t^{(n)}$. Next, disaggregation

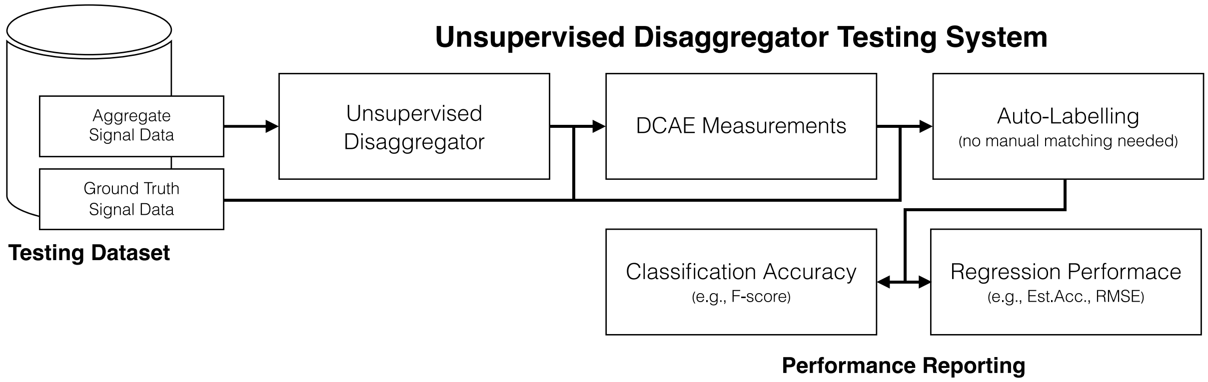


FIG. 1. Generalized unsupervised disaggregation system with the added DCAE measurement step which then replaces *manual signal labeling* to now being *auto-labeling*.

performance is evaluated by comparing the error between the ground-truth appliance power readings and the disaggregated appliance power readings. Finally, labels are applied to the disaggregated appliance power readings by some sort of appliance signature database [2], [9].

Unsupervised learning methods of disaggregation do not require individual appliance data. Moreover, these methods are more robust than supervised learning techniques due to the fact that many different homes have many different appliance mixes and they adapt but tuning a model in an unseen house overtime time [10]. This approach is often called *unsupervised disaggregation*.

Classification Accuracy

Classification accuracy in the context of unsupervised disaggregation is a critical metric used to evaluate the performance of algorithms that aim to separate aggregated data into meaningful components. Unsupervised disaggregation involves dividing a composite signal into its individual sources without prior knowledge about the nature or structure of these sources. This process is essential in various applications, such as energy disaggregation, where the goal is to break down total household energy consumption into specific devices' usage patterns without labeled training data. High-classification accuracy indicates that the disaggregation algorithm can reliably identify and separate each component accurately, thus providing valuable insights into the underlying data.

Achieving high-classification accuracy in unsupervised disaggregation can be challenging due to the complexities and variations in the data. For instance, in the case of energy disaggregation, different household devices can have overlapping usage patterns, varying power consumption levels, and noise interference. These factors make it difficult to pinpoint the exact contribution of each device to the aggregate signal. To overcome these challenges, advanced machine learning algorithms, signal processing techniques, and domain-specific knowledge are often employed. Evaluating these algorithms' performance typically involves comparing the disaggregated results with actual ground truth data, if available, or using metrics such as precision, recall, F1-score, and root mean square error (RMSE) to gauge their efficiency.

Improving classification accuracy in unsupervised disaggregation not only enhances the reliability of the results but also increases the potential applications of these methods. For example, more accurate energy disaggregation can lead to better energy conservation strategies by providing consumers with detailed insights into their consumption patterns. Similarly, in other fields such as audio signal processing, accurately separating different sound sources can enhance speech recognition systems and improve audio quality. As research in this area progresses, the development of more sophisticated models and algorithms is expected to further increase the classification accuracy, making unsupervised disaggregation a more viable solution for real-world problems.

Error Metrics

Error metrics for unsupervised disaggregation tasks are essential to evaluate the performance of models that aim to separate aggregated data into its constituent components without prior labels. These metrics provide insights into how accurately the model can identify and quantify the individual sources from the mixed signal. In unsupervised disaggregation, common error metrics include reconstruction error, signal-to-noise ratio (SNR), and clustering effectiveness. Each of these metrics offers a different perspective on the model's performance, helping researchers and practitioners to fine-tune their algorithms for better accuracy and reliability.

One of the key error metrics is the reconstruction error, which measures the difference between the original aggregated signal and the reaggregated signal from the separated components. Lower reconstruction error indicates that the model has effectively captured the underlying structure of the data and can accurately reconstruct the original signal from its components. Additionally, SNR assesses the proportion of the desired signal to the background noise after disaggregation. High-SNR values suggest that the model has successfully isolated the individual sources with minimal noise, improving the overall quality of the disaggregation process.

Clustering effectiveness is another crucial metric in unsupervised disaggregation. It evaluates how well the separated components are grouped together, reflecting the model's ability to distinguish between different sources in the aggregated data. Metrics such as silhouette score [11], Davies–Bouldin index [12], and Dunn index [13] are commonly used to assess clustering quality. A higher silhouette score indicates that the separated components are well clustered and distinct, which is desirable for accurate disaggregation. These error metrics collectively help in assessing and enhancing the performance of unsupervised disaggregation models, guiding improvements in data preprocessing, feature extraction, and algorithm design.

Motivation

Disaggregation is both a classification problem and a regression problem; thus, measuring performance accuracy is not easily defined—there exists no consensus within the disaggregation community [14], [15]. Many of the performance metrics are adopted from other areas and applied to disaggregation without a lack of deep understanding on the behavior of accuracy and errors metrics.

According to [15], all the classification metrics assume all the events are equally important, which is not realistic since all not appliances require the same energy or even targeted to reduce energy usage. Thus, we can already observe how these metrics result are problematic as they only contain appliance state information without providing any power information. On the other hand, estimation metrics allow us to measure how accurately the disaggregation algorithm performs by providing overall error between the ground-truth and the disaggregated appliance power readings. Note how these metrics assume that the matching between the disaggregated appliance power readings with their respective ground-truth appliance power readings has been previously done.

PROPOSED METHODOLOGY

Measuring the performance of disaggregation is equally a challenge [14], [15]. This challenge ranges from a lack of community consensus to flaws in performance measurement. On the classification side of disaggregation, performance metrics have been studied extensively [15]. Here, we turn the attention to error metrics on the regression side where disaggregation tries to estimate the power consumption of each disaggregated appliance/load. Our dive into regression metrics in disaggregation has lead us to develop a new more robust metric we call disaggregated component assignation error (DCAE). DCAE is a generalized version of the work taken from Rodriguez-Silva's thesis [16] first labeled the *power assignation metric*.

A disaggregation performance metric must have at least these fundamental attributes.

1) *Tracking Power Assignation*: Any metric must measure how well the inferred power of each disaggregated appliance power reading matches the corresponding ground-truth appliance power reading.
2) *Output Permutation Robustness*: Any metric value should be invariant due to permutations in the ground-truth and unsupervised disaggregated appliance power readings.
3) *Input Cardinality Agnostic*: As the number of unsupervised disaggregated outputs does not necessarily match the number of ground-truth appliance power readings, any metric must evaluate the matching of one or more unsupervised disaggregated appliance power reading to a ground-truth appliance.

Given T observations of M appliances $y_t^{(m)}$, the set of ground truth appliances, and the associated N inferred disaggregated appliance power readings $\hat{y}_t^{(n)}$. We define the DCAE error measurement

$$
\begin{aligned}
E(y,\hat{y}) = \min_{\alpha_{mn}} \Bigg[& \overbrace{\sum_{t=1}^{T}\sum_{m=1}^{M}\left(y_t^{(m)} - \sum_{n=1}^{N}\alpha_{mn}\hat{y}_t^{(n)}\right)^2}^{\text{Part 1}} \\
& + \lambda \overbrace{\sum_{m=1}^{M}\sum_{n=1}^{N}\alpha_{mn}}^{\text{Part 2}} \Bigg] \\
& + \underbrace{\min_{\beta_{mn}}\left[\sum_{t=1}^{T}\sum_{m=1}^{M}\left(y_t^{(m)} - \sum_{n=1}^{N}\beta_{mn}\hat{y}_t^{(n)}\right)^2\right]}_{\text{Part 3}}
\end{aligned} \tag{2}
$$

subject to

$$\sum_{m=1}^{M}\beta_{mn} = 1 \quad \forall n \tag{3}$$

$$\alpha_{mn}, \beta_{mn} \in \{0, 1\} \quad \forall m, n. \tag{4}$$

The objective is to find the set of coefficients α_{mn} and β_{mn} that minimize the error expression in (2), which can be formulated as a QP problem [17]—binary assignation. Intuitively, the α_{mn} coefficients in (2), *part 1*, are binary indicators of whether the inferred output appliance power reading $\hat{y}_t^{(n)}$ should be assigned or not to the ground truth appliance power reading $y_t^{(m)}$. We aim to find the minimum error between each of the ground truth appliance power readings and the sum of the disaggregated output appliance power readings by assigning each of the output appliance power readings to one input appliance power reading.

The term in (2), *part 2*, is a penalizing term. It penalizes appliances that are disaggregated in their single-state components. The term *part 1* will aim to assign the inferred appliance power readings to their corresponding ground-truth appliance power readings. However, this comes with a *cost*, as the term *part 2* will increase as more inferred appliance power readings are assigned to the ground-truth appliance power readings. If the parameter λ is set to be very large, all of the terms in *parts 1* and *2* will get zeroed out, except for the ground-truth appliance power readings $y_t^{(m)}$ in the term *part 1*. Therefore, the minimization will be the sum of these $y_t^{(m)}$ terms. On the other hand, if the parameter λ is set to a value close to zero, it lessens the penalization for disaggregating multistate appliances.

The term in (2), *part 3*, operates similarly to *part 1* of the minimization problem. The β_{mn} coefficients are binary indicators to whether the inferred output appliance power reading $\hat{y}_t^{(n)}$ should be assigned or not to the ground truth appliance power reading $y_t^{(m)}$. However, the sum of the β_{mn} overall M the possible ground-truth appliance power readings should be equal to 1, for all the inferred appliance power readings. In simple words, we are forcing the $\hat{y}_t^{(n)}$ inferred appliance power reading to be assigned to one $y_t^{(m)}$ ground-truth appliance power reading (3). In *part 1*, there is no need to constrain the α_{mn} coefficients as doing so would not allow for the regularization term in *part 2* to perform its task.

The error metric in expression (2) can suffer from numerical instability for three reasons: 1) we are working with power appliance power readings; 2) we have a square term for *parts 1* and *3*; and 3) we are summing overall the appliances and all the time instants which means the equation in (2) grows rapidly. A scaling factor γ is introduced to avoid numerical instability. This scaling factor consists of summing overall the total energy square of each of the $y_t^{(m)}$ ground truth appliance power readings as

$$\gamma = \sum_{m=1}^{M}\sum_{t=1}^{T}(y_t^{(m)})^2 + \epsilon \tag{5}$$

where ϵ is a small number, for this work it was set to $\approx 1 \times 10^{-3}$. Therefore, the error metric (2) is divided by the scaling factor introduced in (5) to allow for numerical stability. The numerical solver that was used in "Applied Analysis" section does not support QP with integer constraints, so the condition in (4) has to be relaxed for the α_{mn} and β_{mn} coefficients to take values in the interval $[0, 1]$. The error metric in (2) (i.e., DCAE) can be rewritten by incorporating the scaling factor in (5) and by relaxing the constraint in (2) as

$$E_s(y, \hat{y}) = \min_{\alpha_{mn}}\left[\frac{\sum_{t=1}^{T}\sum_{m=1}^{M}\left(y_t^{(m)} - \sum_{n=1}^{N}\alpha_{mn}\hat{y}_t^{(n)}\right)^2}{\gamma} + \frac{\lambda\sum_{m=1}^{M}\sum_{n=1}^{N}\alpha_{mn}}{\gamma}\right] + \min_{\beta_{mn}}\left[\frac{\sum_{t=1}^{T}\sum_{m=1}^{M}\left(y_t^{(m)} - \sum_{n=1}^{N}\beta_{mn}\hat{y}_t^{(n)}\right)^2}{\gamma}\right] \tag{6}$$

subject to

$$\sum_{m=1}^{M}\beta_{mn} = 1 \quad \forall n \tag{7}$$

$$\alpha_{mn}, \beta_{mn} \in [0, 1] \quad \forall m, n. \tag{8}$$

APPLIED ANALYSIS

We want to test how well the metric introduced in (6) performs under several possible test case scenarios that appear when attempting appliance disaggregation. Although the appliance power readings that were used are composed of short samples from the RAE dataset [18], they help us to illustrate the performance of DCAE in several fundamental cases to test its robustness. In the four test cases as follows, we used the two appliances: fridge and dishwasher.

DCAE was coded and tested in Python 3.7. All tests ran on a Mac Pro (2017 model) with a 2.3 GHz Intel Core i5 processor and 8 GB of memory. The numerical solver `CVXPY` [19], [20] was used for solving the metric in (6) and integrating QP. The `CVXPY` [19], [20] Python solver was constrained to conditions (7) and (8). It allowed us to relax the condition in (4), meaning that coefficients can now be

$$\alpha_{mn} \in \mathbb{R} : 0 \geq \alpha_{mn} \geq 1$$

and

$$\beta_{mn} \in \mathbb{R} : 0 \geq \beta_{mn} \geq 1.$$

Given that the `CVXPY` is convex quadratic solver the total runtime complexity is $O(L^2 n^4)$, where L is input bits and n are the variables—meaning that the solution is solved in weak-polynomial time.

We take data from house-1, block-1, (nine days) from the RAE dataset [18]. In each case study, nonpreprocessed appliance power readings containing 2300, 1 Hz samples where given to the optimizer.

Fundamental Test Case 1

Suppose the aggregate appliance power reading y_t is composed of two appliance appliance power readings $y_t^{(1)}$ and $y_t^{(2)}$. We obtain the inferred appliance power reading $\hat{y}_t^{(1)}$ which is the aggregate of two appliances [see Fig. 2(e)]. This is *disagg1*.

To evaluate how *disagg1* performed, we could simply compute the RMSE which would be zero—concluding the performance without error. However, this is misleading as the disaggregator never disaggregates the two appliance power readings [Fig. 2(f)] when it should have [Fig. 2(g)]. This disaggregator is *disagg2*.

Our experiment found that RMSE inaccurately scored zero for both *disagg1* and *disagg2*, whereas DCAE resulted in an error of 0.08 for *disagg1* and 0.00 for *disagg2*. Table I shows a comparison between the RMSE and DCAE computed to both *disagg1* and *disagg2*. By default, we use the penalizing term $\lambda = 1$. DCAE is able to penalize poor performance that RMSE is not able to detect. DCAE being characterized

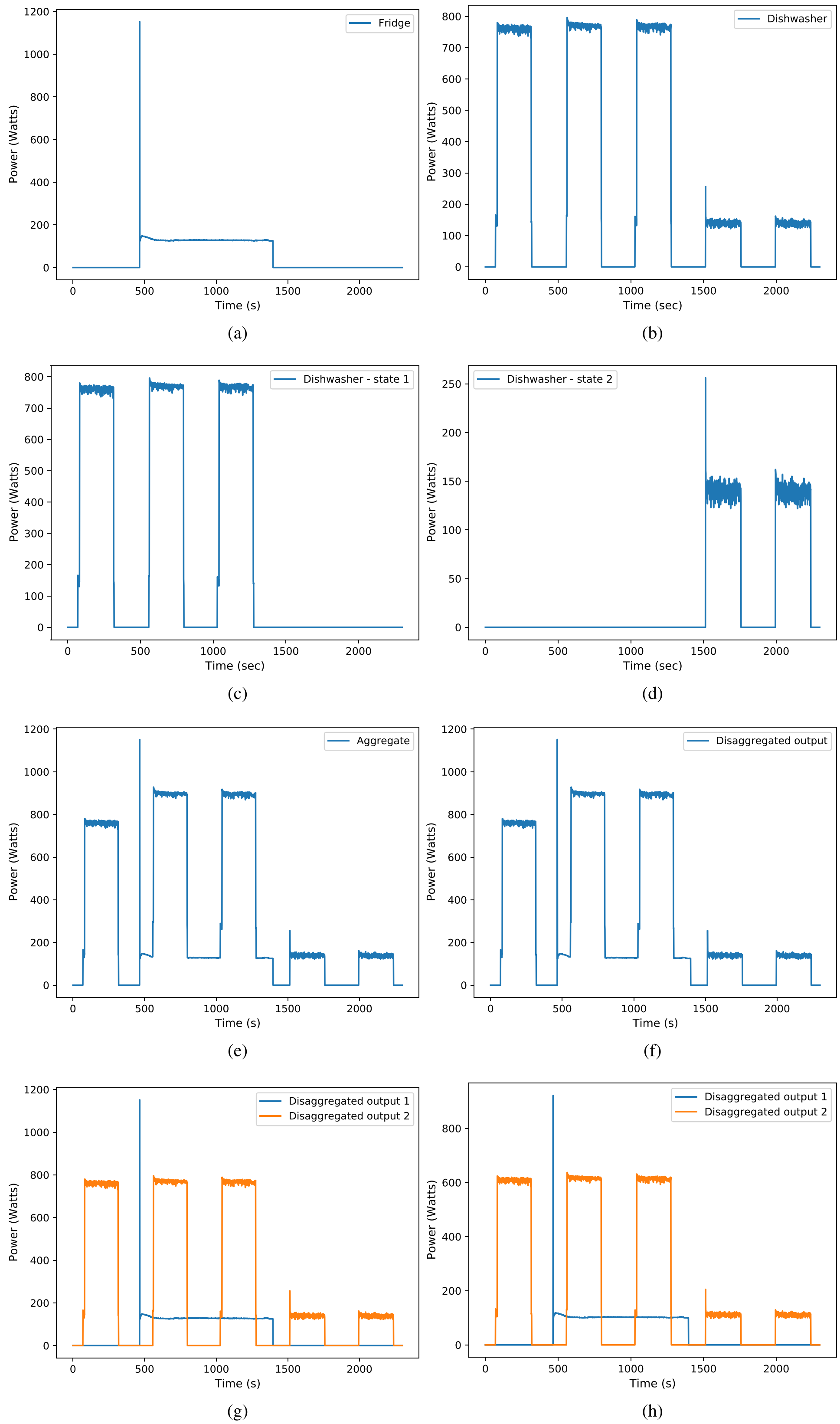


FIG. 2. Ground-truth and inferred appliance power readings for the different *test cases*: (a) one cycle of the fridge; (b) two states of the dishwasher; (c) first operational mode of the dishwasher; (d) second operational mode of the dishwasher; (e) aggregate appliance power reading composed of the fridge and the dishwasher; (f) example of poor disaggregation performance where the inferred appliance power reading is the same as the aggregate fed into the disaggregator; (g) example of optimal disaggregation where inferred appliance power readings match the ground-truth; and (h) example of scaled disaggregation. The inferred appliance power readings are a scaled version of the ground-truth appliance power readings.

TABLE I. **Case 1 Error Results**

Error Metric	disagg1	disagg2
RMSE	0.0000	0.0000
Proposed	0.0843	0.0000

as an assignation problem means β_{mn} coefficients can be interpreted later, as to *what disaggregated appliance power reading corresponds to what ground-truth appliance power reading*; i.e., DCAE automatically assigns the inferred or disaggregated appliance power readings to the ground-truths. We can express these β_{mn} coefficients in a $(M \times N)$ matrix as matrix $\mathbf{B}$ with entries

$$\mathbf{B} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1N} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{M1} & \beta_{M2} & \cdots & \beta_{MN} \end{bmatrix}. \tag{9}$$

The resulting matrix after computing (6) can be expressed as

$$\mathbf{B}_{disagg2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{10}$$

where m denotes ground-truth appliance power readings (rows) and n denotes inferred appliance power reading (columns). For *disagg2*, there are two ground-truth appliance power readings and two inferred appliance power readings. The $\beta_{11} = 1$ coefficient can be interpreted as the inferred appliance power reading $\hat{y}_t^{(1)}$ can be assigned to the ground-truth appliance power reading $y_t^{(1)}$. The coefficient $\beta_{22} = 1$ can be similarly interpreted as the inferred appliance power reading $\hat{y}_t^{(2)}$ can be assigned to the ground-truth appliance power reading $y_t^{(2)}$. Coefficients $\beta_{12}, \beta_{21} = 0$ can be interpreted as the inferred appliance power reading $\hat{y}_t^{(1)}$ cannot be assigned to the ground-truth appliance power reading $y_t^{(2)}$ and the inferred appliance power reading $\hat{y}_t^{(2)}$ cannot be assigned to the ground-truth appliance power reading $y_t^{(1)}$, respectively. Similar reasoning can be applied to different for different M and N values. From this point on, the β_{mn} coefficients will not be written for each of the test cases. Nevertheless, the reader can assume the inferred appliance power readings were matched correctly with their corresponding input appliance power readings. In other words, the β_{mn} coefficients have been assigned correctly.

Fundamental Test Case 2

DCAE should be invariant to the order of the ground-truth and inferred appliance power readings. Suppose the aggregate appliance power reading is composed of two ground-truths $y_t^{(1)}$ [Fig. 2(a)] and $y_t^{(2)}$ [Fig. 2(b)]. After disaggregation we obtain the inferred appliance power readings $\hat{y}_t^{(1)}$ and $\hat{y}_t^{(2)}$ [Fig. 2(g)]. Let us denote a new disaggregator *disagg3* in which we have the same aggregate appliance power reading as in *disagg2*, composed of the same ground-truth appliance power readings. The inferred appliance power readings $\hat{y}_t^{(1)}$ and $\hat{y}_t^{(2)}$ are the same as the ones depicted in Fig. 2(g). Nevertheless, with the distinction that the inferred appliance power readings are permuted (i.e., $\hat{y}_t^{(1)}$ belongs to the dishwasher and $\hat{y}_t^{(2)}$ belongs to the fridge). Both *disagg2* and *disagg3* should have the same error metric regardless of the order, in which the appliance power readings appear in the disaggregator system. The results (Table II) of our experiment show the DCAE score for *disagg2* and *disagg3* is zero.

TABLE II. **Proposed Metric Values for *disagg2* and *disagg3***

Error Metric	disagg2	disagg3
Proposed	0.0	0.0

TABLE III. **Proposed Metric Values for *disagg2* and *disagg4***

Error Metric	disagg2	disagg3
Proposed	0.004497	0.006750

TABLE IV. **Case 4 Accuracy Results**

Error Metric	disagg2	disagg5
F1-Score (fridge)	100%	100%
F1-Score (dishwasher)	100%	100%
Proposed	0.00000	0.06489

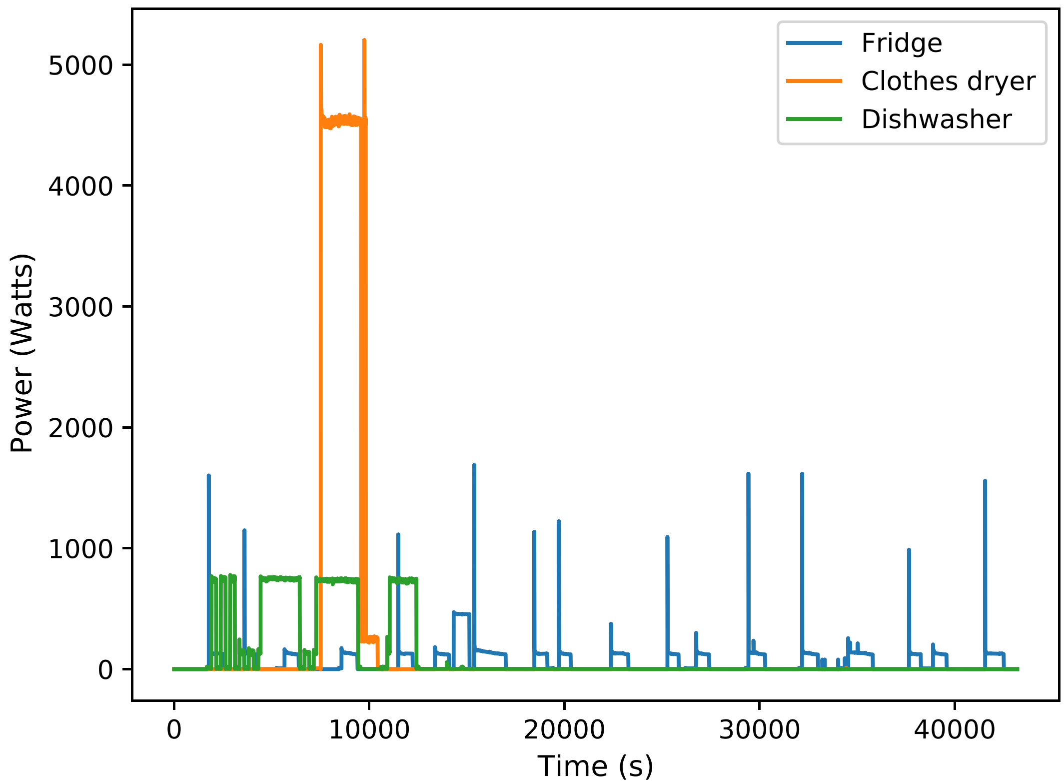
TABLE V. **Matching Disaggregated Test Results versus Ground-Truth**

Inferred Signals	Matched Ground-Truth
Inferred load 1	Clothes dryer
Inferred load 2	Dishwasher
Inferred load 3	Dishwasher
Inferred load 4 (zero vector)	-
Inferred load 5	Dishwasher
Inferred load 6	Fridge
Inferred load 7	Fridge
Inferred load 8	Dishwasher

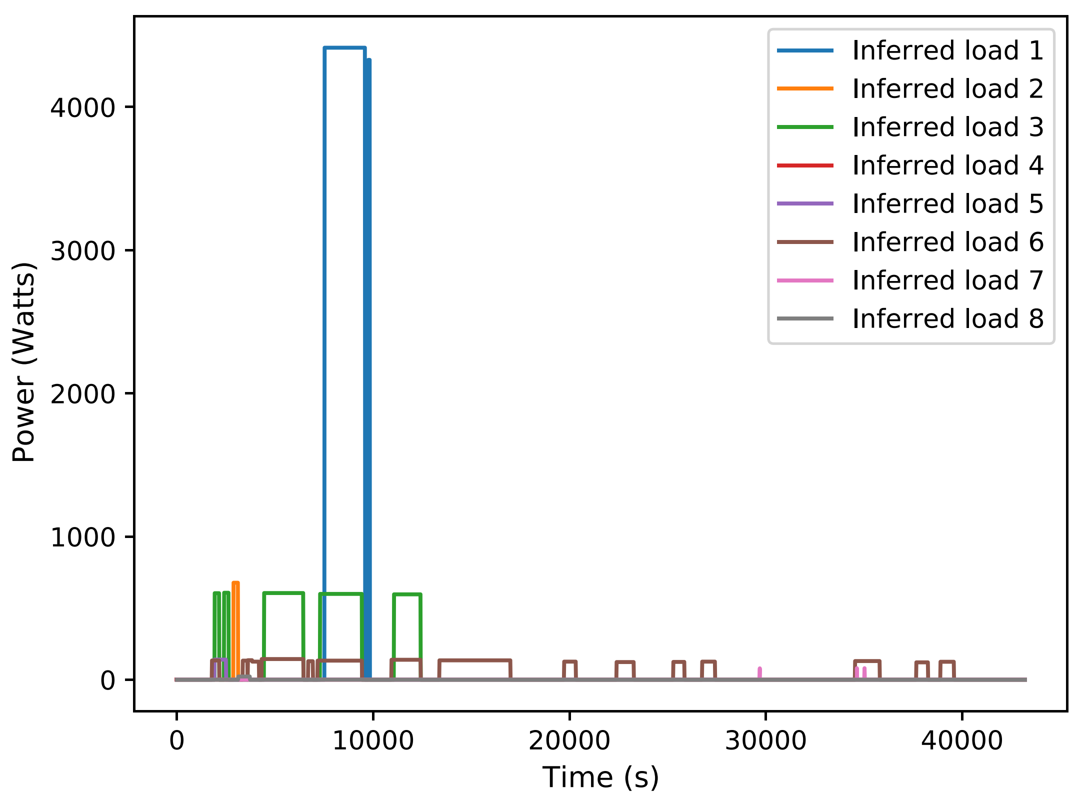
Note: Charts shown in Fig. 3.

Fundamental Test Case 3

DCAE should be robust and penalize scenarios where the appliances are disaggregated in their single state components. Suppose we have the same aggregate as in test cases 1 and 2. The aggregate is composed of the two ground-truths $y_t^{(1)}$ and $y_t^{(2)}$ [Fig. 2(a) and 2(b)]. Our new disaggregator *disagg4* decomposed the aggregate appliance power reading in three inferred appliance power readings: $\hat{y}_t^{(1)}$, $\hat{y}_t^{(2)}$, and $\hat{y}_t^{(3)}$. The first inferred appliance power reading $\hat{y}_t^{(1)}$ belongs to the fridge [Fig. 2(a)]. The inferred appliance power readings $\hat{y}_t^{(2)}$ and $\hat{y}_t^{(3)}$ are the first- and second-operational modes of



(a)



(b)

FIG. 3. (a) Inferred signals are a scaled version of the ground-truth signals. (b) Disaggregation results using [1] as compared to top chart.

the dishwasher [Fig. 2(c) and 2(d)], respectively. Technically speaking, the *disagg4* is not performing poorly. However, it is decomposing one of the appliances contributing to the aggregate into its two operational mode. This should be reflected in a metric. When comparing the performance of *disagg2* and *disagg4* using DCAE the resulting error is 0.004 and 0.006 (respectively) as shown in Table III. It can be observed that the *disagg4* has a larger error value. The penalizing weight was set to $\lambda = 1 \times 10^6$. The penalizing parameter λ can be set according to "how much" we want to emphasize cases where multistate appliances are broken into their single operational modes. Note how clustering algorithms such as the one described in [1], [2], and [3] will disaggregate multistate appliances into a set of single-state appliances, a scenario worth addressing.

Fundamental Test Case 4

DCAE should be a power variant. If a disaggregator decomposes an appliance power reading into a scaled version of the ground-truth appliance power readings, a robust error metric should be able to detect such scaling. Suppose that our aggregate appliance power reading is composed of two ground-truth appliance power readings $y_t^{(1)}$ [Fig. 2(a)] and $y_t^{(2)}$ [Fig. 2(b)]. The aggregate appliance power reading is decomposed into a scaled version of the ground truth appliance power readings. Therefore,

$$\hat{y}_t^{(1)} = s \cdot y_t^{(1)} \text{ and } \hat{y}_t^{(2)} = s \cdot y_t^{(2)}, \text{ s.t. } s \in \mathbb{R}^+.$$

For this example, we set $s = 0.8$ and call this *disagg5*. Classification metrics such as F1-score are invariant to cases where the inferred appliance power readings are ON/OFF at the same time instants as their corresponding ground-truth appliance power readings, yet the power difference between both can be off by a large variation. The inferred appliance power readings $\hat{y}_t^{(1)}$ (scaled version of the fridge appliance power reading) and $\hat{y}_t^{(2)}$ (scaled version of the dishwasher appliance power reading) are depicted in Fig. 2(h). Our experiment resulted in *disagg2* and *disagg5* having and F1-score of 1.0, while DCAE resulted in an error score of 0.00 and 0.06 (see Table IV). The F1-score is not able to detect the differences between the ground-truth and the inferred appliance power readings.

DISCUSSION OF RESULTS

We demonstrated the robustness of DCAE in several fundamental cases. Naturally, we need to test the proposed metric using a state-of-the-art disaggregator [1] with a larger and more complex power signal. Jones [1] reports good performance using unsupervised clustering based on nonparametric Gaussian mixture models (GMMs).

Twelve hours of 1 Hz data (43.2k samples) from the RAE dataset [18] was selected for three appliances: fridge, clothes dryer, and dishwasher. Data were partitioned in three 4-h windows containing 14.4k samples. We selected periods where all the appliances were running simultaneously to better asses DCAE's performance as in Fig. 3(b). The raw aggregate signal passes through a filtering stage, where we selected the method from [21] to perform this task. The filtered aggregate signal is then fed to [1], which produces the signals as shown in Fig. 3(a).

Window 1 results in a low-DCAE score (0.009) because the disaggregator accurately decomposes all appliances for the first four days. Windows 2 and 3 result in high-DCAE scores (0.788 and 0.631) as the disaggregator was not able to reconstruct the fridge signal for the final 8 h. Inferred appliances were matched by comparing binary coefficients (see Table V).

Matrices $\mathbf{B}_{w_1}$, $\mathbf{B}_{w_2}$, and $\mathbf{B}_{w_3}$ contain the binary coefficients for the first, the second, and the third windows, respectively. These coefficients can be interpreted as the matches between inferred and ground-truth signals. For example,

$$\mathbf{B}_{w_1} = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{3} & 0.619 & 0.352 & \frac{1}{3} & 0.193 \\ 1 & 0 & 0 & \frac{1}{3} & 0.157 & 0.180 & \frac{1}{3} & 0.312 \\ 0 & 1 & 1 & \frac{1}{3} & 0.224 & 0.467 & \frac{1}{3} & 0.495 \end{bmatrix}. \quad (11)$$

From the columns, the first three inferred signals have the largest energies Inferred signals 5 and 6 were swapped for the fridge and dishwasher because these two appliances have operational modes with similar power values. The inferred signal 7 has equal weights. Inferred signal 8 was assigned correctly. Manually matching the disaggregated appliances with the ground-truth signals is challenging due to its behavior, resulting in incorrect matching of signals.

CONCLUSION

The DCAE metric presented here is a good alternative to the traditional classification and estimation metrics. Although we presented the metric focusing on disaggregation, it can be used for other research areas where source separation is the main application.

The metric not only proves to be robust under fundamental test cases but also automatically assigns the disaggregated appliance power readings with their corresponding ground-truth appliance power readings, even when the number of disaggregated and ground-truth appliance power readings is not equal. Moreover, assuming the ground-truth appliance power readings are labeled, DCAE also solves the labeling problem in disaggregation, which is always left as future work or completely neglected.

In the case of NILM, DCAE provides a more robust way to measure appliance power estimation errors. The second main contribution is the ability to automatically match disaggregated appliance power readings to the corresponding ground-truth appliance power readings. In unsupervised NILM (as with [9]), this appliance power reading matching is always performed manually, which is time consuming and can suffer greatly from human error. Our third and final contributions are that the proposed metric also solves the labeling problem in unsupervised NILM when ground-truth appliance power reading contains their corresponding labels.

SOURCE CODE AND SCRIPTS

All DCAE-implementation scripts (originally known as PAM) have been made publicly available through a GitHub repository. To access the scripts, please follow the link provided: https://github.com/compsust/DCAE.

ACKNOWLEDGMENT

All authors wrote and reviewed the manuscript.

The authors have declared no conflicts of interest.

REFERENCES

[1] R. Jones, "Non-parametric modeling in non-intrusive load monitoring," Master's thesis, Appl. Sci.: Sch. Eng. Sci., Burnaby, BC, Canada, Burnaby, BC, Canada, 2020.

[2] B. Zhao, L. Stankovic, and V. Stankovic, "On a training-less solution for non-intrusive appliance load monitoring using graph signal processing," *IEEE Access*, vol. 4, pp. 1784–1799, 2016.

[3] M. Qureshi, C. Ghiaus, and N. Ahmad, "A blind event-based learning algorithm for non-intrusive load disaggregation," *Int. J. Elect. Power Energy Syst.*, vol. 129, 2021, Art. no. 106834.

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

[5] R. Jia, Y. Gao, and C. J. Spanos, "A fully unsupervised non-intrusive load monitoring framework," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Miami, FL, USA. Piscataway, NJ, USA: IEEE, 2015, pp. 872–878.

[6] T. Bernard and M. Marx, "Unsupervised learning algorithm using multiple electrical low and high frequency features for the task of load disaggregation," in *Proc. 3rd Int. Workshop NILM*, Vancouver, BC, Canada, website/online, 2016, pp. 1–5.

[7] A. Cominola, M. Giuliani, D. Piga, A. Castelletti, and A. E. Rizzoli, "A hybrid signature-based iterative disaggregation algorithm for non-intrusive load monitoring," *Appl. Energy*, vol. 185, pp. 331–344, 2017.

[8] G. Jacobs and P. Henneaux, "Unsupervised learning procedure for NILM applications," in *Proc. IEEE 20th Mediterr. Electrotech. Conf. (MELECON)*, Palermo, Italy. Piscataway, NJ, USA: IEEE, 2020, pp. 559–564.

[9] A. Rodriguez-Silva and S. Makonin, "Universal non-intrusive load monitoring (UNILM) using filter pipelines, probabilistic knapsack, and labelled partition maps," in *Proc. IEEE PES Asia-Pacific Power Energy Eng. Conf. (APPEEC)*, Macau, China. Piscataway, NJ, USA: IEEE, 2019, pp. 1–6.

[10] R. Bonfigli, S. Squartini, M. Fagiani, and F. Piazza, "Unsupervised algorithms for non-intrusive load monitoring: An up-to-date overview," in *Proc. IEEE 15th Int. Conf. Environ. Elect. Eng. (EEEIC)*, Rome, Italy. Piscataway, NJ, USA: IEEE, 2015, pp. 1175–1180.

[11] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257

[12] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[13] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, 1973.

[14] S. Makonin and F. Popowich, "Nonintrusive load monitoring (NILM) performance evaluation," *Energy Effic.*, vol. 8, no. 4, pp. 809–814, 2015.

[15] L. Pereira and N. Nunes, "Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—A review," *Wiley Interdiscip. Rev.*, vol. 8, no. 6, 2018, Art. no. e1265.

[16] A. Rodriguez-Silva, "Filtering in non-intrusive load monitoring," Master's thesis, Appl. Sci.: Sch. Eng. Sci., Burnaby, BC, Canada, 2021.

[17] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[18] S. Makonin, Z. Wang, and C. Tumpach, "RAE: The rainforest automation energy dataset for smart grid meter data analysis," *Data*, vol. 3, no. 1, p. 8, 2018.

[19] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 83, pp. 1–5, 2016.

[20] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd, "A rewriting system for convex optimization problems," *J. Control Decis.*, vol. 5, no. 1, pp. 42–60, 2018.

[21] R. Jones, A. Rodriguez-Silva, and S. Makonin, "Increasing the accuracy and speed of universal non-intrusive load monitoring (UNILM) using a novel real-time steady-state block filter," in *Proc. IEEE Power Energy Soc. Innovative Smart Grid Technol. Conf. (ISGT)*, Piscataway, NJ, USA: IEEE, 2020, pp. 1–5.