VI-based Appliance classification using aggregated power consumption data

Leen De Baets, Tom Dhaene, Dirk Deschrijver, and Chris Develder
Department of Information Technology,
Ghent University - imec,
Ghent, Belgium
leen.debaets@ugent.be

Mario Berges
Civil & Environmental Engineering
Carnegie Mellon University,
Pittsburgh, PA 15213-3890

***Abstract*—Non-intrusive load monitoring detects active appliances in a household (and their power consumption) from measuring the aggregated power at just one point in that household. Our previous works focused on classifying a single appliance, assuming that the voltage and current trace could be isolated from an aggregated signal by considering the difference in current before and after the event. In this paper, we show that this assumption holds and that it is a viable approach in practice. We experimentally validate this for two classification methods we proposed earlier: (1) random forests using elliptical Fourier descriptors of the appliances' VI trajectories and (2) convolutional neural networks using the appliances' VI images. We benchmark these approaches on the aggregated data from the 2018 version of PLAID. We obtain, respectively for each of these classifiers, a maximal F_{macro}-measure of 85.31% and 87.95 %. We also show that using submetered data for training does not improve the performance.**

***Index Terms*—Non-intrusive load monitoring, appliance classification**

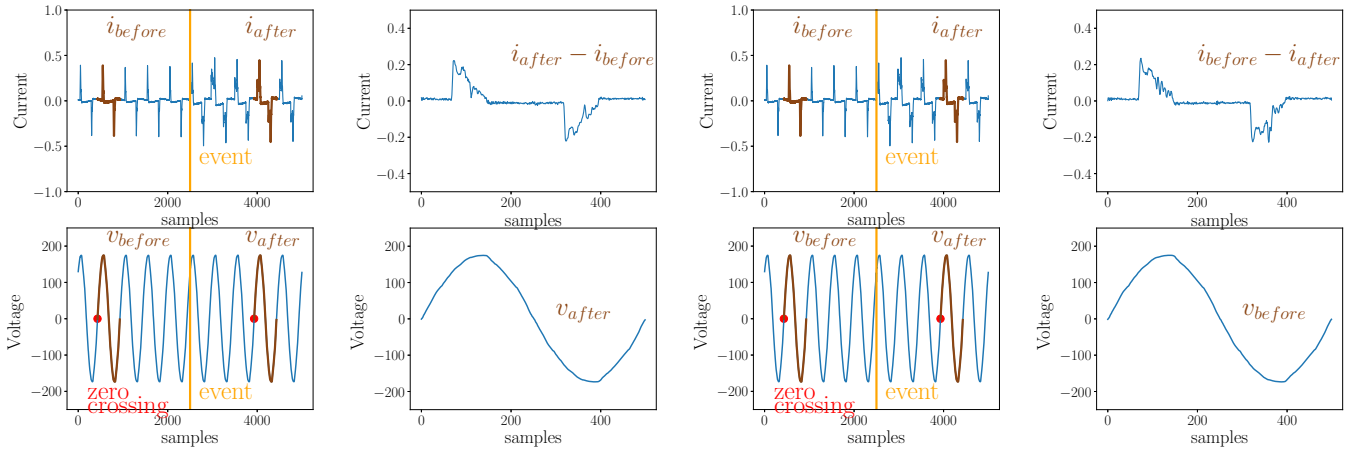
I. Introduction

A basic but crucial step towards increased energy efficiency and savings in residential settings, is to have an accurate view of energy consumption. To monitor residential energy consumption cost-effectively, i.e., without relying on per-device monitoring equipment, non-intrusive load monitoring (NILM) provides an elegant solution. It identifies the per-appliance energy consumption by first measuring the aggregated energy trace at a single, centralized point in the home and then disaggregating this power consumption for individual devices using machine learning techniques. Quite often, two required steps are event detection and appliance classification.

Classifying active appliances for NILM is mostly done by extracting features from the monitored data and training a machine learning classifier. These features are often extracted once it is detected that a device is switched on/off [1]. The type of extracted features heavily depends on the sampling rate of the measurements. When using low frequency data ($\leqslant 1$ Hz), the most common features are the power levels and the on/off durations [2]. A drawback of this approach is that only energy-intensive appliances can be detected. This can be alleviated by performing higher frequency measurements at the cost of an increased data storage rate and more complex data analytics, i.e., the voltage and current signals sampled at a frequency higher than 1 Hz are measured. From these signals, features like the harmonics [3] and other frequency components [4] from the steady-state and transient behavior can be calculated. More recently, the possibility to consider voltage-current (VI) trajectories has also been considered [5]–[7]. Once the features are extracted, they can be fed into different classification methods, like support vector machines (SVM) [8], decision trees [9], or nearest neighbors [10]. In order to distinguish appliances based on their VI trajectories, the voltage and current signals need to be sampled at a relatively high frequency.

In our previous work, the problem of classifying appliances based on their VI trajectories is addressed as an image recognition problem. A first work is based on detecting contours [11]. It represents the trajectory as a pixelated image and describes a classical method for image recognition that: (1) finds the contours, (2) calculates the elliptic Fourier descriptors of the contours, and (3) trains machine learning methods using these elliptical Fourier descriptors. A second work performs image recognition using convolutional neural networks (CNNs) [12]. CNNs are often used for classification tasks in computer vision, due to their excellent discriminative power in classifying images [13]. It is shown that a CNN approach can also be valuable in a NILM context to discriminate active appliances based on the weighted pixelated VI image.

Ideally, to test the methods proposed in our previous papers, a dataset having high frequency aggregated and high frequency sub-metered v and i signals should have been used. However, when these methods were developed, no existing public dataset included both. For this reason, both the 2014 version of PLAID [14] and WHITED [15] were considered as datasets to benchmark the methods as they both contain high frequency sub-metered data. This research on appliance classification was a first step towards a more realistic NILM setting starting from the aggregated power measurements. It was a very meaningful step, as typically appliances are turned on/off one at a time, and the single appliance current (and thus VI trajectory) can be extracted from the aggregated



(a) (b) (c) (d)

Fig. 1: The aggregated current and voltage when an appliance is (a) activated and (c) deactivated, together with the current and voltage of the appliance causing the event (b) and (d).

measurements by considering the difference in current before and after the event. In this paper, we experimentally validate this idea by applying the methods on the high frequency aggregated data of PLAID, which will be published soon. (Note, that also the concurrent work [16] confirmed that single appliance current extracted from aggregated measurements can be successfully used for NILM).

Section II explains first how the current and voltage signal of a single appliance can be extracted from the aggregated data, second it restates how the current and voltage signal are transformed into a pixelated or weighted pixelated image, and at last, it briefly discusses the image recognition methods. For a complete discussion, we refer the reader to the original papers [11], [12]. The evaluation setup is described in Section III. The results of the two earlier published methods on aggregated data are presented in Section IV. Furthermore, we investigate if a better performance is obtained when training uses submetered data instead of aggregated data. Section V concludes this paper.

II. Methodology

This section briefly discusses the methods presented in [11] and [12]. Both methods for appliance classification cast the problem as an image recognition problem. Thus, the VI trajectory of an appliance needs to be transformed into a pixelated or weighted pixelated image, which are respectively taken as input of both methods. First, we describe how the current and voltage signal of a single appliance can be extracted from the aggregated current and voltage signals. Then, the preprocessing to obtain images from the current and voltage signal is discussed. After that, the respective methods are explained.

A. Obtaining submetered current and voltage signal

In order to obtain the pixelated or weighted pixelated VI images of individual appliances from aggregated data, the current and voltage before and after all events are selected. These events can be present in the dataset as labels, or one can detect them using a robust event detection method [17]. The current and voltage before the event (i_{before} and v_{before}) are respectively one current and voltage cycle happening one second before the event. These two cycles are aligned at a zero crossing of the voltage. The extraction of the current and voltage after the event (i_{after} and v_{after}) is performed in the same way for the cycles occurring one second after the event. If the event is caused by the activation of an appliance (the maximum of i_{after} being higher than the maximum of i_{before}) and if only one appliance is activated, then the current i and voltage v of the activated appliance is obtained by:

$$i = i_{\text{after}} - i_{\text{before}} \tag{1}$$

$$v = v_{\text{after}} \tag{2}$$

If the event is caused by the deactivation of an appliance (the maximum of i_{before} being higher than the maximum of i_{after}) and if only one appliance is deactivated, then the current i and voltage v of the deactivated appliance is obtained by:

$$i = i_{\text{before}} - i_{\text{after}} \tag{3}$$

$$v = v_{\text{before}} \tag{4}$$

Figure 1 gives an example. From the obtained per-appliance/submetered i and v signals, the pixelated or weighted pixelated VI image is created.

B. Obtaining VI image from current and voltage

The VI trajectory of an appliance is obtained by first plotting the voltage against the current for a defined time period when the appliance is turned on and in steady state. The VI trajectory is then converted into a VI image ($n \times n$ matrix) by meshing the VI trajectory. If a pixelated VI image is created, each cell of the mesh is assigned a binary value that denotes whether or not it is traversed by the trajectory. If a weighted pixelated VI

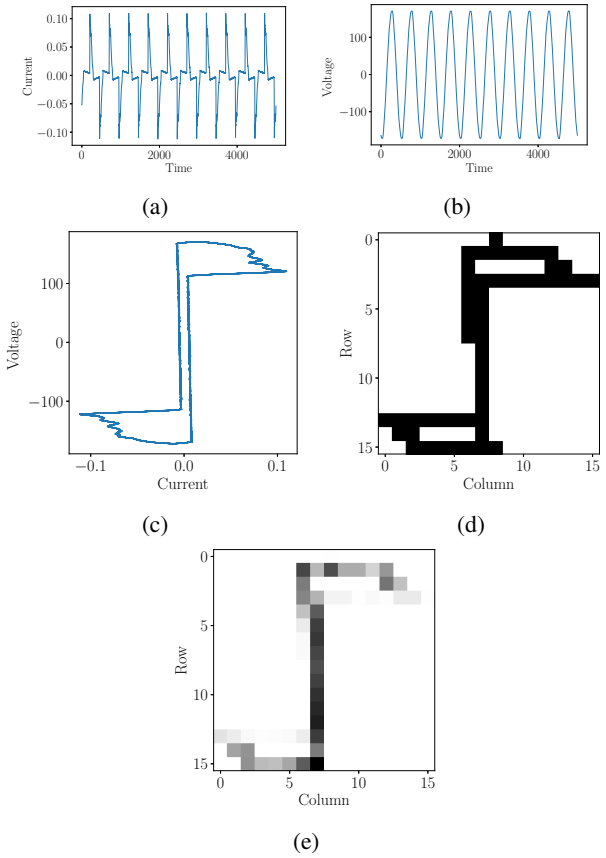


Fig. 2: (a) The voltage and (b) current of a CFL and the transformation from (c) the continuous VI trajectory into (d) the pixelated and (c) weighted pixelated VI image for $n = 15$.

image is created, each cell of the mesh is assigned a value that denotes the number of times it is traversed by the trajectory. This is shown in Figure 2.

C. Elliptical Fourier Descriptors

In [11], the image classification problem is rephrased as an object recognition problem. The contours of an object are identified from the image, characterized by elliptic Fourier descriptors and then classified with a label. In this context, object recognition is used to recognize the contours of a VI trajectory in the pixelated image, and to describe them using elliptical Fourier descriptors. A random forest classifier uses these descriptors to classify the objects.

The contour of an object in an image is a closed curve that forms the boundary of that object. Figure 3 shows the detected contours of the VI trajectory of a compact fluorescent lamp. This example has three contours. To avoid that the trajectory touches the border, extra pixel rows and columns are added to the sides. (Otherwise this would result in two separate outside contours instead of one.) Only one contour can be used to classify the appliances because not all appliances have the same amount of contours, while this is required for the use of machine learning methods. For that reason, the outer counter is chosen, since it is a closed curve that resembles the shape

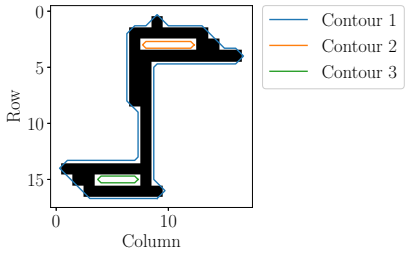


Fig. 3: The identification of the VI trajectory contours of a CFL. Only the outer contour is used for object classification.

of a smoothed VI trajectory. In contrast to the original VI trajectory, all points on the contour are separated uniformly, such that the Euclidean distance between neighbouring points on the contour is the same.

Once the contour is identified, elliptical Fourier descriptors (EFD) are used to characterize the corresponding appliance. EFDs define the contour as the sum of a certain number of ellipses (e) required to mimic the shape, and each ellipse is defined by four parameters (two each for the x- and y-axis). The first ellipse describes the overall shape, location, size, and rotational orientation of the contour. Additionally, more ellipses can be included to capture more detailed information about the contour's complexity. Figure 4 shows the reconstruction of the contour when using up to $e = 4$ harmonics. The approximated contour better resembles the original contour when more harmonics are included. The reader is referred to [18], [19] for mathematical details.

The object recognition results in a vector of size $4 \cdot e$. This vector can be used as input for classification algorithms. As our previous work [11] shows that the random forest obtains the best performance, this is the only classifier that will be used to classify the descriptors. A random forest (RF) is an ensemble technique that classifies the data using a collection of decision trees. Each decision tree is trained on a subset of the dataset that has the same size as the original training set, but samples are drawn with replacement. At each node of the decision tree, a feature is selected and the tree is traversed downward (either following left/right branch) by comparing its value to a threshold. Given a new sample, the output of each decision tree is averaged to obtain the final prediction.

D. Convolutional Neural Networks

Instead of converting the VI trajectory into a pixelated image, it can also be converted into a weighted pixelated image. A CNN can then be applied in order to classify the images. CNNs are a type of neural network (NN) that are often used in computer vision. To create a CNN from a NN, convolutional layers are added. The main difference between a convolutional and fully connected layer is that each node in a convolutional layer is connected to a small region of the input matrix exploiting local correlation, making them highly suitable for classifying images [13]. After a convolutional layer, it is common to implement a pooling

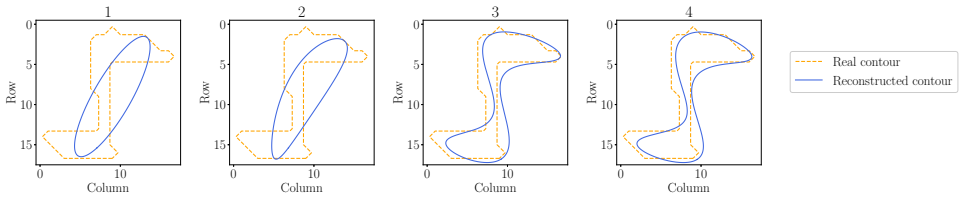


Fig. 4: The original (orange) contour of the VI trajectory of a CFL together with the approximated (blue) contour of the VI trajectory with increasing number of coefficients.

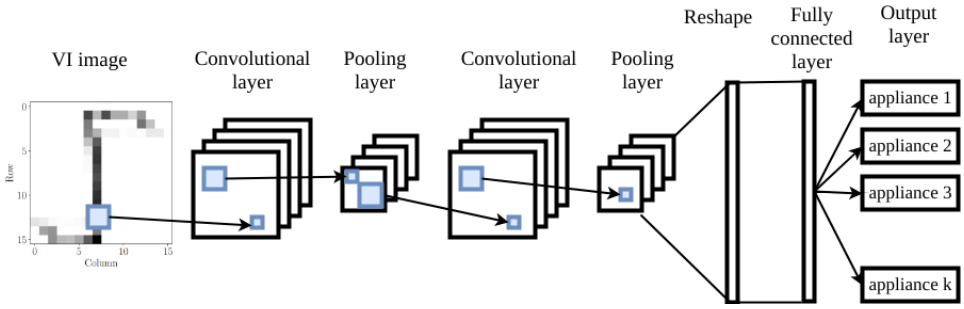


Fig. 5: The architecture of the implemented CNN taking as input the VI image.

layer to downsample the convolved matrix. This reduces the spatial size of the representation and the amount of parameters, and hence also manages overfitting. This downsampling is achieved by sliding a $d \times d$ window over the input (here, with $d = 2$) and each time outputting the largest element of the window.

The CNN implemented in this chapter has the following structure, see Figure 5: it takes as input the weighted pixelated VI image (a $n \times n$ matrix), and has the following hidden layers: a convolutional layer with f filters of size 5, a pooling layer, another convolutional layer with f filters of size 5, another pooling layer, a fully connected layer with n^2 nodes and an output layer with k nodes. The number of filters f is set to 50. The number of output nodes k is determined by the number of different appliances present in the dataset (i.e., the number of classes). An analysis of alternative parameter settings for f showcased no significant changes in the results. Since the class labels are categorical, the cost function of the implemented CNN is defined as the cross-entropy function [20].

III. Evaluation Setup

This section first describes the data on which the proposed methods is benchmarked. After that the used evaluation metric and the research questions are stated.

A. Data

The high frequency aggregated data in the Plug-Load Appliance Identification Dataset (PLAID) is measured at 30 kHz at one location and contains 1478 measurements (activations or deactivations) for 12 different appliances. Additionally, the 12 different appliances are submetered, each 10 times leading to 130 events (the soldering iron leads to two start-up events). In this dataset, the activations and deactivations (events) are labelled making it easy to calculate the current and voltage signal of the appliance causing the event. This data is publicly available.[1] For this paper, the aggregated data is obtained from the files with id ranging from 1 to 324 (included), and the submetered data from the files with id ranging from 1794 to 1925 (included).

It is important to note that although the results presented in [11] and [12] are also obtained using data from PLAID, they cannot be compared with the results obtained in this paper, as a different part of the dataset is used. Here, only the appliances having both submetered and aggregated are used. Whereas for the previous works, the appliances only had submetered data. Thus none of the appliances tested in the previous work, are tested here, making comparison useless.

B. Evaluation metrics

As proposed in [21], the F-measure is used to evaluate the classification performance, which is calculated for each appliance type separately:

$$F_i = 2 \cdot \frac{precision_i \cdot recall_i}{precision_i + recall_i}, \quad \forall i \in [1, \ldots, a] \tag{5}$$

$$precision_i = \frac{TP_i}{TP_i + FP_i} \tag{6}$$

$$recall_i = \frac{TP_i}{TP_i + FN_i} \tag{7}$$

where TP_i, TN_i, FP_i, and FN_i are respectively the true positives, true negatives, false positives, and false negatives for appliance type i. The number of different appliance types is a. The F-measure for a perfect classifier is 1, whereas a random classifier yields an F-measure of 0.5. This measure provides information about the confusion between instances.

[1] www.plaidplug.com

Its magnitude is mainly determined by the number of correctly labeled samples, but tells us nothing about the instances that are correctly labeled with a 0 (the true negatives). In other words, the precision and recall only focus on the positive class [22]. In the end, the average over all the appliance types' F-measure is taken, leading to the so-called macro-average.

$$F_{\text{macro}} = \frac{1}{a}\sum_{i=1}^{a} F_i \tag{8}$$

where a is the total number of different appliance types. Furthermore, the confusion matrix is plotted showing the correct predictions (the diagonal) and the types of incorrect predictions (the rows represent the predicted class and the columns the real class). This matrix gives a clear view on which appliances are confused with each other. The F-measure can be calculated from the confusion matrix.

C. Research Objectives

Objective 1: For each method, does the classification of the appliance types works with these extracted features? We investigate this by calculating the F-measure for the two described classification methods. Additionally, is submetered data necessary for training the algorithms or is aggregated data sufficient? We investigate this by comparing the performance of the algorithms when trained respectively using submetered or aggregated data. To obtain the performance of the first case, the classification algorithms are trained using the submetered data and tested using the aggregated data. To obtain the performance of the second case, the classification algorithms are trained and tested using respectively $3/4$th and $1/4$th of the data. This is repeated 4 times and each fold is created by sampling without replacement (also known as 4-fold cross validation). As a result, each sample of the aggregated data belongs once to the test data. If we store the prediction of each test sample, we are capable to calculate the F_{macro}-measure which can be compared to the F_{macro}-measure calculated in the previous case.

Objective 2: For each method, which parameters lead to the best performance? We investigate this by altering the parameters image size and number of EFDs, and by comparing the obtained F_{macro}-measures.

Objective 3: Which method, the method using the EFDs or the CNN performs the best? We investagte this by comparing the obtained F_{macro}-measures.

IV. Results

This section reports the results obtained by the method using the EFDs and by the CNN and discusses each research qeustion posed in the previous section.

A. Objective 1

Figure 6 shows the F_{macro} for the random forest that uses the EFDs as input and that is trained on submetered and aggregated data for different image sizes. Figure 7 shows F_{macro} for the CNN when using aggregated or submetered data for training, and when using varying images sizes.

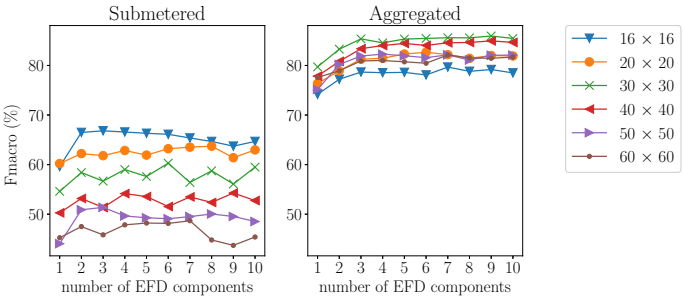


Fig. 6: The F_{macro} of the random forest classifier using an increasing number of EFDs e for different image sizes and when trained on submetered data (left) or aggregated data (right).

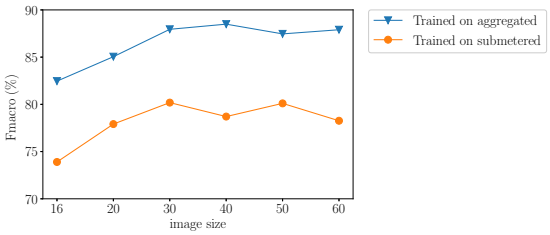


Fig. 7: The F_{macro} for the aggregated data of the 2018 version of PLAID when CNN for $f = 50$ and varying image size n is used. The training is done using submetered or aggregated data.

For both methods, we can state that appliance classification using the extracted submetered voltage and current signals from the aggregated signals works. Additionally for both methods, the results show that training directly on the extracted traces from the aggregated measurements works better than training on the submetered data. This can be explained intuitively by the fact that when training uses aggregated data, the training data contains the same noise (caused by other active appliances) as present in the test data. This is important for practical reasons, as in a household, it is not achievable to submeter all different appliances.

B. Objective 2

In Figure 6, we can see that using three or more EFDs does not significantly impact the accuracy for random forest in terms of the F_{macro}-measure. The pixelated image size is altered between $[16 \times 16, 20 \times 20, 30 \times 30, 40 \times 40, 50 \times 50, 60 \times 60]$. As shown, increasing the image size does not lead to an improvement in the F_{macro} when using EFDs as input for random forest. When trained on submetered data, the EFDs calculated from the contours from the smallest image (16×16) lead to the best F_{macro}, and those from the largest (60×60) to the worst. An intuitive explanation would be that the lower resolution of the images masks the difference between the submetered and aggregated data. When trained on aggregated data, the EFDs calculated from the contours from the image with size 30×30) lead to the best F_{macro}, and those

from the smallest image (16×16) to the worst. We conclude that once a certain resolution is obtained, adding information by increasing the resolution is not useful and leads to a lower performance. The best $F_{\text{macro}}(= 85.31\%)$ is obtained when the random forest is trained on aggregated data and the 3 EFDs are calculated from images of the size 30×30.

In Figure 7, we see that using image sizes larger than 30×30 does not considerably improve the F_{macro}-measure, just like was the case when using EFDs as input. The best $F_{\text{macro}}(= 88.0\%)$ is obtained when the CNN is trained on aggregated data and images of the size 30×30 are used.

Additionally, we also plotted the F-measure per appliance and the confusion matrix for each method when trained using submetered and aggregated data. Figure 8 shows the F-measure per appliance and the confusion matrix for the random forest using as input 3 EFDs extracted from images with size 30×30, and when using submetered and aggregated data for training. When using submetered data for training (Figure 8 (a)), the water kettle and coffeemaker are confused with each other (both resistive heaters). Additionally, some other confusion exists: the CFL is confused with the laptop charger (both non-linear loads) and the AC with the soldering iron. When training uses aggregated data (Figure 8 (b)), a lot of confusion is resolved. Now only the water kettle and the coffeemaker are confused with each other, and the CFL with the laptop charger.

Figure 9 shows the F-measure per appliance and the confusion matrix when using submetered and aggregated data for training, and when the image size of 30×30 is used. When using aggregated data for training (Figure 9b), only the the water kettle and the coffeemaker are confused with each other (both resistive heaters) in respectively 45.3% and 31.9% of the samples. When using submetered data (Figure 9a)), 46.6% of the coffeemaker samples are confused with the water kettle and 13.3% the other way around. Additionally, also the ILB and AC are confused sometimes with the coffeemaker (respectively 15.7% and 9.7%). Further research is necessary to explain why there is an asymmetry in the confusion and why the confusion is reduced when using aggregated data for training compared to using submetered data.

C. Objective 3

When using submetered data for training, the F_{macro} of the CNN (80.4%) is significantly higher than the one obtained by the method based on EFDs (72.5%). This difference is caused by the fact that the CNN is better in classifying the AC and there is less confusion between the water kettle and the coffee maker. When using aggregated data for training, the F_{macro} of the CNN (88.0%) is slightly higher than the one obtained by the method based on EFDs (85.3%). Both the CNN and the method based on EFDs, confuse the water kettle and coffee maker with each other, but the CNN is better in classifying the CFL. This difference in performance is also visible in the previously published work when training and testing was performed on submetered data: the F_{macro} is 66.2% when using 3 EFDs as input for a random forest [11], and 77.6% when using CNN [12]. Again, the CNN outperforms the method using the EFDs. Note that these last two results can not be compared in terms of absolute performance measures with the results mentioned in this paper, as submetered data is used for training and testing.

V. Conclusion

In this paper, we validate that the single appliance current and voltage can be extracted from the aggregated measurements by considering the difference in current before and after the event, assuming that only one appliance is turned on/off one at a time. We tested appliance classification on such submetered signals extracted from aggregated measurements and evaluated two classification methods: (1) the random forest using elliptical Fourier descriptors of the appliances' VI trajectory [11] and (2) the CNN using the appliances' VI images [12], on the aggregated data in PLAID. An F_{macro}-measure of 85.3% and 88.0% are obtained respectively by the two methods, validating that appliance classification using the extracted single appliance current and voltage works reasonably well.

An F_{macro}-measure of 72.5% and 80.4% is obtained respectively by the two methods when submetered data is used for training instead of aggregated data. Using aggregated data for training leads thus to a better performance, indicating that the gathering of submetered data is unnecessary.

In addition for both methods, it was also found that increasing the image size above 30×30 does not lead to better performance. A similar conclusion is found when the number of EDFs exceeds three.

When comparing the two methods, it is found that the CNN performs better than the method using the EFDs.

References

[1] L. De Baets, J. Ruyssinck, C. Develder, T. Dhaene, and D. Deschrijver, "On the bayesian optimization and robustness of event detection methods in nilm," *Energy and Buildings*, vol. 145, pp. 57–66, 2017.

[2] N. Henao, K. Agbossou, S. Kelouwani, Y. Dubé, and M. Fournier, "Approach in nonintrusive type i load monitoring using subtractive clustering," *IEEE Transactions on Smart Grid*, 2015.

[3] W. Wichakool, Z. Remscrim, U. A. Orji, and S. B. Leeb, "Smart metering of variable power loads," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 189–198, 2015.

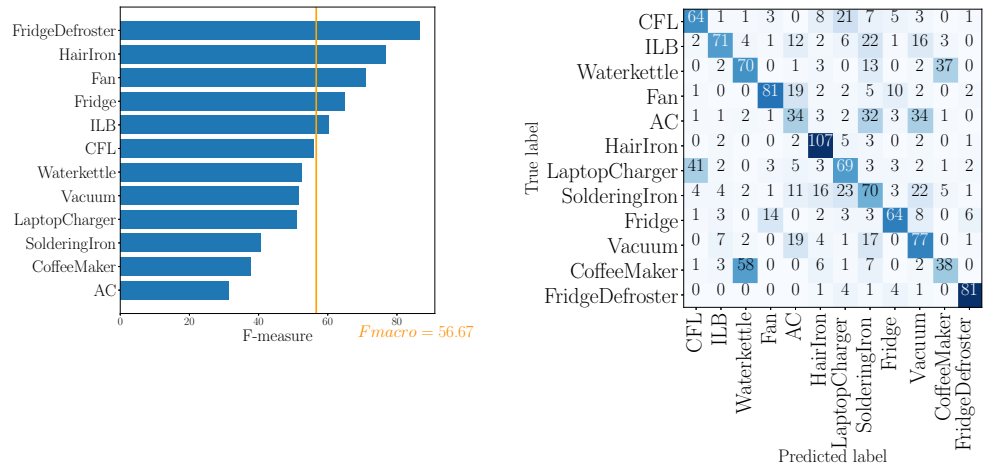
[4] H.-H. Chang, K.-L. Chen, Y.-P. Tsai, and W.-J. Lee, "A new measurement method for power signatures of nonintrusive demand monitoring and load identification," *IEEE Transactions on Industry Applications*, vol. 48, no. 2, pp. 764–771, 2012.

[5] T. Hassan, F. Javed, and N. Arshad, "An empirical investigation of vi trajectory based load signatures for non-intrusive load monitoring," *IEEE Transactions on Smart Grid*, vol. 5, no. 2, pp. 870–878, 2014.

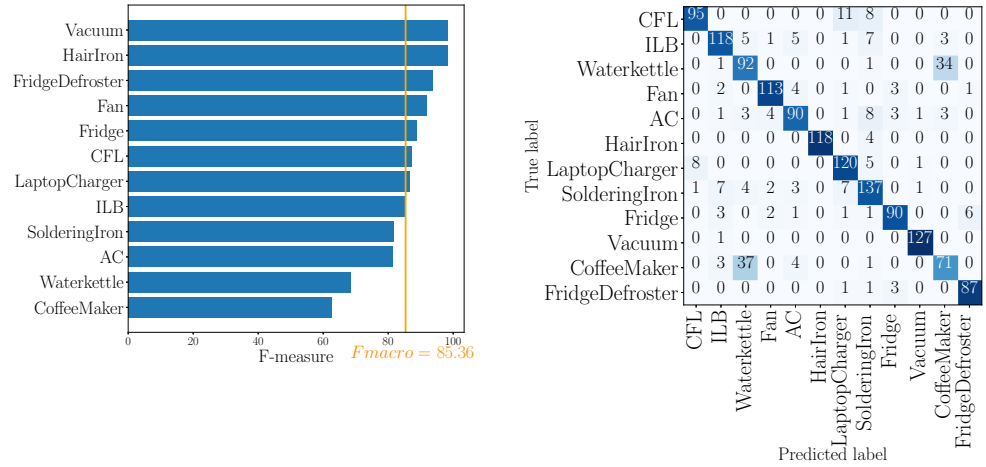
[6] L. Du, D. He, R. G. Harley, and T. G. Habetler, "Electric load classification by binary voltage–current trajectory mapping," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 358–365, 2016.

[7] J. Gao, E. C. Kara, S. Giri, and M. Bergés, "A feasibility study of automated plug-load identification from high-frequency measurements," in *Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on*. IEEE, 2015, pp. 220–224.

[8] H. Altrabalsi, V. Stankovic, J. Liao, and L. Stankovic, "Low-complexity energy disaggregation using appliance load modelling," *AIMS Energy*, vol. 4, no. 1, pp. 884–905, 2016.



(a) submetered



(b) aggregated

Fig. 8: The F-measure per appliance and confusion matrix for the aggregated data in the 2018 version of PLAID with $e = 3$ EFD components, the image size is 30×30 and the random forest is trained using (a) submetered, and (b) aggregated data. AC = air conditioning, CFL = compact fluorescent lamp, ILB = incandescent light bulb

[9] M. Nguyen, S. Alshareef, A. Gilani, and W. G. Morsi, "A novel feature extraction and classification algorithm based on power components using single-point monitoring for nilm," in *Electrical and Computer Engineering (CCECE), 2015 IEEE 28th Canadian Conference on*. IEEE, 2015, pp. 37–40.

[10] K. Basu, A. Hably, V. Debusschere, S. Bacha, G. J. Driven, and A. Ovalle, "A comparative study of low sampling non intrusive load dis-aggregation," in *Industrial Electronics Society, IECON 2016-42nd Annual Conference of the IEEE*. IEEE, 2016, pp. 5137–5142.

[11] L. De Baets, C. Develder, D. Deschrijver, and T. Dhaene, "Automated classification of appliances using elliptical fourier descriptors," in *In the proceedings of the IEEE International Conference on Smart Grid Communications*, 2017, pp. 1–6.

[12] L. De Baets, J. Ruyssinck, C. Develder, T. Dhaene, and D. Deschrijver, "Appliance classification using vi trajectories and convolutional neural networks," *Energy and Buildings*, vol. 158, pp. 32–36, 2018.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[14] J. Gao, S. Giri, E. C. Kara, and M. Bergés, "Plaid: a public dataset of high-resoultion electrical appliance measurements for load identification research: demo abstract," in *proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*. ACM, 2014, pp. 198–199.

[15] M. Kahl, A. U. Haq, T. Kriechbaumer, and H.-A. Jacobsen, "Whited-a worldwide household and industry transient energy data set," in *3rd International Workshop on Non-Intrusive Load Monitoring*, 2016.

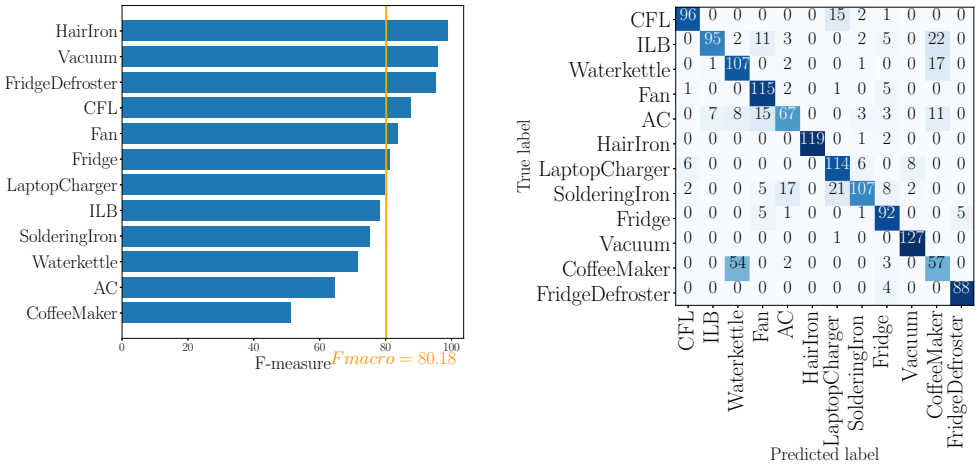
[16] A. L. Wang, B. X. Chen, C. G. Wang, and D. Hua, "Non-intrusive load monitoring algorithm based on features of v–i trajectory," *Electric Power Systems Research*, vol. 157, pp. 134–144, 2018.

[17] L. De Baets, J. Ruyssinck, C. Develder, T. Dhaene, and D. Deschrijver, "On the bayesian optimization and robustness of event detection methods in nilm," *Energy and Buildings*, vol. 145, pp. 57–66, 2017.

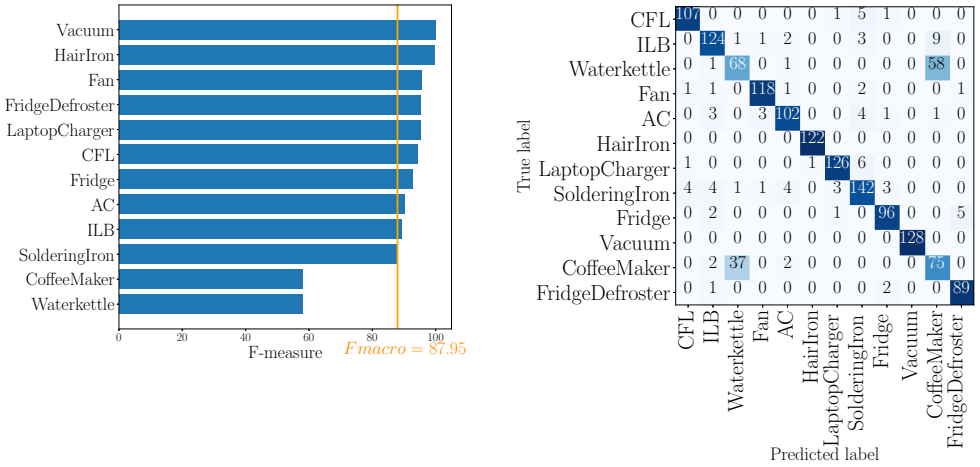
[18] F. P. Kuhl and C. R. Giardina, "Elliptic fourier features of a closed contour," *Computer graphics and image processing*, vol. 18, no. 3, pp. 236–258, 1982.

[19] J. Caple, J. Byrd, and C. N. Stephan, "Elliptical fourier analysis: fundamentals, applications, and value for forensic anthropology," *International Journal of Legal Medicine*, pp. 1–16, 2017.

[20] "Nielsen, m., 'neural networks and deep learning'," http://neuralnetworksanddeeplearning.com/, accessed: 2017-06-17.



(a) submetered



(b) aggregated

Fig. 9: The F-measure per appliance and confusion matrix for the aggregated data in the 2018 version of PLAID when the CNN for $n = 30$ and $f = 50$ is used, and is trained using (a) submetered, and (b) aggregated data . The number of samples per appliance type is mentioned between the brackets. AC = air conditioning, CFL = compact fluorescent lamp, ILB = incandescent light bulb

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

[22] N. Japkowicz, *Assessment metrics for imbalanced learning*. Wiley Online Library, 2013.