Transparent AI: Explainability of deep learning based load disaggregation

David Murray
Lina Stankovic
Vladimir Stankovic
david.murray@strath.ac.uk
lina.stankovic@strath.ac.uk
vladimir.stankovic@strath.ac.uk
Dept. of Electronic and Electrical Engineering, University of Strathclyde
Glasgow, Glasgow, UK

ABSTRACT

The paper focuses on explaining the outputs of deep-learning based non-intrusive load monitoring (NILM). Explainability of NILM networks is needed for a range of stakeholders: (i) technology developers to understand why a model is under/over predicting energy usage, missing appliances or false positives, (ii) businesses offering energy advice based on NILM as part of a broader energy home management recommender system, and (iii) end-users who need to understand the outcomes of the NILM inference.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Machine learning**; • **General and reference** → **Reliability**; **Verification**; **Validation**.

KEYWORDS

datasets, neural networks, reliability, validation

ACM Reference Format:
David Murray, Lina Stankovic, and Vladimir Stankovic. 2021. Transparent AI: Explainability of deep learning based load disaggregation. In *The 1st ACM SIGEnergy Workshop of Fair, Accountable, Transparent, and Ethical AI for Smart Environments and Energy Systems (FATEsys '21), November 17–18, 2021, Coimbra, Portugal.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3486611.3492410

1 INTRODUCTION

Recent years have seen significant research in defining, at a high level, how inference models can be designed to be explainable to end-users. Explainablity leads to trust in data-driven AI systems ensuring that complex machine learning (ML) models underpinning these systems are understandable to the end user and decisions or recommendations are transparent. Despite a large number of publications from different disciplines, including many tutorial and feature articles [1, 3, 5, 6, 8, 11], most explainable AI implementations focus on technology designs, e.g., for the purpose of removing the bugs in the code or improving the models, while other potential users of the technology are neglected. Furthermore, the bulk of the literature tends to focus on explainability of image processing and natural language processing, while raw time-series sensor signals processing, e.g. energy measurements, is almost non-existent.

In this paper, we focus on the explainability of deep-learning based non-intrusive load monitoring (NILM) [7] of electrical smart meter data that provides feedback to householders or building managers about energy consumption of individual appliances [14], [4]. NILM has been researched for over 40 years and has been embedded in energy management recommender systems by providing appliance-specific energy outcomes. We demonstrate, using the popular sequence2point deep learning NILM architecture [4], how heatmaps can be used to explain NILM outputs.

We refer to a model being interpretable if it is possible to mathematically predict its output, and interpretability as the ability to support user comprehension of the model decision making process and predictions. We refer to explainability as the ability to explain the underlying model and its reasoning with accurate and user comprehensible explanations. Explainablity is essential when assessing effects of biases in the data, degrees of fairness and other ethical implications of research, since the methods need to be replicated and tested in a new environment (using different, potentially biased dataset), and its decisions need to be mathematically tractable [5].

There have been only few attempts to explain time-series data models [12], where it is challenging to relate decisions to raw signals, and hence explanations have mainly been related to quantifying the importance of each feature; however, with deep learning models that take raw signals and integrate the feature engineering steps, this is often impossible. Similarly, there have been no attempts to explain NILM specifically besides [10], which targeted tech developers by visualising trained network weights at the early layers.

NILM or load disaggregation refers to estimating individual appliance load contributing to the metered household aggregate energy consumption without submetering. Numerous approaches for NILM have been used previously, and a review can be found in [14]. To illustrate explainability tools for NILM, we use a sequence2point network of [4] that is a widely used for benchmarking deep learning based NILM work. We note that the approaches presented apply to other architectures also. The architecture of [4, 13], is a novel

FATEsys '21, November 17–18, 2021, Coimbra, Portugal

ACM ISBN 978-1-4503-9114-6/21/11…$15.00
https://doi.org/10.1145/3486611.3492410

sequence2point approach for NILM, based on convolutional neural networks (CNN) that extracts meaningful latent features with appliance transfer learning and cross-domain transfer learning. A sliding window of the aggregate is mapped into a single middle value point of the targeted appliance, this way predicting the appliance consumption value for each sample in time. [4] presents results showing excellent performance of the proposed approach for a range of appliances on three datasets.

2 EXPLAINABILITY OF NILM

We illustrate how NILM deep learning models can be interpreted and explained using the washing machine, considered a challenging appliance to disaggregate, due to multiple consumption states, with power values similar to numerous other appliances. To explain how the model makes decisions, we occlude (null values) part of the raw input and slide the occlusion window across the data. For each window position, we estimate the model's singular point output. This is used to generate a heat map as shown in Figure 1 (bottom). This methodology makes no changes to the network's internals unlike methods such as Attention Networks which require the addition new methods/layers.

In Figure 1 (top) we show an example of the input aggregate signal, target signal (washing machine), and predicted (non-occluded) signal. The horizontal axis shows sample number and vertical, consumed power. In this case we show a true positive result on the ECO dataset [2], the model being trained on the REFIT dataset [9]. The occlusion window blocks 50 consecutive samples and is stepped across the input window from index 0 to 549. This is then used to generate the heat map in Figure 1 (bottom). For a fixed sample point (horizontal axis), vertically, the values in the map correspond to the network output for different starting positions of the occluding window (from 0 to 549).

The heat map should be read diagonally to keep the occluded window stationary as we move along the sample axis, due to the network targeting the centre point of the window. The heat map is aligned with the top plot along the x-axis to better indicate where the target point is, with the colour representing consumption estimation at a given point. The horizontal bar across the centre of the heat map represents where the centre point of the input sequence window is occluded, e.g., samples 249 to 299. When this occurs the network struggles to predict, as the input at the target sample is null. Importantly, this leaves the model vulnerable should errors occur around the window centre, and makes a case for explaining how data is filled/processed to end users.

The ellipses in the heat map represent three key features. Ellipse 1 shows what we consider the main feature of the washing machine, the heating element turn on, around sample 705700; when this is occluded the predicted load drops significantly. Inversely, when occluding the area just before the heating element turns on, we see the highest load values, higher than the non-occluded input. This appears as multiple true and false positives, and highlights a weakness with multiple high power appliances being used one after another. After the heating and spin cycles, there is a period where draining occurs with the occasional spin. Shown in Ellipse 2, the lowest load estimate occurs when this feature is fully occluded highlighting its importance. The third ellipse highlights the spin cycle. In this case, the occluded section, which is not fully captured by ellipse 3, shows distinct band of blue which begins shortly after the washing machine heating element has turned off, but not immediately, indicating that the model is taking time into account.

In Figure 2, we show an example that illustrates the limitation of the trained model to handle overlapping activations, where a number of appliances are used simultaneously. In this example, another appliance usage occurs at the end of the washing machine heating cycle (sample 575500). When occluding this area, we expect a truer estimate of the previous load. Indeed, the result is a much higher network prediction, shown by ellipse 1 in the heat map. Ellipse 2 shows the importance of the draining cycle in order to detect washing machine uses. If this segment is even partially occluded, the estimated consumption drops to near 0. Additionally, there is another appliance which overlaps this feature (Sample 575700, end of area 2); this, along with the second overlapping appliance, helps us to explain why the network likely missed this activation. Finally, Ellipse 3 corresponds to the detection of the spin cycle that in Figure 2 has a number of unknown appliance uses causing network confusion. Ellipse 3 in the heat map plot shows a false positive occurring if the end of the second appliance is occluded, e.g., the network thinks that a second heating cycle is in progress. This mistake (the network does not detect this activation unless an occlusion window is applied) shows the trained networks inability to cope with overlapping appliances.

Heat maps provide a model agnostic way to visually interpret time series results, working with both sequence-to-sequence and sequence-to-point style networks. Depending on the complexity of the input and target signals, the number of learned features will become apparent when occluding the input signal. Depending on the size of the target signal, the size of the occluding signal can highlight features, and shrinking the occluding window can show what the model considers the most impactful features. This methodology could also be used to discover adversarial examples in which outputs are vastly influenced by a single point in the input window. The visualisation of stacked plots, allows those not familiar with the field to understand features which are considered important, and could help to create a "stress testing" set of tricky examples to be used for trained model benchmarking.

Figures 1 and 2 also report the MAE, SAE and NDE performance measures for these particular uses as defined in [4]. The values of all three metrics are lower for Figure 1 compared to those of Figure 2, which is expected since the former is a TP sample whilst the latter is a FN. However, these measures do not clearly have a range of values that are comparable. While interpretability explains the decisions made by the model, it is often not understandable by the end user, e.g., a householder trying to understand their appliance consumption estimate in regards to their electricity bill. Thus we provide explainability by attempting to explain the measures in relation to the top plot in Figure 2. Clearly, the predicted of the consumption of the appliance is under-estimated compared to the actual. The MAE is the only metric that captures this wide difference in reconstructing the signal, compared to the other two metrics but does not necessarily explain the underestimation, which would not provide a realistic consumption to the end-user trying to understand the real consumption of their appliance. Over the entire dataset, however, MAE is less explainable as the MAE value becomes lower

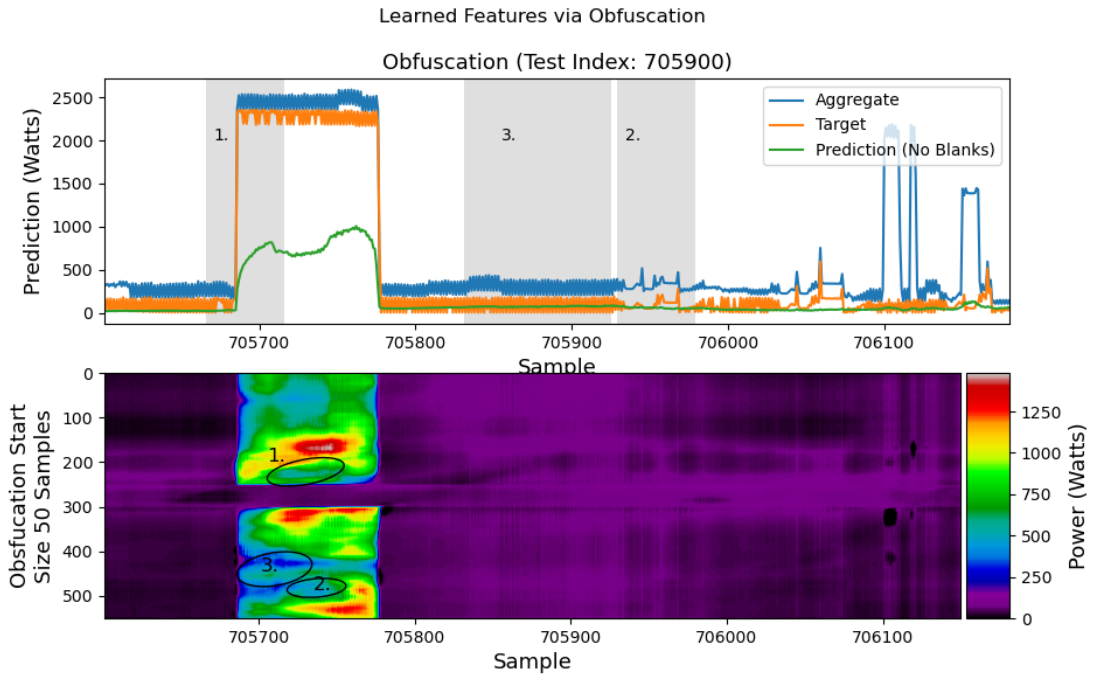

Figure 1: The heat map generated for the Washing Machine in the ECO dataset, house 1. The model is trained using the entire REFIT dataset. The obtained performance measures for this sample are: MAE:292.73, SAE:0.62, NDE:0.45.

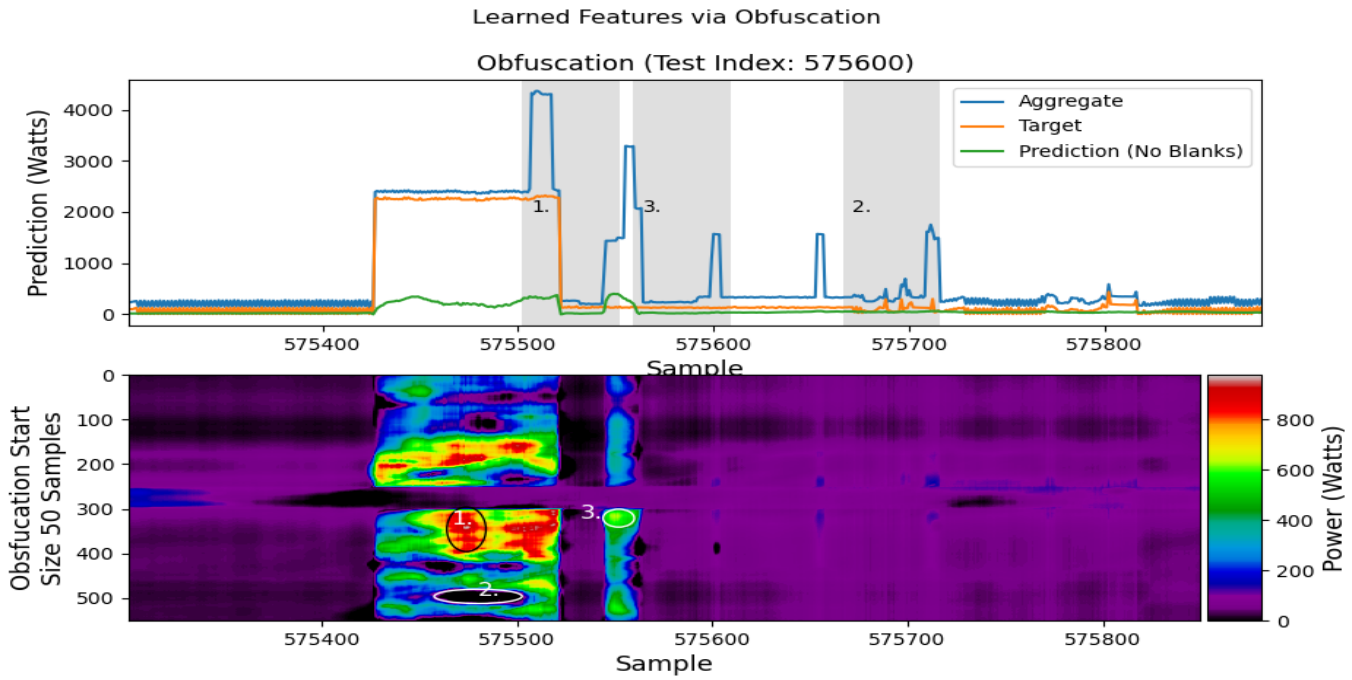

Figure 2: The heat map generated for the Washing Machine in the ECO dataset, house 1. The model is trained using the entire REFIT dataset. The obtained measures for this sample are: MAE:383.18, SAE:0.83, NDE:0.79.

due to the fact that appliances spend a significant period turned off. Therefore, we wish to highlight that these metrics, commonly used for evaluating the performance of deep learning approaches in the recent NILM literature, are not truly explainable since they are not necessarily intuitive.

3 CONCLUSION

In summary, we propose heat maps to help explain performance metrics and reconstructed appliance signatures. For the purposes of explainability, metrics for validation and evaluation of performance need to be application-specific for comparison with other methods and understandable by the end-user.

However, we recognise that heat maps may be difficult to explain to the end-user, who has little to no domain knowledge. Therefore, further studies with end users and non-AI specialist building systems experts will be needed, e.g. through interactive workshops to evaluate different levels of explainability. Furthermore, a separate study to analyse input data to address bias, in terms of patterns of use for example, is needed especially when testing on unseen datasets.

ACKNOWLEDGMENTS

This work was supported in part by the European Commission under Horizon2020 MSCA-RISE-2016 Grant Agreement No 734331 (SENSIBLE project), and MSCA-ITN-2020 Grant Agreement No 955422 (GECKO project).

REFERENCES

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58, October 2019 (6 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[2] Christian Beckel, Wilhelm Kleiminger, Romano Cicchetti, Thorsten Staake, and Silvia Santini. 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 - BuildSys '14*. ACM Press, New York, New York, USA, 80–89. https://doi.org/10.1145/2674061.2674064

[3] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8 (7 2019), 832. https://doi.org/10.3390/electronics8080832

[4] Michele D'Incecco, Stefano Squartini, and Mingjun Zhong. 2020. Transfer Learning for Non-Intrusive Load Monitoring. *IEEE Transactions on Smart Grid* 11, 2 (3 2020), 1419–1429. https://doi.org/10.1109/TSG.2019.2938068

[5] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* (2 2017). http://arxiv.org/abs/1702.08608

[6] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 80–89. https://doi.org/10.1109/DSAA.2018.00018

[7] George William Hart. 1992. Nonintrusive appliance load monitoring. *Proc. IEEE* 80, 12 (1992), 1870–1891. https://doi.org/10.1109/5.192069

[8] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2018. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *arXiv* 1, 1 (2018).

[9] David Murray, Lina Stankovic, and Vladimir Stankovic. 2017. An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study. *Scientific Data* 4 (2017). https://doi.org/10.1038/sdata.2016.122

[10] David Murray, Lina Stankovic, and Vladimir Stankovic. 2020. Explainable NILM Networks. In *Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring* (Virtual Event, Japan) *(NILM'20)*. Association for Computing Machinery, New York, NY, USA, 64–69. https://doi.org/10.1145/3427771.3427855

[11] Wojciech Samek, Thomas Wiegand, and Klaus Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv* (2017).

[12] Udo Schlegel, Hiba Arnout, Mennatallah El-Assady, Daniela Oelke, and Daniel A. Keim. 2019. Towards a rigorous evaluation of XAI methods on time series. In *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*. https://doi.org/10.1109/ICCVW.2019.00516

[13] Chaoyun Zhang, Mingjun Zhong, Zongzuo Wang, Nigel Goddard, and Charles Sutton. 2018. Sequence-to-point learning with neural networks for non-intrusive load monitoring. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*.

[14] Bochao Zhao, Lina Stankovic, and Vladimir Stankovic. 2016. On a Training-Less Solution for Non-Intrusive Appliance Load Monitoring Using Graph Signal Processing. *IEEE Access* 4 (2016), 1784–1799. https://doi.org/10.1109/ACCESS.2016.2557460