Toward Transparent Load Disaggregation—A Framework for Quantitative Evaluation of Explainability Using Explainable AI

Djordje Batic, *Student Member, IEEE*, Vladimir Stankovic, *Senior Member, IEEE*, and Lina Stankovic, *Senior Member, IEEE*

***Abstract*—Load Disaggregation, or Non-intrusive Load Monitoring (NILM), refers to the process of estimating energy consumption of individual domestic appliances from aggregated household consumption. Recently, Deep Learning (DL) approaches have seen increased adoption in NILM community. However, DL NILM models are often treated as black-box algorithms, which introduces algorithmic transparency and explainability concerns, hindering wider adoption. Recent works have investigated explainability of DL NILM, however they are limited to computationally expensive methods or simple classification problems. In this work, we present a methodology for explainability of regression-based DL NILM with visual explanations, using explainable AI (XAI). Two explainability levels are provided. Sequence-level explanations highlight important features of predicted time-series sequence of interest, while point-level explanations enable visualising explanations at a point in time. To facilitate wider adoption of XAI, we define desirable properties of NILM explanations - faithfulness, robustness and effective complexity. Addressing the limitation of existing XAI NILM approaches that don't assess the quality of explanations, desirable properties of explanations are used for quantitative evaluation of explainability. We show that proposed framework enables better understanding of NILM outputs and helps improve design by providing a visualization strategy and rigorous evaluation of quality of XAI methods, leading to transparency of outcomes.**

***Index Terms*—Deep neural networks, explainable AI (XAI), non-intrusive load monitoring, load disaggregation.**

I. Introduction

LOAD disaggregation or Non-intrusive load monitoring (NILM) is the process of algorithmically inferring the energy consumption of individual electrical appliances from the aggregate metered power consumption of a residential building [1]. There is a growing interest in NILM deployment due to growing energy costs, energy efficiency initiatives and national smart metering roll-outs. Deep learning based implementations for NILM have grown sharply over the past few years with very good performance demonstrated via domain-agnostic accuracy metrics, such as the popular Mean Absolute Error, across a wide range of real-world datasets [2]. However, using accuracy metrics as a standalone determinant for selection of an AI technology is inadequate for wider consumer adoption, as put forth in [3] and [4]. The latter recommends that, in order to ensure Trustworthy AI, robustness, fairness, transparency, and privacy need to be addressed. Indeed, the European Commission has recently published seven principles of Trustworthy AI [5], which include transparency as one of the key elements of trustworthy AI systems. Transparency is closely linked to traceability of the datasets, as well as explainability of the technical processes of the AI system and the related AI decisions, and finally communication of AI system's level of accuracy and limitations to the end-users and system developers.

For AI-based NILM, the majority of work has focused on addressing technical robustness in the form of accuracy, reliability and reproducibility across different datasets [2], [6], [7] and data transparency through the use of public, peer-reviewed and well-documented datasets [8], [9], with limited research in the area of privacy protection [10], [11], [12] and technical explainability [13], [14], [15]. The majority of deep learning-based NILM approaches are designed as "black-box" systems due to their inherent algorithmic complexity and absence of explainability. Since the underlying mechanics resulting in NILM predictions are not interpretable or explainable, deep learning (DL) based NILM cannot be fully trusted, which somewhat hinders wider deployment of NILM systems [3]. As the adoption of smart home devices and energy management systems continues to grow, the necessity to ensure these technologies are both transparent and understandable to consumers grows concurrently. By developing and evaluating XAI methods for NILM, the research community can contribute to design of AI solutions that adhere to consumer standards such as the EU's vision of ethical and responsible AI [5] and foster consumer trust in these emerging technologies, empowering users to make informed decisions about their energy consumption. Furthermore, understanding the produced outputs can help improve the design, provide a better overview of the model accuracy, and facilitate better understanding of failure scenarios. Thus, the role of explainability is to ensure a transparent inference process of the AI system by providing decisions that are understood and traceable. As a result, algorithmic transparency facilitated by explainability has been

Manuscript received 9 May 2023; revised 23 May 2023; accepted 28 July 2023. Date of publication 1 August 2023; date of current version 26 April 2024. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie under Grant 955422. *(Corresponding author: Djordje Batic.)*

The authors are with the Electronic and Electrical Engineering Department, University of Strathclyde, G1 1XW Glasgow, U.K. (e-mail: djordje.batic@strath.ac.uk).

Digital Object Identifier 10.1109/TCE.2023.3300530

identified as a paramount challenge in the present landscape of NILM research [3].

The wider problem of explainability of DL models has recently gained traction, leading to the emergence of the field of Explainable AI (XAI). Recent literature [16], [17], [18], [19], [20], [21], [22] suggest that XAI can facilitate trust by providing algorithmic transparency, support assessment of levels of bias, and improve the overall understanding of the inner workings of deep learning models. The majority of XAI work, predominantly tackling computer vision tasks, primarily centers around the integration and development of techniques that analyse the outputs of the model and visualise the importance of the input features. Such work frequently illustrates that explainability can enhance the understanding of the model and foster trust in the AI systems [21]. However, many existing XAI techniques can lead to unstable explanations in real-world scenarios due to limited, qualitative evaluation [23], [24], [25], [26]. Addressing such issues is particularly important for systems that can reveal personal information, such as temporal appliance patterns of use, generated by NILM. XAI approaches for NILM are still in their infancy, with limited literature available [13], [14], [15], [27]. As XAI-based solutions for NILM continue to grow, it is of vital importance to properly evaluate their explainability components. This assessment can serve as a way to assert that the used explainability techniques are truly able to be deployed in the real-world scenarios and help with understanding of model outputs. Therefore, XAI system design that incorporates robust qualitative and quantitative evaluation procedures for explainability techniques used in the real-world environment is of crucial importance for the successful adoption in NILM.

The main contributions of this work are summarized as follows:

- A new multi-temporal XAI visualisation technique for regression-based DL NILM, taking into account the need for different levels of visualisation granularity.
- Definition of three core properties for evaluation of explainable NILM system: faithfulness, robustness, and complexity, that quantify the quality of XAI NILM visualisations with respect to the ability to identify important features of the signal, deal with noisy inputs, and be human understandable, respectively.
- Demonstration that the proposed approach can provide visualisations and quantify well the quality of XAI NILM systems using two public, well documented datasets and five XAI approaches.

The rest of the paper is organised as follows. A detailed literature review is presented in Section II to position our contributions with respect to the state-of-the-art. The proposed explainability framework is described in Section III followed by the experimental results and key findings in Section IV, before we conclude in Section V.

II. Literature Review

A. NILM Problem Formulation

Let $y = (y_1, y_2, \ldots, y_T)$ be a sequence of aggregated power consumption from M appliances, captured at time $t = \{1, 2, \ldots, T\}$. Given a measurement of aggregate power $y(t)$, the goal of a NILM algorithm is to determine the individual power contribution $x_i(t)$ of appliance $i \in \{1, 2, \ldots, M\}$, such that the aggregate can be represented as:

$$y(t) = \sum_{i=1}^{M} x_i(t) + n(t), \tag{1}$$

where $n(t)$ denotes noise caused by the measuring equipment and unknown appliances contributing to the aggregate. NILM can be treated as a regression problem if the task is to directly infer $x_i(t)$ based on the aggregate signal $y(t)$. On the other hand, it can be regarded as a binary classification problem if the task is to determine the on/off state of appliance i at time t, based on the aggregate signal $y(t)$. Formulated in this manner, NILM can be solved in a range of supervised and unsupervised approaches and eliminates the need for appliance submetering, leading to a reduction in costs [28], while still enabling a diverse set of applications such as energy usage feedback [29], anomaly detection [30], and load shifting [31].

In terms of algorithmic approaches, CNNs are the most widely used architectures in the latest NILM literature according to the recent review of [32]. Reference [33] use an event-driven CNN for load disaggregation of residential appliances, while [34] employ a CNN to perform unsupervised domain adaptation. However, of all CNN-based works, sequence-to-point (seq2point) learning represents one of the most cited approaches [35]. Given an input sequence of aggregate signal, the seq2point algorithm predicts the midpoint of the output (i.e., appliance) signal instead of the whole sequence. This approach has shown to be a better approximation of the target distribution compared to previous approaches and consequently provides advantageous predictive performance [36].

B. XAI for NILM

Algorithmic transparency in AI systems is often characterized by the clarity of decision making processes implemented by AI algorithms [37]. From an engineering perspective, works focusing on algorithmic transparency fall in the category of XAI [17], [18], [19], [20]. Despite the increased need for algorithmic transparency and extensive research in XAI, the majority of current AI systems lack the ability to provide clear explanations of how the AI model generated an output.

XAI for decision-tree based NILM was demonstrated in [27], whereby Partial Dependance Plots and Individual Conditional Expectation were used to explain the predictions of the NILM multi-class classifier by highlighting feature importance for individual appliances. However, the remaining XAI approaches for NILM focus on explainability of DL-based NILM. The first XAI approach for NILM, proposed in [13], focuses on occlusion sensitivity, and provides visual insight into important features of the prediction of a regression-based NILM AI algorithm. Explanations are generated by first occluding parts (i.e., setting to zero value) of the signal with a sliding window across the time series. Then, for each window position, model output at a single point is calculated. The information about the resulting outputs is used to determine the importance of individual time steps and create the explanation heatmap. The sliding window is slid over the whole sequence that largely contains

power levels under 500 W. However, the proposed approach suffers from issues of computational complexity due to the nature of the sliding window approach. Furthermore, occlusions that are set as zero values are rarely observed in practice due to the baseload presence, making the presented methodology exposed to potential out-of-distribution inference scenarios, which can result in unstable predictions. A comparison between a gradient-based technique, GradCAM [19], and an occlusion-sensitivity approach for visualizing the important features of a NILM classifier is examined in [14]. However, [14] uses a less challenging NILM approach based on multi-class CNN to only determine the existence of an appliance in the input time-series, without detecting on/off states, using a single dataset. [15] propose a learning mechanism that utilizes XAI techniques for training of DL NILM models using the paradigm of knowledge distillation [38]. Authors explored the transfer of knowledge in the Teacher-Student scenario, identified the main inconsistency in the transfer of explainable knowledge, and proposed a modification to the knowledge distillation loss function to improve the model performance by minimizing the inconsistencies between the Teacher and Student explanations.

Despite the recent advancements, there are several gaps in the literature with respect to XAI for NILM that warrant further exploration. One notable gap is the scalability and computational complexity of current XAI visualization methods for regression-based NILM. For instance, existing techniques for regression-based NILM, such as occlusion sensitivity [13], are computationally heavy, limiting their feasibility for large-scale datasets or real-time applications. Another critical gap is the lack of standardized evaluation metrics for assessing the quality and usefulness of XAI techniques in the context of NILM. All existing work in NILM relies solely on qualitative evaluation of XAI methods. However, developing a comprehensive set of benchmarks for domain-relevant aspects of explainability would enable better comparisons between different XAI approaches and facilitate the identification of best practices. Even though the aforementioned works can be considered as an entry-point towards explainability in NILM systems, to the best of our knowledge, there is still no work in NILM literature that evaluates XAI methods in a quantitative manner. This suggests a lack of rigorous evaluation of the quality of generated explanations, which is a requirement to ensure trust in the explanation outputs of XAI-based AI systems [22], [24], [25], [26], [39], [40], [41], [42], [43].

C. Explainability Methods

In this study, our focus lies on post-hoc XAI methods that aim to explain outputs of a trained DL model by assigning attribution or relevance values to each input feature. Given an input to a DL model and a target concept, attribution-based XAI aims to map the importance of each input feature to the target concept. The target concept is either a class of interest in classification tasks or an output value in regression-based problems. We refrain from using feature-based approaches such as LIME [44] and SHAP [45], due to their instability and computational complexity [24], [46]. Instead, we examine five popular families of methods that best exemplify the variety of algorithmic approaches contained in the field of XAI, namely GradCAM [19], LRP [20], SmoothGrad [18], and Integrated Gradients [17].

1) Gradient-Weighted Class Activation Mapping (GradCAM): GradCAM is an XAI technique used to create an explanation for a prediction of a target concept (e.g., a class or a signal sequence) by computing its gradient w.r.t the final convolutional layer of a CNN network [19]. In order to generate an explanation map $h^c \in \mathbb{R}^{W \times H}$ of width W and height H for a target concept c, the gradient of the output for the target concept y^c w.r.t the kth feature map activations A^k of the last convolutional layer is computed, i.e., $\frac{\partial y^c}{\partial A^k}$. Next, a global average pooling operation is applied over the height and width (indexed by i and j, respectively) on the computed gradients, to obtain neuron importance weights [19]:

$$\omega_k^c = \frac{1}{W \times H} \sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k}. \tag{2}$$

The generated weights represent the importance of feature map k for the target concept c. In order to compute the explanation map h^c, weighted combination of feature map activations, followed by ReLU function, is performed [19]:

$$h^c = ReLU\left(\sum_k \omega_k^c A^k\right). \tag{3}$$

2) Improved Gradient-Weighted Class Activation Mapping (GradCAM++): GradCAM++ is an extension of the original GradCAM method that has been shown to provide better visual explanations for CNN models [47]. The main improvement lies in the calculation of the neuron importance weights, which now considers not only the first-order partial derivatives but also the second-order partial derivatives to capture higher-order interactions among feature maps. The updated neuron importance weights for the target concept c in GradCAM++ are computed as follows [47]:

$$\omega_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \cdot ReLU\left(\frac{\partial y^c}{\partial A^k ij}\right), \tag{4}$$

such that the partial derivatives w.r.t. A_{ij}^k are as follows:

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 y^c}{(\partial A^k i,j)^2}}{2\frac{\partial^2 y^c}{(\partial A^k i,j)^2} \sum_a \sum_b A_{ab}^k \frac{\partial^3 y^c}{(\partial A^k i,j)^3}}. \tag{5}$$

Where the final explanation map h^c is computed as in Eq. (3). Comparing with Eq. (2) and (3), GradCAM++ reduces to GradCAM if $\forall i, j$, $\alpha_{ij}^{kc} = \frac{1}{W \times H}$. GradCAM++ has been shown to produce higher quality and more precise visual explanations compared to the GradCAM method, allowing for better interpretation of CNN models [47].

3) Integrated Gradients (IG): IG [17] aims to generate an explanation for a prediction of a target concept, via counterfactual reasoning. Absence of a cause for a certain prediction informs the generation of the importance features by creating a single baseline value used to compare the outcomes. Generally,

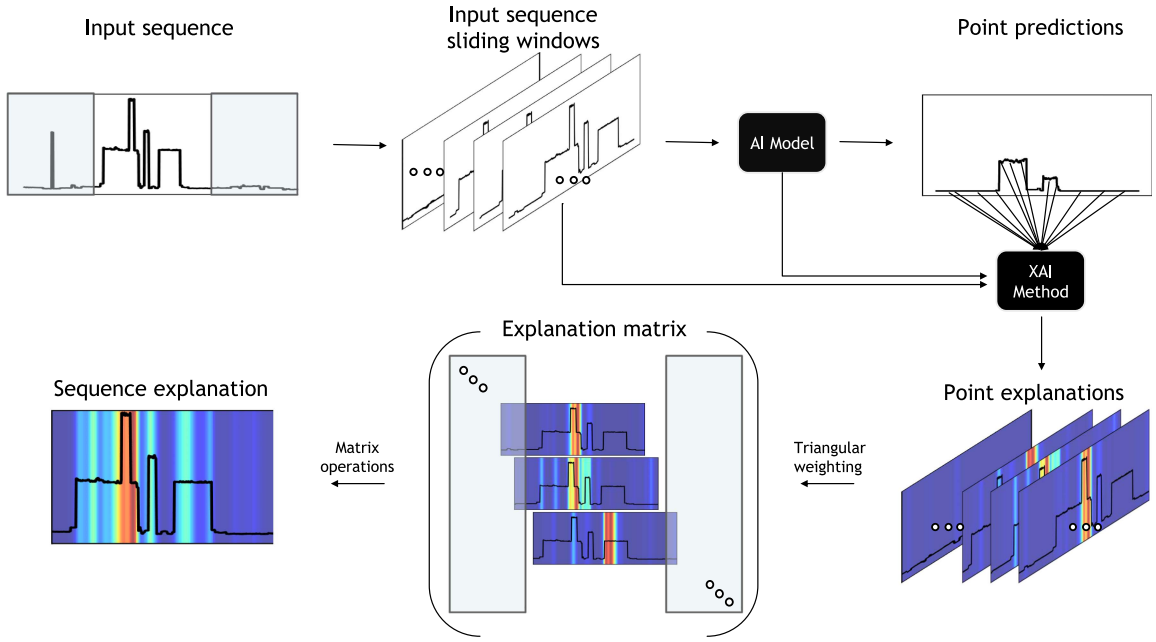

Fig. 1. Visual outline of the proposed approach showcasing the mechanism for visualization of importance at two levels of specificity, leading to point-level and sequence-level explanations for an input sequence of interest.

the baseline is modeled as a space where predictions are neutral. In computer vision, this would typically be a black image, while in the case of time-series data this can be represented as absence of the signal. Formally, explanation map $h^c \in \mathbb{R}^{W\times H}$ of width W and height H for a target concept c, considering input $x \in \mathbb{R}^{W\times H}$ and baseline value $\hat{x} \in \mathbb{R}^{W\times H}$, is created by constructing a set of interpolations along the i^{th} dimension between x and $\hat{x}$ [17]:

$$h_i^c = (x_i - \hat{x}_i) \odot \int_{\alpha=0}^{1} \left[\nabla f(\hat{x} + \alpha \cdot (x - \hat{x}))\right]_i d\alpha, \tag{6}$$

which can be approximated by a finite summation of gradients at small intervals along the path from $\hat{x}$ to input x [17]:

$$h_i^c \approx (x_i - \hat{x}_i) \odot \frac{1}{N} \sum_{k=1}^{N} \nabla f\left(\hat{x} + \frac{k}{N} \cdot (x - \hat{x})\right). \tag{7}$$

where N is the number of steps in the Riemann approximation of the integral.

4) SmoothGrad: Driven by the premise that instability of gradient-based explanation maps can be corrected by smoothing of a gradient with a Gaussian kernel over a large number of local perturbations, SmoothGrad calculates an average of gradients w.r.t N alterations of the input, by adding a small amount of random noise [18]. Given that the method computes gradients with respect to input x, i.e., $m^c(x) = \frac{\partial y^c}{\partial x}$, explanation map h^c is calculated as [18]:

$$h^c = \frac{1}{N} \sum_{i=1}^{N} m^c\left(x + \mathcal{N}\left(0, \sigma^2\right)\right), \tag{8}$$

This technique aims to reduce the visual noise, and can be combined with other methods to create smoother heatmaps.

5) Layer-Wise Relevance Propagation (LRP): LRP [20] computes the explanation heatmaps by using the layered structure of the neural network to produce relevance scores in an iterative manner. Given two consecutive layers, j and k, propagation of a relevance score R from a higher to a lower neuron is achieved by means of purposely designed local propagation rules. For example, given an input activation a_j and weight w_{jk} connecting neuron j to neuron k, LRP-ϵ rule is defined as [20]:

$$R_j = \sum_k \frac{a_j \omega_{jk}}{\epsilon + \sum_{0,j} a_j \omega_{jk}} R_k. \tag{9}$$

ϵ is a regularization term - high ϵ values help stabilize the relevance scores when contribution to the activation of neuron k is weak or unclear, leading to less noisy explanation maps.

D. Evaluation of Explainability

Traditionally, the quality of attribution-based explainability has been evaluated by qualitative, subjective assessment. This constitutes determining subjective levels of satisfaction with the usefulness of explanation, which is evaluated by a developer or end-user of an AI system. However, driven by the need for more rigorous and objective evaluation strategies, recent advancements in the field have focused on the development of quantitative metrics for assessing the degree of the quality and trust of XAI methods.

A key challenge in evaluating XAI methods is the lack of ground truth. Given that the information about how a model generates a prediction can rarely be known a priori, efforts in evaluating the quality of explanations tend to approach the problem indirectly. Concretely, with the end goal of measuring if explanations correspond to the predictive performance of the model, [25], [39], [40] propose various methods for measuring *faithfulness*, based on the notion that removing or obscuring

important input features should have a significant negative effect on performance, or confidence of the prediction. The degree of faithfulness is quantified by measuring the difference between the probability scores of a classifier predicting on perturbed and original input, where more faithful methods lead to larger differences in scores. Faithfulness has also been referred to as *sensitivity-n* [22], *selectivity* [41], *fidelity* [42].

Unreliability of backpropagation-based XAI methods has long been an issue, as discussed in [22], where concerns over the fact that XAI methods can lead to unstable and unintelligible explanations are discussed. To mitigate the issue, *sanity checks* are proposed [43], comprising a set of techniques geared towards evaluating the trustworthiness of explainability methods by comparing the results of applying them to trained and randomly initialized models. Furthermore, with the aim of addressing the aforementioned issues of unreliability, the notion of the *robustness* of explainability methods has been suggested [39]. Their findings suggest that slight changes in the input, simulating adversarial noise, could lead to dramatic differences in generated explanations, while retaining the same predictions. Driven by the need to formulate the relationship between input data and reliability of XAI methods, [39] evaluate robustness of explanation functions under slight perturbations of the input, and derive measures for determining their ability to deal with small modifications of the input. The notion of robustness has been explored in other works and referred to as *sensitivity* [42], *continuity* [39] and *stability* [41].

The end goal of explanations is to be understandable to humans who are interpreting them. As a result, explanations that deem all of the features as important, even if faithful, have limited utility as their interpretation might be too difficult for a human to understand. As a way of measuring the conciseness of explanations, authors in [25] proposed a measure of *complexity*. The low complexity of generated explanations suggests that they highlight only the most relevant features and that understanding them does not present a difficult task. Complexity has also been presented as *sparseness* in [26].

III. NILM Explainability Framework

The backbone of our proposed XAI framework for NILM is the proposed visualization procedure, illustrated in Fig. 1, that facilitates the generation of human-interpretable explanations of NILM model outputs. Since the desired granularity of explanations can vary, the visualization procedure offers an ability to generate explanations for both sequential-level, as well as point-level predictions. The sequence-level explanations highlight the areas of the signal most responsible for the prediction, while the point-level explanations display the reasoning behind a prediction of a particular point in time. These two layers of explainability can be used interchangeably as they offer varying degrees of specificity. In the visualization procedure, we utilize five distinct XAI techniques to formulate explanations. Subsequently, the created explanations are subjected to a quantitative evaluation of quality. Taking into consideration a diverse set of needs and possible deployment scenarios, the quality of an explanation is defined as alignment with three desirable properties of explanations, specifically: faithfulness, robustness, and low complexity.

A. Visualization via Heatmaps

We demonstrate how to integrate XAI in the popular seq2point DL-NILM implementation of [35] trained for load disaggregation of various appliances, via regression, on two popular datasets: UK-DALE [9] and REFIT [8]. The full procedure is illustrated in Fig. 1. First, to account for the nature of the seq2point algorithm, sliding windows are used to split the input signal into small, overlapping segments, and generate the point output predictions. Then, for a seq2point model with input size δ, for each generated point along the sliding window, a point explanation heatmap of size δ is created via GradCAM, LRP, SmoothGrad, or IG, as per Section II-C. If a step size of 1 is used, and the length of activation window of interest is ω, the total number of generated heatmaps is:

$$N = \omega - \delta + 1. \tag{10}$$

Following this procedure, we observe that a single time step along the activation window ω can receive up to δ importance scores. However, this does not hold for all points in ω, in particular the ones at the edges of the window. For example, two points at the far edge (left and right) of the activation window receive only one computed importance score. To ensure that each point along ω captures δ importance scores, we expand the activation window by $\delta - 1$ on both sides. Thus, we create a window of size:

$$\omega' = \omega + 2 * (\delta - 1). \tag{11}$$

Given that the size of activation window of interest, ω, is larger than the model input size, δ, to map the N resulting heatmaps to a single, sequence-level heatmap of size ω, which corresponds to the activation of interest, we need to transform the results into a new representation. To create a heatmap of size ω, we first generate a zero matrix of size $\omega' \times (N + 2 * (\delta - 1))$. Each generated heatmap is added to the matrix based on its position relative to the activation of interest. For example, the first row of the matrix contains the first heatmap that is followed by zero values, acting as padding, until reaching ω' samples. The first element in the second row is set to zero, followed by the second heatmap, and finally zero values afterward until reaching ω' samples. This procedure is repeated until the last row.

Before populating the matrix, we apply a weight function to mitigate the presence of noise and promote smoothness of heatmaps. Given that the temporal dimension of the middle point of the input corresponds to the output point of prediction, and is highly influential to the prediction, we apply a triangular weight function to the heatmap defined as:

$$\psi(x) = \begin{cases} \frac{x}{m}(p_{max} - p_{min}) + p_{min} & \text{if } 0 \le x \le m \\ \frac{x-m}{m}(p_{min} - p_{max}) + p_{max} & \text{if } m < x \le 2m \end{cases} \tag{12}$$

where m represents the middle point value, and p_{min} and p_{max} are the lowest and highest weight values, respectively. The maximum value p_{max} is placed at the middle point, while the

values drop linearly in both directions when moving away from the middle point, with the lowest value p_{min} at points 0 and $2m$. For the purpose of this work, the weight function holds the maximum value of 1 at the middle point, with the two furthest points holding a weight of 0.8.

To further reduce the noise, we aggregate the results by first sorting the matrix column-wise in descending order, corresponding to the time step in the window of interest, and then creating a vector of size ω' by computing the non-zero mean value of the top 40% of values per each column of the matrix. In the last step, we transform the window to size ω by clipping the generated vector by $\delta - 1$ on both sides. Following this procedure, the importance heatmap of the target window of interest is obtained, containing the cumulative importance for each of the predicted points of the signal.

B. Property of Faithfulness

The proposed faithfulness evaluation strategy quantifies the extent to which explanations attest to the predictive performance of a model. In other words, faithfulness aims to determine if the feature importance scores, generated by the visualization procedure, are indicative of importance w.r.t. prediction. Given that a ground truth explanation can rarely be known, faithfulness is measured indirectly, by observing the impact of a feature removal on the generated prediction. To measure the faithfulness of an XAI-enabled NILM approach, the following steps are taken:

1) Generate a sequence-level feature importance map of an input signal of interest, as in Section III-A.
2) Partition the sequence-level maps into sorted, non-overlapping segments based on the sum of importance scores over a certain period, to determine the most important areas of the input signal.
3) Evaluate the faithfulness of the derived explanations by performing an iterative perturbation of features by changing the input signal values in the segments of interest, starting with the segments of highest relevance. The perturbation of input segment is performed by replacing the power level of the initial signal by the signature of low consuming appliances (e.g., a combination of TV, Lights and Fridge, equaling around 250W). This perturbation ensures that the activation signal is attenuated, while keeping the input data distribution within the space that the model has learned on, as opposed to setting the power level to zero, which would constitute an unfavorable case of an out-of-distribution scenario.
4) To establish whether there is a significant impact on the predictive performance, after each perturbation of features we measure the difference between the performance metrics calculated on predictions of non-perturbed and perturbed signals.
5) To convey the degradation of performance, we consider both classification and regression-based performance metrics. As a way of capturing the classification performance, we convert the regression output to a step function and calculate the F_1 score as:
$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}, \tag{13}$$
where TP stands for True Positives, FP for False Positives, and FN for False Negatives. To quantify the disaggregation performance, we utilize mean absolute error (MAE) between the true (E_i) and predicted ($\hat{E}_i$) consumed energy of the appliance of interest where MAE is calculated as follows:
$$MAE = \frac{1}{T} \cdot \sum_{i=1}^{T} |\hat{E}_i - E_i|. \tag{14}$$
6) After each perturbation step, compute the difference between performance metrics of altered and original input. The faithfulness score is the resulting area under curve (AUC) after a set number of iterations, where more faithful XAI methods correspond to a higher AUC score. The classification faithfulness showcases the difference in F_1 score values, while regression faithfulness depicts the difference in MAE values. Iterative perturbation of features that leads to a sharper increase in the difference between the F_1 and/or MAE scores (and thus higher faithfulness score) suggests that the feature importance scores generated by the XAI method successfully assign scores to the highly relevant input features and are indeed indicative of predictive performance of the model.

C. Property of Robustness

The growing body of literature in deep learning theory [48] suggests that robustness of neural networks is closely related to the value of its local Lipschitz constant. Intuitively, a Lipschitz constant represents the value by which neural network's output is allowed to change relative to its input. It has been used as a hard constraint to enable adversarial robustness, better generalization and training of generative adversarial networks. Moreover, it has been suggested as a technique for evaluating the robustness of explanations [39]. Given a slight modification of input, and consequently negligibly small effect on the prediction, a robust explanation should not differ drastically compared to those created from the unmodified input. We aim to investigate the (in)stability of existing XAI methods w.r.t. slight modifications of household aggregated consumption signal. Given an explanation function $h(\cdot)$ and input aggregate signal x, we expose the signal to zero-mean Gaussian noise with controlled standard deviation σ to create modified input aggregate signal, $\hat{x}$. We define local Lipschitz constant estimate as [39]:
$$\hat{L} = \frac{\|h(x) - h(\hat{x})\|}{\|x - \hat{x}\| + \mu}, \tag{15}$$
where μ represents a small value added for numerical stability ($\mu = 1e^{-6}$). For validity, we repeat this procedure n times and report the averaged robustness score (RS). Methods with low Lipschitz value scores display a characteristic of being stable under the presence of noise and should be favoured. In the context of NILM-like data it is important to note that we assume

bounded input space, i.e., that maximum change in the function value is finite, which can be assumed for NILM signals as the magnitude of the aggregate power signal is bounded.

D. Property of Complexity

One of the core principles of XAI is to provide human understandable explanations. Previous studies in the area of research focusing on applying XAI in the energy sector have reported mixed results when applying XAI tools to real-world energy data [49]. Yet, none of these studies have delved into the evaluation of explainability methods, particularly the complexity of explanations. We argue that this property is one of the most desirable ones, as it quantifies the entropy of the XAI output. If most of the input features are deemed important, it does not provide an adequate level of clarity and lowers the human interpretability of explanation. To measure the conciseness of explanation output, we measure the statistical dispersion of the output map. The output map is first sorted in ascending order, and indices of the sorted values are determined. Finally, the conciseness of explanation is formulated as a Gini index computation [26]:

$$Gini = \frac{\sum_{a=1}^{\omega}(2a-\omega-1)\cdot h_a}{\kappa+\omega\cdot\sum_{a=1}^{\omega}h_a}, \tag{16}$$

where h_a is the a-th point in the sorted XAI output of length of ω, i is the rank of values in the ascending order, and $\kappa = 1e^{-8}$ is a small value added for numerical stability. A *Gini* coefficient takes values in the range of $[0-1]$, with coefficient of 0 expressing equal contribution of all features, and 1 expressing that only one feature contributes to the resulting heatmap.

Evaluation of explainability is in general a two-step process, where at first an explanation result is generated using an XAI method considering the input of the model and the model itself, followed by the measurement of the desirable property of explanation result. In this sense, explanation sparseness points to the dispersion of the distribution of the output of the XAI method (i.e., the complexity of explanation). However, it disregards information about the complexity of the input variable. We argue that this is highly important for systems that include time-varying data, as the presence of noise is a common phenomenon, and the system's ability to deal with it is of particular interest. Consequently, explanation sparseness in the context of NILM does not reflect one of the most common challenges of working with time-series. One of the existing measures that capture the percentage of noise in data sample, noise-aggregate measure (NAR) [50], is defined as:

$$NAR = \frac{\sum_{i=1}^{T}|y(t)-\sum_{i=1}^{N}x_i(t)|}{\sum_{t=1}^{T}y(t)}. \tag{17}$$

We adapt the formula to measure the noisiness of one particular window and appliance i of interest defined as:

$$NAR^{(i)} = \sum_{i=1}^{T}\left|1-\frac{x_i(t)}{y(t)}\right|. \tag{18}$$

We observe that the explanation complexity is often similar for inputs with varying degrees of noise. To establish the relationship between the complexity of an input variable and the complexity of explanation, we introduce an additional term to the explanation complexity that reflects the "noisiness" of the input. Thus, to quantify the complexity of explanation in the context of NILM, we define the "effective complexity" measure as:

$$EC^{(i)} = \frac{Gini}{1-NAR^{(i)}}. \tag{19}$$

IV. Experimental Results: Qualitative and Quantitative Explainability

A. Experimental Setup: Datasets and Model Training

For transparency, we used the most widely used [2] and well documented REFIT [8] and UK-DALE [9] public datasets. These datasets contain real-world active power measurements obtained from residential buildings, exhibiting a realistic spectrum of appliance ownership and usage patterns. To evaluate explainability across appliance activations with different levels of power and activation periods, we focus our attention on popular multi-state and single-state appliances, namely: Washing Machine, Dishwasher, Microwave, and Kettle. The aggregate data were pre-processed using normalization with mean and standard deviation values computed from the training set. All models were trained and evaluated by reproducing the procedure outlined in [35]. Houses were chosen based on the condition that they must contain measurements of all four aforementioned appliances. For UK-DALE, we use houses 1, 3, 4, and 5 for training, while house 2 is used for testing. In the case of REFIT, houses 2, 3, 6, 11, 13, and 15 were used for training, while the test set contains data from house 5.

The explainability dataset is created by randomly sampling 30 days when appliances of interest are running and selecting a window of size ω samples centered around the appliance activation window from each chosen day. Given a dataset with granularity of 8 seconds, ω is determined from the typical operation time of the appliance of interest. For appliances with lengthy duration, i.e., Washing Machine (WM) and Dishwasher (DW), activation length $\omega = 1024$ is chosen, which represents roughly 2 hours and 15 minutes of measurements, in line with the average length of a duty cycle of most WM and DW devices. For the Microwave (MW), activation length $\omega = 80$ was chosen, which corresponds to around 10 minutes. Finally, Kettle (KT) activation length ω is set at 40, corresponding to around 5 minutes. If the total length of the activation length of interest is larger than ω, the first ω data samples are selected.

B. Interpretation of Faithfulness, Robustness and Complexity Scores

Faithfulness is of particular importance to an algorithm designer, as it facilitates understanding of how feature importance scores influence the prediction. Conversely, robustness provides an indication of the change in prediction if the input to the DL model changed slightly (e.g., due to appliance model fluctuations, appliance settings and influence of unknown appliances), which is a crucial indicator of scalability. Finally, complexity reflects the human comprehensibility of the visualization. The relative significance

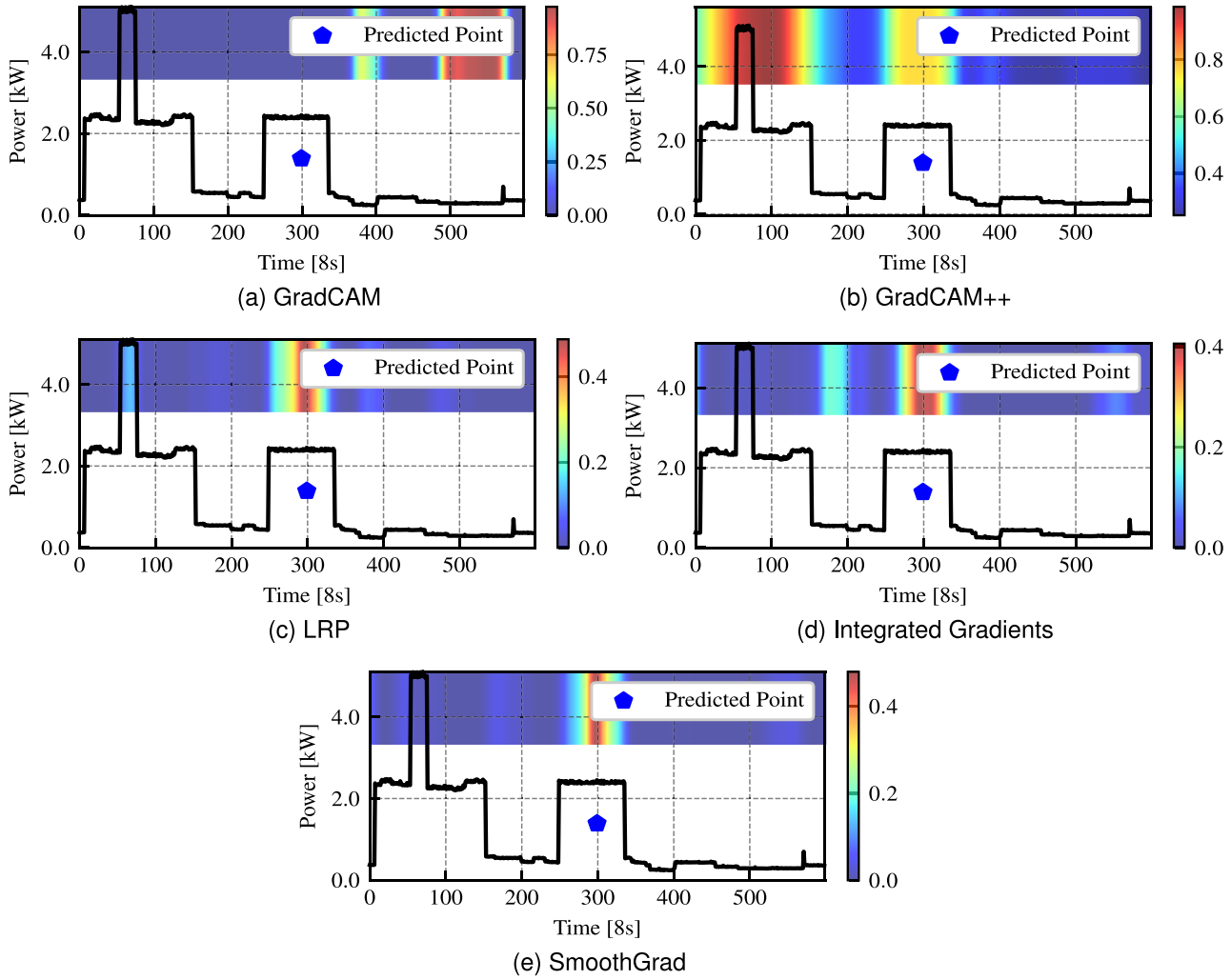

Fig. 2. Explanations generated for positive activation of dishwasher in UK-DALE dataset. We can observe unreliable results from GradCAM, while other methods offer more accurate and concise explanations.

TABLE I
COMPARISON OF EXPLAINABILITY AND PREDICTIVE PERFORMANCE OF SEQ2POINT MODEL FOR UK-DALE DATASET

Appliance	XAI Method	R. Faithfulness	C. Faithfulness	Robustness	Gini	Eff. Complexity
Washing Machine	GradCAM	1413.384	19.800	0.485 ±0.308	0.485	0.833
	GradCAM++	1908.446	17.183	0.602 ±0.161	0.189	0.325
	LRP	**2466.142**	**23.560**	**0.113 ±0.113**	**0.880**	**1.510**
	IG	1888.325	20.253	0.393 ±0.168	0.412	0.708
	SG	1889.292	19.454	0.306 ±0.119	0.500	0.859
Dishwasher	GradCAM	37.942	5.934	1.606 ± 0.734	0.486	0.658
	GradCAM++	2014.717	20.704	0.617 ± 0.216	0.342	0.462
	LRP	3186.500	**26.973**	0.517 ± 0.289	**0.784**	**1.061**
	IG	2375.636	12.329	0.699 ± 0.433	0.592	0.801
	SG	**3262.523**	19.823	**0.459 ± 0.175**	0.662	0.897
Kettle	GradCAM	**1721.840**	**1.699**	0.062 ±0.060	0.421	0.476
	GradCAM++	1653.429	1.667	**0.034 ±0.034**	0.432	0.488
	LRP	1386.882	1.478	0.225 ±0.140	**0.692**	**0.782**
	IG	1235.617	1.205	0.309 ±0.159	0.490	0.554
	SG	516.182	0.394	0.129 ±0.081	0.428	0.484
Microwave	GradCAM	598.240	4.298	0.155 ± 0.150	0.478	0.74
	GradCAM++	**602.853**	**4.456**	**0.055 ± 0.045**	0.490	0.759
	LRP	479.337	3.810	0.127 ± 0.085	**0.798**	**1.236**
	IG	547.137	4.450	0.148 ± 0.085	0.756	1.171
	SG	435.108	3.983	0.128 ± 0.081	0.775	1.200

of each score is determined by the use-case, i.e., which property is most desirable to an algorithm designer, system developer, consumer or technology enthusiast. Explainability scores (see Sections III-B–III-D) obtained for four different appliances are presented in Tables I and II, for the UK-DALE and REFIT datasets, respectively. Regression (R) and

TABLE II
COMPARISON OF EXPLAINABILITY AND PREDICTIVE PERFORMANCE OF SEQ2POINT MODEL FOR REFIT DATASET

Appliance	XAI Method	R. Faithfulness	C. Faithfulness	Robustness	Gini	Eff. Complexity
Washing Machine	GradCAM	517.966	0.454	1.070 ± 0.667	0.405	1.257
	GradCAM++	339.881	0.213	0.532 ± 0.240	0.165	0.514
	LRP	**1794.590**	**4.751**	**0.434 ± 0.357**	**0.661**	**2.052**
	IG	1381.301	2.561	0.847 ± 0.296	0.394	1.224
	SG	1098.127	2.001	0.700 ± 0.301	0.461	1.431
Dishwasher	GradCAM	2773.987	9.538	1.323 ± 1.017	0.539	1.242
	GradCAM++	2934.133	10.385	0.940 ± 0.942	0.276	0.635
	LRP	4312.670	14.862	**0.367 ± 0.235**	**0.683**	**1.572**
	IG	**6530.439**	**26.035**	0.764 ± 0.369	0.577	1.329
	SG	5469.436	17.727	0.804 ± 0.451	0.575	1.324
Kettle	GradCAM	1158.721	2.161	0.188 ± 0.234	0.472	0.671
	GradCAM++	**1325.057**	**2.415**	**0.059 ± 0.049**	0.355	0.503
	LRP	1160.369	2.073	0.205 ± 0.170	**0.608**	**0.862**
	IG	1011.075	1.667	0.197 ± 0.099	0.598	0.849
	SG	910.304	1.637	0.172 ± 0.081	0.562	0.797
Microwave	GradCAM	628.539	3.520	0.296 ± 0.239	0.512	0.754
	GradCAM++	**720.156**	**4.069**	**0.116 ± 0.140**	0.443	0.666
	LRP	672.775	3.712	0.229 ± 0.124	**0.785**	**1.180**
	IG	677.663	3.857	0.272 ± 0.132	0.528	0.794
	SG	634.402	3.363	0.282 ± 0.195	0.482	0.724

Classification (C) scores are calculated as the AUC for MAE and F1 scores, as described in Section III-B. For long duration appliances (WM and DW), we perform 75 perturbation steps, while for MW and KT we perform 10 and 5 steps, respectively. To calculate the sorted, non-overlapping segments of importance (as per Section III-B), for appliances with a long activation period, we choose segments containing 40s of measurements, while other appliances contain 24s of measurement. High faithfulness score indicates that the explainability method is able to correctly identify the important features of the input signal, thus leading to a large drop in prediction accuracy after perturbation. The Robustness score is calculated as mean and standard deviation of $n = 35$ computations of Lipschitz constant estimate, defined in Eq. (15), where μ and σ values of Gaussian noise are 0 and 0.1, respectively. Low robustness score indicates the ability of the explainability method to deal with noise. The Effective complexity is calculated as per Eq. (19). High effective complexity suggests that the explainability method is able to generate explanations that are concise and human understandable.

Tables I and II suggest that LRP-ϵ achieved the most success across the proposed properties that explainable NILM systems based on sequence-to-point learning should satisfy. This can largely be attributed to the ability to deal with gradient noise as the relevance is propagated through the layers of the network. We report a strong relationship between the choice of parameter ϵ and the results in performance metrics, where ϵ value should be guided by the noisiness of the dataset. As the REFIT dataset is known to be significantly noisier than UK-DALE, we set the parameter ϵ to be a large value ($\epsilon = 1$) compared to UK-DALE ($\epsilon = 0.1$). Contrary to previous studies in the energy sector that recommended GradCAM as the best XAI method [49], our analysis indicates that GradCAM is not the ideal XAI approach for time-series NILM applications employing sequence-to-point architectures. Notably, GradCAM's faithfulness scores for dishwashers were significantly lower compared to other methods, implying an inability to identify crucial signal features. This observation is further supported by Fig. 2 and the results for the noisier REFIT dataset in Table II, where faithfulness scores for both washing machines (WM) and dishwashers (DW) were unsatisfactory. In an attempt to improve the score, we explored guided gradient technique used for GradCAM, but our findings point to further degradation of performance. On the other hand, our findings reveal that GradCAM++ method does outperform the original GradCAM, achieving better faithfulness and robustness. However, while the results demonstrate significant enhancements of GradCAM++ over GradCAM in these two aspects, the complexity of explanations generated by Grad-CAM++ is observed to be less than ideal. This finding suggests that the enhancements in faithfulness and robustness of GradCAM++ may come at the cost of increased complexity. Intriguingly, IG exhibited excellent performance for the complex signals (i.e., WM and DW) within the REFIT dataset. This implies that a zero signal is an appropriate choice for the baseline value of the IG algorithm for NILM-like data. Meanwhile, SmoothGrad (SG) produced robust results across most scenarios due the nature of the algorithm.

We acknowledge certain limitations in our work that necessitate further exploration. A primary constraint of the proposed evaluation framework is its inability to present specific steps for enhancing the effectiveness of explainability techniques. Nonetheless, our approach facilitates the comparison of various XAI methods, which remains valuable for identifying their strengths and weaknesses and guiding future research and development efforts. Furthermore, a crucial aspect involves examining the relationship and trade-offs between faithfulness, robustness, and complexity in XAI for NILM systems. Striking a balance among these metrics is vital for ensuring the utility, transparency, and, ultimately, trust in XAI NILM systems. Additionally, a key assumption in the context of XAI methods that were used in this work are that the proposed methods assume feature independence, which is a well-known issue

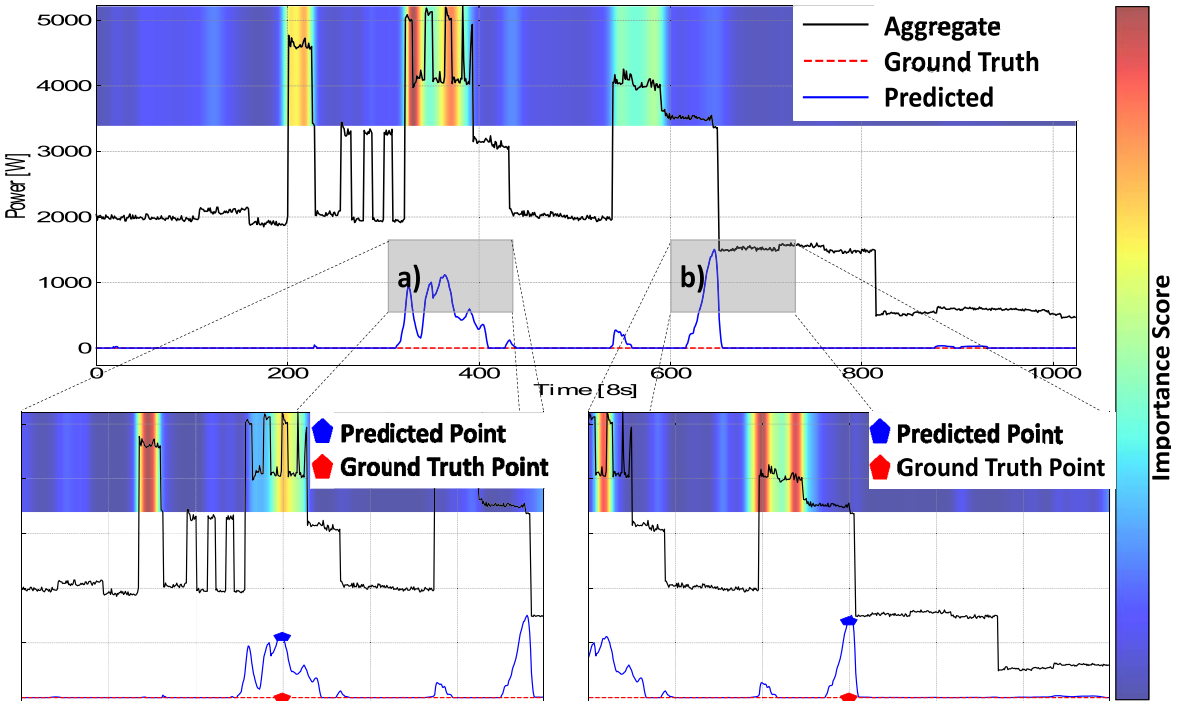

Fig. 3. Visual outline of the proposed approach showcasing an example of false positive prediction of washing machine for UK-DALE dataset, and the derived explanations using LRP. Two levels of explainability provide general, sequence-level (top image), and specific, point-level explanations (a and b), under a test scenario of signal incorrectly predicted as a washing machine.

in the field of XAI. To mitigate this, a new field of causal discovery has emerged; however the field is in infancy and its practical usefulness is still limited. Another assumption is related to robustness measure where we assume continuity, i.e., that small changes in the input (through introduction of Gaussian noise) will lead to small changes in the output explanation. Furthermore, to calculate the robustness score, we assume bounded input space, i.e., that maximum change in the explanation function is finite, which can be assumed for NILM signals as the aggregate function is bounded.

C. Visualisation via Heatmaps

The proposed approach enables two levels of explainability. On one hand, point-level explainability provides visual understanding of how a prediction of a single time step was made. It is specific to a point of reference. On the other hand, the visualization algorithm generates another, sequence-level explanation, showcasing the aggregate importance of the input signal for the prediction of the output, and acting as a more general representation of the importance. Point-level explanation is preferred to illuminate the features that have contributed to an individual point of the prediction especially if that point prediction is an outlier. Sequence-level explanations are more appropriate if trying to comprehend the decision on inference of a complete appliance duty cycle, such as why a time-series sequence was predicted as a Washing Machine.

Our visualization approach offers several advantages over the previously proposed methods. We tackle the more challenging regression scenario for the NILM problem compared to earlier work, which utilized a multi-class CNN for the simpler task of detecting appliance presence without recognizing on/off states [14]. Moreover, our method has been rigorously validated on numerous real-world datasets, demonstrating its adaptability and generalizability across diverse contexts. Unlike previous work that relied on a single dataset, our approach handles varied energy consumption patterns and appliance configurations, ensuring its practicality and resilience. In comparison to the regression-based XAI visualization method in NILM [13], our approach is more computationally efficient, as gradient-based methods require fewer iterations and calculations than occlusion sensitivity, making them well-suited for real-time applications and large-scale datasets. Additionally, our approach avoids the introduction of out-of-distribution scenarios caused by setting parts of the input signal to zero, ensuring that the generated explanations are more faithful to the model's behavior. A key strength of our method lies in its ability to provide multi-temporal explanations, offering insights into both local and global patterns at various levels of granularity, such as point-level and sequence-level explanations. This enhanced interpretability facilitates a better understanding of the NILM model's decision-making process and allows users to make more informed decisions based on the model's output. Furthermore, the gradient-based XAI methods can be applied to a wider range of DL-based NILM algorithms.

Fig. 2 provides an example of point explanations for a Dishwasher signal prediction from the UK-DALE dataset. This is a true positive prediction where the primary features contributing to the prediction of the middle point (marked with a blue pentagon) are displayed in a form of heatmap. We observe that most XAI methods highlight the true positive part of the

input signal. However, different XAI methods produce varying heatmap visualizations, underscoring the necessity for their quantitative quality evaluation. Comparing the results in Fig. 2 with the results displayed in Table I, LRP and SmoothGrad indeed showcase the best performance. We observe that both heatmaps highlight the truly important parts of the signal, suggesting high faithfulness, and that explanations are concise, pointing to low complexity. On the other hand, GradCAM shows the lowest faithfulness score, which can be observed from Fig. 2 as the GradCAM visualised explanation highlights an area that is not related to high activity of the dishwasher signal, suggesting a case of instability. To a smaller extent, this phenomenon is also observed in the case of IG. While the localization of feature importance scores in GradCAM++ improved compared to GradCAM, we observe a higher complexity of generated explanation. Comparing to LRP and SmoothGrad, we observe that the explanation heatmaps of GradCAM, GradCAM++, and IG cover a larger area of the input signal, and are of noticeably higher complexity, which is a finding that is reinforced by the complexity evaluation scores. Another scenario showcasing the mechanism behind a false positive prediction of a NILM DL model is presented in Fig. 3. In this example, a DW signature is incorrectly predicted as WM. We observe that the general explanation (on the top) enables us to assign the importance scores to the areas of the signal that the network deemed as indicative of a WM duty cycle. Looking further, the point-level explanations (a and b) enable us to understand that the DL model recognizes that there may be multiple cycles in a typical WM signature, which is supported by high importance score assigned to past signal spikes that look similar to a WM duty cycle. This can help the algorithm designer to improve the training and tuning process or adopt a multi-classification approach to better distinguish these multistate appliances with similar power level, duty cycle and duration.

V. Conclusion and Future Work

This paper proposes a methodology for determining the explainability of a time-series deep neural network regression non-intrusive load monitoring (NILM) problem. Specifically, we propose visualization via heatmaps approach by integrating XAI methods into the DL NILM and quantify explainability via faithfulness, robustness and complexity scores. As a way of overcoming the problem of transparency inherent to DL algorithms, the proposed approach provides a dual mode of explainability, one at a general, sequence level, and other at a specific, point level. Both levels of explainability can be used interchangeably based on the use case, as they provide varying degrees of specificity, i.e., they can deal with different scenarios when the decisions of NILM systems are unclear or difficult to explain. We show that this can be achieved without changing the architecture of the model. Furthermore, we define the core properties that should be considered when designing explainable NILM systems, and provide a strategy for quantitative evaluation of their explainability. We show that XAI methods, such as LRP, that have an inherent ability of dealing with noise, can lead to explanations that satisfy properties of being faithful to the performance of the model, robust to slight changes of input, and offer unambiguous interpretation of resulting heatmaps. The choice of the most appropriate methods should be guided by the target user of explanation, be it a domain expert, researcher, or customer, considering the trade-off between the aforementioned properties. By using the proposed method, the diverse set of needs of various users of the system can be satisfied, while maintaining the predictive performance and facilitating trust in the NILM system deployed in a real-world scenario.

In future work, it is important to extensively explore the relationship and trade-offs between the properties of faithfulness, robustness and complexity in XAI NILM approaches. For example, a highly faithful explanation that closely reflects the model's behavior may be more complex and harder to understand. Conversely, a simpler explanation may be more accessible but less faithful to the model's true decision-making process. Similarly, there may be cases where faithful explanations are sensitive to small changes in input data, resulting in a trade-off between faithfulness and robustness. Thus, striking the right balance between the metrics of explanation quality is crucial to ensure the usefulness of the XAI system. Our research focused on applying XAI to a CNN NILM algorithm. Future studies can extend this work to other NILM algorithms, including other deep learning-based approaches, to better understand the impact on the explainability performance and the generalisability of our findings. Another possible area of research could be combining different XAI techniques to create hybrid explanations, which may offer more comprehensive insights into NILM model behavior. Additionally, as one of the challenges in deploying NILM systems is the need for real-time processing and interpretation of energy consumption data, investigating the feasibility of real-time XAI methods for NILM applications would be a valuable contribution to the field, enabling more practical and actionable insights for users. Further work might also explore the relationship between visualizations and explainability performance for multi-appliance classification and regression. Finally, this framework can be extended to other applications in the energy sector to further promote reliable and safe integration of XAI in the smart grid.

References

[1] M. Zeifman and K. Roth, "Nonintrusive appliance load monitoring: Review and outlook," *IEEE Trans. Consum. Electron.*, vol. 57, no. 1, pp. 76–84, Feb. 2011.

[2] P. Huber, A. Calatroni, A. Rumsch, and A. Paice, "Review on deep neural networks applied to low-frequency NILM," *Energies*, vol. 14, no. 9, p. 2390, Jan. 2021.

[3] M. Kaselimi, E. Protopapadakis, A. Voulodimos, N. Doulamis, and A. Doulamis, "Towards trustworthy energy disaggregation: A review of challenges, methods, and perspectives for non-intrusive load monitoring," *Sensors*, vol. 22, no. 15, p. 5872, 2022.

[4] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Mach. Intell.*, vol. 1, no. 9, pp. 389–399, 2019.

[5] *Ethics Guidelines For Trustworthy AI*, Eur. Comm., Brussels, Belgium, 2019.

[6] C. Klemenjak, A. Faustine, S. Makonin, and W. Elmenreich, "On metrics to assess the transferability of machine learning models in non-intrusive load monitoring," 2019, *arXiv:1912.06200*.

[7] A. Vavouris, B. Garside, L. Stankovic, and V. Stankovic, "Low-frequency non-intrusive load monitoring of electric vehicles in houses with solar generation: Generalisability and transferability," *Energies*, vol. 15, no. 6, p. 2200, 2022.

[8] D. Murray, L. Stankovic, and V. Stankovic, "An electrical load measurements dataset of united kingdom households from a two-year longitudinal study," *Sci. Data*, vol. 4, no. 1, pp. 1–12, 2017.

[9] J. Kelly and W. Knottenbelt, "The U.K.-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five U.K. homes," *Sci. Data*, vol. 2, no. 1, pp. 1–14, 2015.

[10] C. Thoma, T. Cui, and F. Franchetti, "Secure multiparty computation based privacy preserving smart metering system," in *Proc. North Amer. Power Symp. (NAPS)*, 2012, pp. 1–6.

[11] H. Cao, S. Liu, L. Wu, Z. Guan, and X. Du, "Achieving differential privacy against non-intrusive load monitoring in smart grid: A fog computing approach," *Concurr. Comput. Pract. Exp.*, vol. 31, no. 22, p. e4528, 2019.

[12] Y. Zhang et al., "FedNILM: Applying federated learning to NILM applications at the edge," *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 2, pp. 857–868, Jun. 2023.

[13] D. Murray, L. Stankovic, and V. Stankovic, "Transparent AI: Explainability of deep learning based load disaggregation," in *Proc. 8th ACM Int. Conf. Syst. Energy-Efficient Build. Transp.*, 2021.

[14] R. Machlev, A. Malka, M. Perl, Y. Levron, and J. Belikov, "Explaining the decisions of deep learning models for load disaggregation (NILM) based on XAI," in *Proc. IEEE Power Energy Soc. General Meeting (PESGM)*, 2022, pp. 1–5.

[15] D. Batic, G. Tanoni, L. Stankovic, V. Stankovic, and E. Principi, "Improving knowledge distillation for non-intrusive load monitoring through explainability guided learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[16] D. Shin, "Why does explainability matter in news analytic systems? proposing explainable analytic journalism," *Journal. Stud.*, vol. 22, no. 8, pp. 1047–1065, 2021.

[17] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328

[18] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[20] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, 2015, Art. no. e0130140.

[21] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. A. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, 2018, pp. 80–89.

[22] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," 2017, *arXiv:1711.06104*.

[23] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," *Int. J. Human-Comput. Stud.*, vol. 146, Feb. 2021, Art. no. 102551.

[24] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in *Proc. AAAI/ACM Conf. AI Ethics Soc.*, 2020, pp. 180–186.

[25] U. Bhatt, A. Weller, and J. M. Moura, "Evaluating and aggregating feature-based model explanations," 2020, *arXiv:2005.00631*.

[26] P. Chalasani, J. Chen, A. R. Chowdhury, S. Jha, and X. Wu, "Concise explanations of neural networks using adversarial training," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1383–1391.

[27] R. S. Mollel, L. Stankovic, and V. Stankovic, "Explainability-informed feature selection and performance prediction for nonintrusive load monitoring," *Sensors*, vol. 23, no. 10, p. 4845, 2023.

[28] S. Ghosh and D. Chatterjee, "Artificial bee colony optimization based non-intrusive appliances load monitoring technique in a smart home," *IEEE Trans. Consum. Electron*, vol. 67, no. 1, pp. 77–86, Feb. 2021.

[29] Q. Liu et al., "Low-complexity non-intrusive load monitoring using unsupervised learning and generalized appliance models," *IEEE Trans. Consum. Electron.*, vol. 65, no. 1, pp. 28–37, Feb. 2019.

[30] B. Buddhahai, W. Wongseree, and P. Rakkwamsuk, "An energy prediction approach for a nonintrusive load monitoring in home appliances," *IEEE Trans. Consum. Electron.*, vol. 66, no. 1, pp. 96–105, Feb. 2020.

[31] J. Bartman and T. Kwater, "Identification of electrical appliances using their virtual description and data selection for non-intrusive load monitoring," *IEEE Trans. Consum. Electron*, vol. 67, no. 4, pp. 393–401, Nov. 2021.

[32] G.-F. Angelis, C. Timplalexis, S. Krinidis, D. Ioannidis, and D. Tzovaras, "NILM applications: Literature review of learning approaches, recent developments and challenges," *Energy Build.*, vol. 261, Apr. 2022, Art. no. 111951.

[33] D. Yang, X. Gao, L. Kong, Y. Pang, and B. Zhou, "An event-driven convolutional neural architecture for non-intrusive load monitoring of residential appliance," *IEEE Trans. Consum. Electron.*, vol. 66, no. 2, pp. 173–182, May 2020.

[34] Y. Liu, L. Zhong, J. Qiu, J. Lu, and W. Wang, "Unsupervised domain adaptation for nonintrusive load monitoring via adversarial and joint adaptation network," *IEEE Trans. Ind. Informat.*, vol. 18, no. 1, pp. 266–277, Jan. 2022.

[35] C. Zhang, M. Zhong, Z. Wang, N. H. Goddard, and C. Sutton, "Sequence-to-point learning with neural networks for non-intrusive load monitoring," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 32, 2018, pp. 2604–2611.

[36] J. Jiang, Q. Kong, M. Plumbley, and N. Gilbert, "Deep learning-based energy disaggregation and on/off detection of household appliances," *ACM Trans. Knowl. Discov. Data*, vol. 15, no. 3, pp. 1–21, 2021.

[37] D. Shin and Y. J. Park, "Role of fairness, accountability, and transparency in algorithmic affordance," *Comput. Human Behav.*, vol. 98, pp. 277–284, Sep. 2019.

[38] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[39] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," 2018, *arXiv:1806.08049*.

[40] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, Nov. 2017.

[41] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018.

[42] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, "On the (in)fidelity and sensitivity for explanations," in *Proc. NeurIPS*, Nov. 2019.

[43] J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 9525–9536.

[44] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.

[45] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4765–4774.

[46] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. NIPS*, Dec. 2018, pp. 7786–7795.

[47] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018, pp. 839–847.

[48] Y.-Y. Yang, C. Rashtchian, H. Zhang, R. Salakhutdinov, and K. Chaudhuri, "A closer look at accuracy vs. robustness," in *Proc. Adv. Neural Inf. Proc. Syst.*, vol. 33, 2020, pp. 8588–8601.

[49] R. Machlev, M. Perl, J. Belikov, K. Y. Levy, and Y. Levron, "Measuring Explainability and trustworthiness of power quality disturbances classifiers using XAI—Explainable artificial intelligence," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5127–5137, Aug. 2022.

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