Engineering Applications of Artificial Intelligence 141 (2025) 109766

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

XNILMBoost: Explainability-informed load disaggregation training enhancement using attribution priors

Djordje Batic [*], Vladimir Stankovic, Lina Stankovic

University of Strathclyde, Department of Electronic and Electrical Engineering, 204 George St., Glasgow, G1 1XW, Scotland, UK

ARTICLE INFO

Keywords:
Explainable deep learning
Load disaggregation
Non-intrusive load monitoring
Trustworthy artificial intelligence

ABSTRACT

In the ongoing energy transition, characterized by increased reliance on distributed renewable sources and smart grid technologies, the need for advanced and trustworthy artificial intelligence (AI) in energy management systems is crucial. Non-intrusive load monitoring (NILM), a method for inferring individual appliance energy consumption from aggregate smart meter data, has gained prominence for enhancing energy efficiency. However, advanced deep neural network models used in NILM, while effective, raise transparency and trust concerns due to their complexity. This paper introduces a novel explainability-informed NILM training framework, specifically designed for low-frequency NILM. Our approach aligns with principles for trustworthy AI, focusing on human agency and oversight, technical robustness, and transparency, incorporating explainability directly into the training phase of a NILM model. We propose a novel iterative, explainability-informed NILM training algorithm that uses attribution priors to guide model optimization, including implementation and evaluation of the framework across multiple state-of-the-art NILM architectures, namely, convolutional, recurrent, and dilated causal layers. We introduce a novel Robustness-Trust metric to measure joint improvement in predictive and explainability performance, utilizing explainability metrics of faithfulness, robustness and effective complexity while analyzing model predictive performance against NILM-specific regression and classification metrics. Results broadly show that robust models achieve better explainability, while explainability-enhanced models can lead to improved model robustness. Together, our results demonstrate significant improvements in robustness and transparency of NILM systems across various appliances, model architectures, measurement scales, types of buildings, and energy usage patterns. This work paves the way for more transparent and trustworthy deployments in AI-driven energy systems.

1. Introduction

Modern energy systems rely on the capacity to gather and discover insights from real-time consumption data, facilitating enhanced monitoring of energy distribution and consumption, reducing operational costs, and improving energy efficiency (Hossain et al., 2016). Smart meters can be used to extract useful information about energy consumption, which is relayed to utility companies, consumers, prosumers, and other parties focused on achieving energy efficiency objectives (Kabalci, 2016). Smart meters enable accurate billing and increased awareness of energy usage patterns on the user side, promoting energy-efficient behavior. On the other hand, smart meters can facilitate the integration of demand response measures through variable tariffs, as well as an increased understanding of customer needs on the utility side (Siano, 2014). To extract energy consumption data of individual loads, Non-intrusive load monitoring (NILM) has shown promising results compared to intrusive submetering of energy usage of individual devices. NILM aims to algorithmically infer the energy consumption of individual electrical appliances from the aggregate metered power consumption collected by a smart metering device (Huber et al., 2021). In recent years, NILM has experienced rapid development due to advancements in artificial intelligence (AI) techniques such as deep neural network (DNNs); see recent review papers (Huber et al., 2021; Angelis et al., 2022). Although DNN-based NILM algorithms demonstrate good disaggregation performance, there are still major challenges to address before large-scale deployment and adoption. One of the core challenges is related to high complexity of DNN models, often called “black-box” models, which leads to diminished understanding of the decisions that the model makes. As “black-box” DNN-based NILM systems are deployed at scale, it is of high importance to ensure that there is a procedure for explaining the outputs of these systems (Batic et al., 2023b).

There are many reasons why explainability in DNN-based NILM systems is desirable: explainability can lead to user confidence that the

[*] Corresponding author.
E-mail address: djordje.batic@strath.ac.uk (D. Batic).

https://doi.org/10.1016/j.engappai.2024.109766
Received 5 April 2024; Received in revised form 24 October 2024; Accepted 27 November 2024

AI algorithms reliably infer their energy consumption; helps reveal potential biases; enables developers to better understand how the systems behave in various conditions; helps in the assessment of vulnerabilities; and enables alignment with policy requirements and legal standards. This is particularly important for a system such as NILM that leverages on processing sensitive user data that can reveal various household activities and has implications on the financial security of the users. As a result, designing explainable NILM systems is key to facilitating trust and wider adoption of NILM (Kaselimi et al., 2022; Batic et al., 2023b).

To facilitate trust in AI, the European Commission has established a regulatory framework (Commission and Directorate-General for Communications Networks, and Technology, 2019) that emphasizes the 7 core principles of Trustworthy AI: Human agency and oversight, Technical robustness and safety, Privacy and data governance, Transparency, Diversity, Non-discrimination and fairness, Societal and environmental well-being, and Accountability. NILM should benefit people by enabling them to track their personal energy usage data, increasing the accessibility of education about consumption habits, and helping to make more informed choices about demand flexibility, choice of tariffs, energy saving and climate goals. At the same time, proper oversight mechanisms need to be ensured by enabling humans to shape the behavior of the AI by encoding prior beliefs about how the model should behave, all of which is encompassed in the principle of Human agency and oversight. Next, AI-based NILM should ensure the principle of technical robustness and safety through accurate and reliable results, by providing correct predictions that are meticulously evaluated in a wide range of input scenarios. Furthermore, NILM systems should be transparent and provide traceability and documentation of the data collection and generation process, communication about the capabilities and limitations of the system, as well as explainability of the AI decision-making process.

Previous studies in the field of Explainable AI (XAI) primarily focus on the development of techniques that increase the model transparency by quantifying the importance of individual input features through explanations. These methods, also known as feature-attribution methods, are effective in revealing problems in a model, understanding the model decisions, or revealing dataset bias. However, feature attribution methods may place too much importance on undesirable features, provide unstable explanations under the presence of input noise, or rely on too many features when low complexity of explanations is desired (Bhatt et al., 2020). As a result, more recent literature has emphasized the need for a mathematical definition of explanation quality and evaluation of feature attribution methods (Alvarez-Melis and Jaakkola, 2018; Ancona et al., 2017; Bhatt et al., 2020).

XAI approaches for NILM are still in their infancy, with limited literature available (Murray et al., 2021; Mollel et al., 2023; Batic et al., 2023b,a). As XAI-based solutions for NILM continue to grow, it is of crucial importance to properly account for transparency property outlined in the EU requirements for Trustworthy AI. As such, explainability evaluation is one of the core components that enables an overview of the real-world performance of feature attribution methods. In addition, it remains unclear how existing NILM architectures, with demonstrated high accuracy, can be made more explainable, for example, by considering model explainability during the training process. Notably, to the best of our knowledge, combining the use of explainability during the training phase with a comprehensive quantitative evaluation of explainability in the context of NILM, has not been attempted before. This gap in the literature presents a significant opportunity to enhance both the interpretability and performance of NILM models.

Building upon recent advances in AI research, recent work has made significant strides in various aspects of Trustworthy NILM. AI-based NILM has leveraged on various architectures to provide accuracy and reliability of predictions (Murray et al., 2019), embedding human oversight through inclusion of user or expert knowledge in the learning process (Todic et al., 2023), or XAI methods for transparency (Mollel et al., 2023; Batic et al., 2023b,a). However, there has been no work that aims to unite the three aforementioned principles of technical robustness, transparency, and human oversight in a single system. In this work, we propose the first explainability-informed NILM training framework for low-frequency NILM. The proposed framework aims to directly mitigate shortcomings of existing NILM approaches in line with EU guidelines for Trustworthy AI, by prioritizing robustness, transparency, and human oversight during the learning process, leveraging on prior human intuition about the behavior of explanations of AI outputs to constrain the model explanations during training and help the model be more accurate and reliable. The vital benefit of our approach is the ability to directly train the NILM neural network to be more explainable, by manipulating the gradients during the training process. In addition, we show that such enhancement can improve the technical robustness of the system by improving the predictive performance across multiple real-world scenarios. Lastly, we generalize our findings by evaluating the predictive and explainability performance across multiple and distinct model architectures and show the link between architectural choices and explainability performance.

In summary, the contributions of this study are as follows:

- We propose the first explainability-informed learning framework for load disaggregation/NILM systems that jointly promotes Trustworthy AI principles of Human agency and oversight, Transparency, and Technical robustness and reliability.
- We present attribution prior NILM training, an iterative algorithm that leverages on human intuition to constrain the NILM model towards better explainability by preventing incorrect assignment of feature attributions.
- We demonstrate how the proposed explainability-informed learning framework can improve the robustness of NILM models by improving their predictive performance.
- We demonstrate how the proposed explainability-informed learning framework can improve the transparency of NILM models by improving their explainability performance across various NILM-specific explainability evaluation metrics.
- We present a comprehensive evaluation of explainability and predictive performance across three state-of-the-art NILM architectures: convolutional, recurrent, and causal networks, as well as four distinct XAI methods by utilizing three publicly available datasets comprising real-world measurements from households in the UK, USA, and Greece.

The rest of the paper is organized as follows. Section 2 discusses prior work in explainable NILM. The proposed explainability-informed training methodology is described in Section 3, while the experimental results and key findings are presented in Section 4. Finally, Section 5 provides concluding remarks, as well as directions for future work.

2. Problem statement and literature review

2.1. NILM problem statement and low-frequency NILM algorithms

Given a sequence of aggregated power consumption $\mathbf{y} = (y_1, y_2, \dots, y_T)$, captured at time $t = \{1, 2, \dots, T\}$, the goal of a NILM algorithm is to determine the individual power contribution x_t^i of appliance $i \in \{1, 2, \dots, M\}$, such that the aggregate can be represented as a combination of individual power consumption of M appliances and a term ϵ_t, which denotes noise from unknown appliances contributing to the aggregate signal and measurement noise:

$$y_{t=1\dots T} = \sum_{i=1}^{M} x_t^i + \epsilon_t \tag{1}$$

To extract the power consumption of a selected appliance $i \in \{1, 2, \dots, M\}$, the majority of NILM approaches are focused on filtering the noise term ϵ as well as all other appliance signals, which is a non-trivial problem due to statistical differences in activation length, time of use, frequency, and peak power usage. To detect an appliance of interest, NILM can be treated as either a classification or regression problem. Classification-based NILM infers the on/off state of an appliance i at time t, based on the aggregate signal y_t. On the other hand, regression-based NILM aims to directly infer x_t^i.

Very early NILM research primarily utilized high-frequency power measurements, using sampling frequencies in the order of kHz or higher. However, the landscape has shifted significantly with national rollouts worldwide of standard smart meters, for which data stored is in the order of 1 s to 30 min. This transition to lower-frequency measurements was driven by several practical factors: reduced privacy concerns, more manageable data storage requirements, and simpler data handling processes. Additionally, previous research has shown that appliance recognition capability varies with sampling frequency, with long-duration activations actually benefiting from reduced sampling rates compared to high-frequency (sub-second measurements) (Armel et al., 2013; Huchtkoetter and Reinhardt, 2020). Furthermore, high-frequency NILM, has already demonstrated very good disaggregation accuracy, leveraging on ability to identify transient features and harmonic content, with little room for further improvement unlike low-frequency NILM. As a result, the challenges of low-frequency NILM has been the main focus of research in recent years due to the abundance of smart meter measurement data and advancements in machine learning (Angelis et al., 2022).

In order to infer individual appliance consumption, various machine learning approaches have been proposed in the recent literature, where DNN approaches form the basis of state-of-the-art implementations according to the recent review of Angelis et al. (2022). Convolutional Neural Networks (CNNs) form the majority of implementations - Zhang et al. (2018) propose a sequence-to-point (seq2point) learning approach using a CNN; Pan et al. (2020) address the high computational complexity of seq2point and propose a CNN architecture for sequence-to-subsequence learning; Chen et al. (2019) use a two subnetworks that are connected in order to infer both regression and classification outputs; Murray et al. (2019) propose a CNN model that provides generalizability to new domains; Massidda et al. (2020) perform multilabel classification using a CNN architecture with pooling layers at different time scales. Another common DNN approach to NILM are Recurrent Neural Networks (RNNs); Zhang et al. (2022) use a multi-quantile RNN to disaggregate the loads and improve the demand side management of solar energy; Krystalakos et al. (2018) propose a Gated Recurrent Units (GRU) approach that reduces memory usage and computational complexity while achieving good disaggregation performance, while Tanoni et al. (2022) proposes a Convolutional RNN approach for multi-label classification of appliances. Lastly, other literature attempts to introduce new learning mechanisms include generative adversarial networks (GANs) (Pan et al., 2020), temporal-causal networks (Harell et al., 2019) and attention mechanisms (Yue et al., 2020). A DNN-focused review for low-frequency NILM (Huber et al., 2021) provides a detailed review of current DNN NILM approaches, where GRU and CNN architectures and their variants, including WaveNet with dilated convolutions (Harell et al., 2019), have been shown to achieve good performance over a range of appliances with well documented publicly available code for reproducibility, and therefore inform the architectures we consider in our proposed work.

2.2. Explainable AI for low-frequency NILM

Explainability refers to the ability to explain both the technical processes of an AI system and the related human decisions (e.g., application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. The use of DNNs generally negatively impacts our understanding of how the decisions are made by the system. In NILM, previous studies have used explainability tools to determine local and global feature importance of decision tree approaches to design a methodology that informs feature selection for each appliance class (Mollel et al., 2023). However, when translating to a regression-based task where the usage of DNNs is more common, explainability presents a larger problem due to the naturally less interpretable nature of DNNs compared to decision tree algorithms. Authors in Murray et al. (2021) propose the first XAI methodology for NILM by using occlusion sensitivity to offer a visual understanding of significant features for the prediction of DNN-based NILM model. This method involves occluding random regions of a signal and analyzing the impact it has on prediction performance. However, this method poses sizeable computational challenges, primarily because of its sliding window mechanism. Moreover, this approach occludes parts of the signal by setting the consumption power values to zero, which is not a realistic scenario and might represent an out-of-distribution scenario where the model can struggle to produce intelligent outputs. A recent study (Machlev et al., 2022) compares the success of using the GradCAM XAI technique against occlusion sensitivity for visualizing significant input features of a NILM classifier. However, authors define a significantly simpler problem statement where a multi-class CNN is used to determine solely the existence of an appliance in the input time-series, without inferring the on/off state or the sample-by-sample energy consumption values typical for regression approaches. Furthermore, they focus solely on a single XAI method, which is a major concern, as XAI methods can generate unreliable explanations, contributing to a diminished understanding and opportunities to exploit the vulnerabilities of the NILM system. In order to promote the adoption of XAI in NILM, Batic et al. (2023a) propose a visualization procedure that explains the outputs of a seq2point algorithm on multiple levels of time granularity. 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. Lastly, the authors in Batic et al. (2023b) show that explainability can be used to improve the performance of knowledge distillation in NILM. However, this approach limits the assessment of explainability improvement to subjective evaluation, lacking a more rigorous evaluation approach.

2.3. Explainability evaluation for NILM

In recent years, researchers and policy makers have warned about the use of explainable AI in medium to high-risk scenarios as there are pitfalls associated with the use of different XAI methods due to their sometimes unreliable outputs which might lead to incorrect assumptions about model behavior or opportunities to exploit the vulnerabilities of the system. As a result, there has been a drive for more rigorous and objective evaluation strategies of XAI methodologies that assess their quality and promote higher transparency of the AI system.

Recent work in NILM (Batic et al., 2023a) has proposed three core properties that facilitate the evaluation of explainability of NILM-like methods: faithfulness, robustness, and low complexity. First, faithfulness represents a property that ensures that the provided explanations accurately correspond to model performance, based on the notion that removing or obscuring important input features discovered by an XAI method should have a significant negative effect on the predictive performance or model confidence. In other words, faithfulness enables understanding of how feature importance scores influence the prediction — a high faithfulness score suggests that the explainability method is able to correctly identify the important features of the input signal, indicated by a large drop in prediction accuracy after obfuscation of important features. Second, to determine the reliability of an XAI method, robustness evaluates how the XAI method performs under slight changes of the input. Recent XAI research suggests (see Batic et al. (2023a) and references therein) that slight changes in the input, similar to adversarial noise, can lead to significant changes in the generated explanation outputs, while retaining the same or similar predictions. To define the relationship between the input data and reliability of XAI methods for NILM, Batic et al. (2023a) estimate

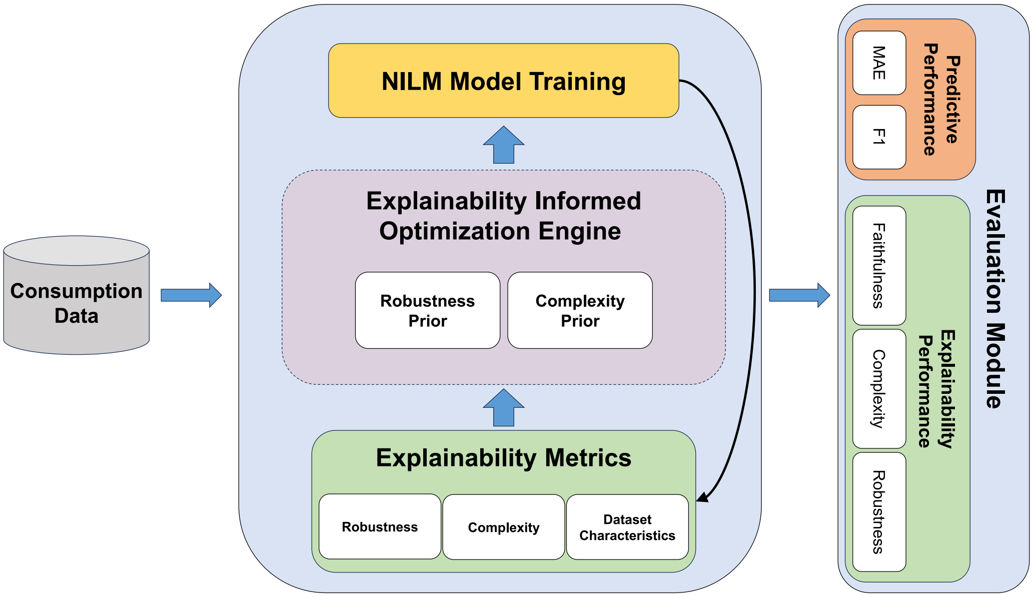


Fig. 1. Overview of the proposed explainability-informed NILM training framework.

the value of Lipschitz constant, where methods with low Lipschitz value scores display more stable behavior under the presence of input noise, leading to higher reliability of XAI system. Finally, explanation complexity indicates the degree of quality w.r.t. human understanding of the produced explanation, as it quantifies the entropy of the XAI output. A highly undesirable scenario is an explanation that does not provide an adequate level of clarity and conciseness. Thus, Batic et al. (2023a) evaluate the complexity of an XAI NILM system by quantifying the relationship between the entropy of explanation, defined by the Gini index, and the level of dataset "noise" in the input, defined by a measure of noise-to-aggregate ratio proposed in Makonin and Popowich (2015).

2.4. Summary

From the above literature review, it is evident that explainability is still a new concept in NILM. Still, there has been work aimed to formalize the concept of explainability in the context of NILM. However, to the best of our knowledge, there is very limited work on joint improvement of predictive performance and explanability of NILM architectures. We address this gap by proposing XNILMBoost, a framework for explainability-informed NILM training enhancement using attribution priors. We show that optimizing for explanability is possible and that it can also lead to improved predictive performance. Moreover, the whole procedure is performed during the training of the model. We propose an iterative training method that can be used to select an optimal prior and fine-tune any NILM model using XNILMBoost, by optimizing for both predictive and explainability performance without a need for labeled data. We generalize our findings by using XNILMBoost on three distinct architectures – GRU, CNN, and WaveNet – for which publicly available code was available for reproducibility, and three real-world NILM datasets – REDD and UK-DALE – that are the most popular datasets on which various NILM DNN architectures were demonstrated (Huber et al., 2021). Additionally, to extend the empirical results of our study, we includeed results from additional Plegma dataset from Greece, which contains appliances rarely shown in NILM datasets, such as AC and Boiler appliances. Lastly, we show the explainability improvement using various NILM-specific explainability metrics (Batic et al., 2023a) — faithfulness, robustness, and complexity.

3. Methodology

Our explainability-informed learning framework for low-frequency NILM is shown in Fig. 1. The backbone of our approach is the explainability-informed optimization engine, which is responsible for the optimization of explainability performance depending on the training requirements. The proposed framework iteratively trains a NILM neural network by proposing an explainability-informed training enhancement strategy by first receiving the information related to dataset statistics, as well as explainability evaluation results for the properties of robustness and complexity, which can be inferred without any labeled data. The training is performed incrementally until the explainability improvement requirements are met.

To diversify the experimental evaluation and generalizability of our proposed approach, we train on three different state-of-the-art architectures, with the aim of incorporating a broad set of techniques including convolutional, recurrent, and dilated causal layers. Lastly, we perform a rigorous experimental evaluation of explainability performance under various real-world scenario datasets, including an ablation study. The following subsections provide a detailed overview of the proposed techniques, as well as the explainability-informed training workflow.

3.1. Explainability evaluation dataset

The explainability evaluation dataset is sampled per appliance. First, to gather the appliance activations, we gather dataset characteristics and define the power-on threshold of appliance activation, as well as minimum on and off duration. Next, after applying the threshold and computing the on/off events, we calculate the distance between the subsequent on and off events to obtain the appliance activation duration. Finally, we select $n = 30$ random samples of activations that are longer than a predefined appliance-specific length and select a window of size ω centered around the appliance activation window. Given a dataset with a granularity of 8 s, 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 h and 15 min of measurements, in line with the average length of a duty cycle of most WM and DW devices. For the Microwave (MW), activation length $\omega = 80$ samples was chosen, which corresponds to around 10 min. Finally, if the total length of the activation length of interest is larger than ω, the first ω data samples are selected. In the case of Plegma dataset, which contains 10 s granularity measurements, activation length of WM appliance is set to $\omega = 820$, while Boiler and AC appliances have activation length set to $\omega = 700$ and $\omega = 1000$, respectively.

3.2. Low-frequency NILM algorithms

For the purpose of demonstrating the adaptability and generalizability of our proposed methodology across diverse contexts, we employ three distinct DNN architectures. To best exemplify the variety of algorithmic approaches for NILM, we use CNN-based (Zhang et al., 2018), GRU-based (Rafiq et al., 2021), and WaveNet-based (Harell et al., 2019) NILM networks, as seen in Fig. 2. One of the most cited CNN-based approaches for NILM is seq2point architecture (Zhang et al., 2018). The seq2point algorithm slides a window across the input aggregate signal to predict the energy consumption at the central point of the sliding window. Previous studies show that this produces a favorable approximation of the target distribution compared to previous NILM approaches (Jiang et al., 2021). On the other hand, RNN-based approaches have been consistently popular in the NILM literature. In this paper, we use a GRU architecture, a variant of the Long Short Term Memory (LSTM) network, that is designed for time series data. Compared to LSTMs, GRU networks deal better with the vanishing gradient problem and are designed to be more computationally efficient. Lastly, given varying activation periods and lengths of appliances, WaveNet-based networks that employ dilated causal convolutions have proven to achieve good disaggregation performance (Harell et al., 2019). To capture various input time steps, dilated causal layers have various dilation factors that grow in depth and allow the network to capture very long-range dependencies. For more details on the selected NILM architectures, readers are referred to Zhang et al. (2018), Rafiq et al. (2021), and Harell et al. (2019).

3.3. Explainability enhancement using attribution priors

The proposed explainability-informed training using attribution priors refers to the process where the model's gradients are altered during the model training process to optimize the explainability performance of attribution methods used for visualization of important features of the model. Rather than considering explainability as a post-processing step of model development, this approach enables learning of correct assignment of input feature attributions. Since it is often unknown which input features will contribute highly to the prediction of a model, we define an attribution prior that captures human oversight and guides model towards correct attribution assignment.

In the context of training a typical DNN model, the primary objective is to learn a non-linear function f characterized by a set of parameters θ. This learning process utilizes a dataset comprising n samples, each represented as a pair (x, y). The goal is to minimize a loss function $\mathcal{L}$, which can be formally expressed as:

$$f = argmin_{\theta} \frac{1}{n} \mathcal{L}(\theta; x, y) + \alpha \mathcal{R}(\theta), \tag{2}$$

In this formulation, α represents a scalar value that modulates the influence of the regularization function $\mathcal{R}$. This approach is commonly employed in supervised learning scenarios, where the regularization term helps prevent overfitting and improves the model's generalization capabilities.

The concept of attribution prior can be formalized for a given feature attribution method $m(\theta, x)$ as a function $p : \mathcal{R} \rightarrow \mathcal{R}$. This function assigns a scalar weight to the attribution features of the function f with input x. Incorporating this notion, the attribution prior-based training can be expressed mathematically as:

$$f = argmin_{\theta} \frac{1}{n} \mathcal{L}(\theta; x, y) + \alpha \mathcal{R}(\theta) + \beta p(m(\theta, x)), \tag{3}$$

In this formulation, β serves as a scalar value that modulates the impact of the attribution prior p. To optimize computational efficiency and reduce training time, the function m is calculated using the standard approach of multiplying the input with the gradient. Within the scope of this research, we explore and implement two distinct types of attribution priors p, each offering unique characteristics and potential benefits to the training process.

Our first approach is motivated by the observation that explainability methods become less effective and human-interpretable when they deem most input features as important. To address this, we introduce a low-complexity prior that encourages models to assign importance to a limited number of input features during training of a model. This approach improves the clarity and interpretability of explanations by highlighting only the most crucial features. To quantify the conciseness of the explanation output, we employ a differentiable function that calculates the Gini coefficient, measuring the statistical dispersion of the generated attribution values. This choice is supported by previous research (Chalasani et al., 2020) indicating that the Gini coefficient serves as a reliable indicator of model explanation complexity. Formally, given a feature attribution method m, we define a low complexity attribution prior that promotes more focused and interpretable explanations while maintaining model performance:

$$p(m(\theta, x)) = \frac{\sum_{a=1}^{\omega} (2a - \omega - 1) m(\theta, x)}{k + \sum_{a=1}^{\omega} m(\theta, x)}, \tag{4}$$

where k is a small value added for numerical stability. This complexity prior penalizes neural networks for creating complex attributions that assign high importance to numerous input features.

Additionally, we propose an alternative method focused on gradient smoothness to reduce incorrect feature attribution. This approach, which we term the robustness prior, applies a total variation denoising algorithm to feature attribution maps. It is defined as:

$$p(m(\theta, x)) = \sum_{i} |m_{i+1}(\theta, x) - m_i(\theta, x)|. \tag{5}$$

The robustness prior aims to minimize unstable attributions and promote gradient smoothness, encouraging attribution maps that are faithful to model outputs and predictive performance. The complexity and robustness priors, though distinct in their immediate objectives, function as complementary approaches to enhance the interpretability and reliability of feature attributions in neural network models. The complexity prior aims to reduce the number of important features, promoting concise explanations, while the robustness prior focuses on smoothing the gradient to ensure stable and consistent attributions. Together, they guide the model towards simpler, more stable decision boundaries. This synergy can lead to models that are both more interpretable and more robust to input variations. Both priors can be viewed as regularization techniques in the attribution space, contributing to the broader goal of regularizing explanations in interpretable machine learning.

3.4. Explainability-informed training

Finding the optimal attribution prior that represents the best tradeoff between explainability and predictive performance can be a tedious task. To address this, we propose an explainability-informed selection process using a novel metric: the Robustness-Trust (ROTR) metric. This approach enables us to iteratively determine the optimal prior for a given NILM model while considering multiple performance aspects simultaneously. Instead of evaluating metrics independently, we consider multiple metrics within a single term that exemplifies the improvement in transparency of a trained model. ROTR metric can be defined as:

$$ROTR = \frac{XF_{prior}}{XF_{base}} \frac{XR_{base}}{XR_{prior}} \frac{XC_{prior}}{XC_{base}}, \tag{6}$$

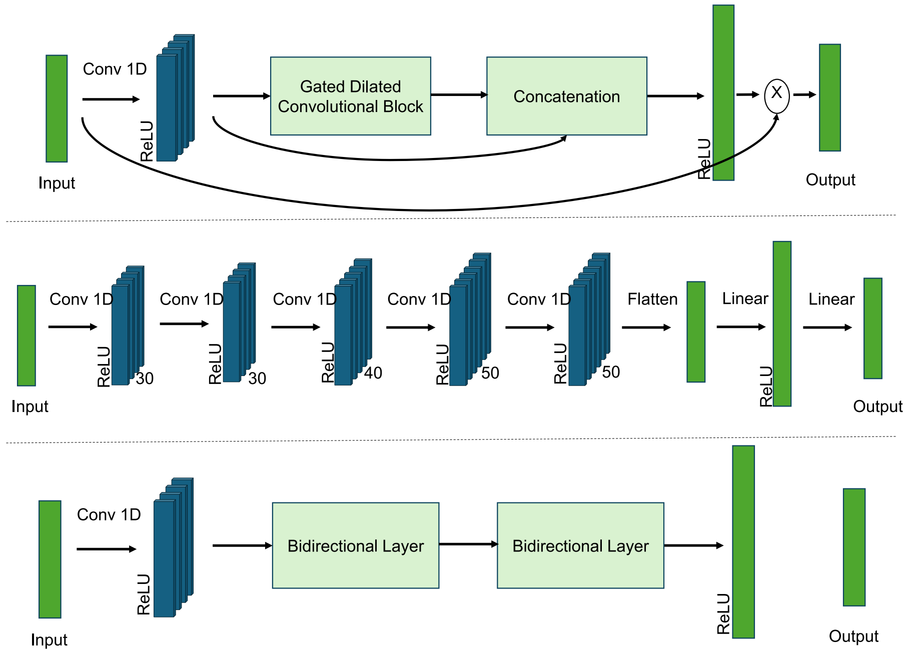


Fig. 2. Model architecture for the four NILM models used in this study. The upper subfigure describes a CNN network (Zhang et al., 2018), whereas the middle and bottom subfigures indicate GRU (Rafiq et al., 2021) and WaveNet (Harell et al., 2019) architectures, respectively.

where XF, XR, and XC represent the faithfulness, robustness, and effective complexity metric scores, respectively.

This metric quantifies the improvement in explainability performance, with scores above 1 indicating beneficial improvement. For more information related to considered explainability metrics, readers are referred to Batic et al. (2023b). To thoroughly explore the trade-offs between the smoothness and low complexity priors, we employ an iterative optimization process. This process involves systematically varying the influence of both priors through their β hyperparameters and analyzing their combined impact on model performance. We have observed that while the smoothness prior enhances gradient stability, it may occasionally conflict with identifying sharp feature boundaries. Conversely, the low-complexity prior promotes concise explanations but might oversimplify complex data relationships. The optimal balance between these priors often yields the best ROTR scores, though this balance can vary depending on the specific NILM task and the characteristics of the data set.

To gather data for $ROTR$ computation, the procedure described in Section 3.1 is used with a distinction that activation samples are collected from the trained base model instead of the ground truth, which allows evaluation without ground truth labels. This is beneficial because, under such a framework, any existing NILM architecture that theoretically supports explainability-informed training can be retrained or fine-tuned.

$ROTR$ is a metric that determines the overall improvement of explaianbility performance of a NILM model trained in our proposed framework. For each metric contained in $ROTR$, the values are calculated before and after applying the prior. $ROTR$ combines multiple metrics in a multiplicative way to indicate the overall improvement of the model. This is achieved by computing the relative change of individual metrics of explainability and aggregating them under a single term that balances all contributing metrics. $ROTR$ score greater than 1, indicates that the proposed prior achieves a beneficial improvement. In such a case, the model is considered "explainability enhanced" and can be passed to the evaluation module. On the other hand, scores below 1 indicate no change, or degradation of performance, thus triggering a new iteration of the optimization engine. Therefore, $ROTR$ indicates the relative change in the improvement of explainability-informed training, which considers both predictive and explainability performance as an indication of performance quality.

$ROTR$ formulation enables iterative training of an explainability-informed NILM model. To frame the problem, an expert needs to define the optimal starting priors, given the real-world scenario. For the purpose of this work, we utilize the aforementioned smoothness and low complexity priors. Then, iterative training is performed by selecting a range of β hyperparameter values that indicate the relative importance of the prior during training. Recognizing the interdependency of these parameters, we adopt a grid search optimization strategy within a predefined parameter space, guided by the computed $ROTR$ values. The metric is defined such that values exceeding 1 signify a net improvement in accuracy-explainability trade-off, providing a unified criterion for model optimization. This iterative process begins with an initial set of β hyperparameters, which are incrementally adjusted based on their impact on $ROTR$. During each iteration, the model undergoes training and evaluation, after which $ROTR$ is calculated to assess the joint improvement. If $ROTR > 1$, the adjustments are considered to have contributed positively, and the hyperparameters are further fine-tuned in the direction that maximizes $ROTR$. In contrast, if $ROTR < 1$, it indicates stagnation or deterioration in explainability, prompting a reevaluation of hyperparameter adjustments. This feedback loop creates a mechanism in which the model self-adjusts, seeking hyperparameter configurations that elevate $ROTR$ above the threshold of 1. To ensure a thorough exploration of the hyperparameter space while avoiding local optima, we employ adaptive hyperparameter selection. This method not only facilitates a granular optimization but also embeds a learning paradigm where the model iteratively converges towards an optimal balance between explainability and predictive accuracy, improving the design of DL-based NILM systems. By systematically varying the influence of the two priors on the training process, we identify the optimal combination that minimizes the objective function, a composite measure of performance accuracy, gradient smoothness, and explanation complexity, thereby demonstrating the effectiveness of our dual-hyperparameter regularization framework that aims to improve the explainability performance without comprising the predictive performance.

3.5. Explainability methods

To quantify the explainability performance of the networks used in this work, we adapt the explainability methods and evaluation methodology described in Batic et al. (2023a). To accommodate to different architectures used in this work, the visualization procedure is modified for GRU and WaveNet networks. GRU network performs prediction of the last point of the input signal. Thus, to compile the sequence-level explanation, a triangular weighting function gives the highest importance to the end of the window. On the other hand, the WaveNet architecture computes sequential output of the same length as the input, thus sequence-level explanation is inherently provided. In this work, we utilize 4 popular XAI methods: GradCAM (Selvaraju et al., 2017), GradCAM++ (Chattopadhay et al., 2018), IntegratedGradients (Sundararajan et al., 2017), and SmoothGrad (Smilkov et al., 2017). The created sequence-level explanations are subjected to a quantitative evaluation of quality. Considering a diverse set of needs and possible deployment scenarios, the explainability evaluation is defined as alignment with three desirable properties of explanations presented in Section 2.3 -faithfulness, robustness, and low complexity.

4. Experimental results

This section provides descriptions of the datasets used to conduct experiments, metrics used to evaluate the proposed methodology, as well as parameters to enable reproducibility of results.

4.1. Datasets and appliances

To evaluate our approach, we conducted experiments on appliances from UK-DALE (Kelly and Knottenbelt, 2015), REDD (Kolter and Johnson, 2011) and Plegma (Athanasoulias et al., 2024) datasets. All three datasets contain aggregate and appliance level energy consumption, where UK-DALE contains measurements from five houses in the UK with up to 4.3 years of data, REDD contains measurements from six different houses in the United States with up to 6 weeks of data, while Plegma contains measurements from 13 different houses in Greece over a period of 12 months. Energy consumption was sampled at a 6 s, 1 s, and 10 s intervals for UK-DALE, REDD, and Plegma, respectively. For the purpose of this study, the data for UK-DALE and REDD datasets were resampled to 8 s resolution, while Plegma kept the original resolution of 10 s. Detailed dataset characteristics and selection of houses for training data is described in Table 4. We evaluate our approach by training appliance-level models for Dishwasher, Washing Machine, Microwave, Refrigerator, AC, and Boiler appliances. The models were tested on unseen houses excluded from the training set. In UK-DALE, houses 1, 3, 4, and 5 were used for training and house 2 for testing, while in REDD houses 2, 3, 4, 5, and 6 were used in the training set while house 1 was preserved for model evaluation. For Plegma dataset, all houses except 10 and 2 were used for training, while house 10 was used for validation, and house 2 for testing. Aggregate measurements were normalized using z-normalization $z = \frac{x-\mu}{\sigma}$, where x represents the recorded power measurement (in Watts), μ mean power value in the whole training dataset, while σ represents the standard deviation of the values in the training dataset.

4.2. Model architectures and training

To enhance the generalizability and robustness of our proposed framework, we base our evaluation on three distinct NILM model architectures: a convolutional network (Zhang et al., 2018), a recurrent network (Rafiq et al., 2021), and a WaveNet neural network (Harell et al., 2019) network, as illustrated in Fig. 2. Reccurent architectures process sequential data by iterating through the input elements and maintaining a hidden state. This allows them to capture temporal dependencies in the data. However, RNNs often struggle with long-term dependencies due to the vanishing gradient problem. More advanced variants like GRU networks address this issue by introducing gating mechanisms to better control information flow. CNNs, on the other hand, use convolutional layers that apply filters across the input data, typically in a sliding window fashion. This allows them to detect local patterns regardless of their position in the input. CNNs also often include pooling layers to reduce dimensionality and increase robustness to small translations. Lastly, WaveNet networks use dilated causal convolutions to create very large receptive fields to model long-range temporal dependencies in time series data while maintaining computational efficiency. For further details on selected NILM architectures readers are referred to Zhang et al. (2018), Rafiq et al. (2021), and Harell et al. (2019).

We selected model hyperparameters based on optimal validation performance across all considered parameters. All models are trained using Adam optimizer with a predefined learning rate of 0.001, and a batch size of 64 samples. Input window lengths of the three selected networks are kept the same as in the original work. The training of the prior model maintains the same learning rate as the initial baseline model, with the β parameter chosen through a grid search of values on a logarithmic scale ranging from $[10^{-10}, 10^{0}]$. To thoroughly explore the trade-offs between the smoothness and low complexity priors, we implemented an iterative optimization process. This approach involves systematically varying the influence of both priors through their respective β hyperparameters and analyzing their combined impact on model performance. Our observations reveal that the smoothness prior, while enhancing gradient stability, may occasionally conflict with the identification of sharp feature boundaries. In contrast, the low-complexity prior promotes concise explanations but risks oversimplifying complex data relationships. In particular, we found that the optimal balance between these priors often yields the best ROTR scores, although this equilibrium can vary significantly depending on the specific NILM task and the characteristics of the dataset.

4.3. Computational complexity

For the purpose of performing the experiments, a PC with the following specifications is used: Intel(R) Core(TM) i9-10980XE CPU @ 3.00 GHz, 258 GB RAM, and two NVIDIA GeForce RTX 3080 GPUs. In analyzing the computational complexity of our framework across different architectures, we observed that the incorporation of priors affects the training speed across all architectures. The recurrent architecture experiences the most significant impact, with training time increasing by 50% when using priors compared to the baseline model without priors. The convolutional architecture shows a 39% increase in training time, while the dilated causal network exhibits a 32% increase. These variations in computational overhead can be attributed to the additional calculations required for prior computation and their interaction with each architecture's unique structure. The recurrent network's higher computational cost may be due to the complex interplay between its sequential processing nature and the prior calculations. The convolutional architecture's moderate increase likely stems from the integration of priors with its feature extraction process, while the dilated causal network's smaller overhead might result from its inherent ability to handle temporal dependencies more efficiently when combined with priors. These findings underscore the trade-off between improved explainability and increased computational cost, highlighting the importance of considering both model architecture and prior implementation when optimizing for NILM applications, especially in scenarios where training time and resources are limited.

4.4. Evaluation metrics

Finding the optimal model requires an objective metric that quantifies the predictive performance. Since the models used in this work are primarily developed for a regression task, we quantify the regression performance using the Mean Absolute Error (MAE) measure. MAE between the true (E_i) and predicted $(\hat{E}_i)$ consumed energy of the appliance of interest is calculated as:

$$MAE = \frac{1}{T} \cdot \sum_{i=1}^{T} |\hat{E}_i - E_i|. \tag{7}$$

Table 1
Comparison of XNILMBoost performance for REDD.

Appliance	AI Model	MAE	F1-Score
Dishwasher	GRU	24.20	0.427
	GRU + Prior	**20.74**	**0.538**
	CNN	19.55	0.696
	CNN + Prior	**17.23**	**0.775**
	WaveNet	24.91	0.408
	WaveNet + Prior	**24.42**	**0.477**
Microwave	GRU	**16.87**	**0.538**
	GRU + Prior	17.11	0.523
	CNN	19.18	0.362
	CNN + Prior	**17.12**	**0.516**
	WaveNet	**16.54**	0.603
	WaveNet + Prior	16.97	**0.619**
Refrigerator	GRU	**33.35**	0.805
	GRU + Prior	**33.35**	**0.806**
	CNN	28.47	0.84
	CNN + Prior	**27.53**	**0.843**
	WaveNet	38.31	0.758
	WaveNet + Prior	**36.69**	**0.765**

Table 2
Comparison of XNILMBoost performance for UKDALE.

Appliance	AI Model	MAE	F1-Score
Washing Machine	GRU	6.39	0.77
	GRU + Prior	**5.68**	**0.78**
	CNN	**6.82**	**0.63**
	CNN + Prior	8.55	0.62
	WaveNet	7.00	0.65
	WaveNet + Prior	**6.61**	**0.69**
Dishwasher	GRU	30.78	0.67
	GRU + Prior	**25.15**	**0.73**
	CNN	35.4	0.7
	CNN + Prior	**34**	**0.74**
	WaveNet	30.38	0.66
	WaveNet + Prior	**30.12**	0.68
Microwave	GRU	6.63	0.18
	GRU + Prior	**6.38**	**0.28**
	CNN	5.85	0.51
	CNN + Prior	**5.30**	**0.63**
	WaveNet	6.36	0.44
	WaveNet + Prior	6.36	**0.45**

Table 3
Comparison of XNILMBoost performance for Plegma Dataset.

Appliance	AI Model	MAE	F1-Score
AC	GRU	38.41	0.773
	GRU + Prior	**38.20**	**0.792**
	CNN	42.49	0.745
	CNN + Prior	**39.64**	**0.772**
	WaveNet	58.15	0.662
	WaveNet + Prior	**53.68**	**0.699**
Boiler	GRU	**4.42**	**0.970**
	GRU + Prior	7.48	0.929
	CNN	4.44	**0.939**
	CNN + Prior	**4.04**	0.929
	WaveNet	**18.27**	0.837
	WaveNet + Prior	18.98	**0.867**
Washing Machine	GRU	2.63	0.543
	GRU + Prior	**1.96**	**0.590**
	CNN	3.23	0.481
	CNN + Prior	**2.97**	**0.560**
	WaveNet	3.42	0.586
	WaveNet + Prior	**3.17**	**0.620**

Whilst MAE is the most common measure for evaluating regression or disaggregation performance (Huber et al., 2021), the F1-score measure is typically used in the NILM literature to evaluate the classification performance (Angelis et al., 2022). To generate events from the regression output, we apply a threshold, as explained in Section 3.1. Specifically, as a way of capturing the classification performance, we convert the regression output to a step function and calculate the F_1 score as:

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}, \tag{8}$$

where TP stands for True Positives, FP for False Positives, and FN for False Negatives.

In terms of explainability evaluation, we quantify the relationship between attribution quality and predictive performance using a faithfulness algorithm defined in Batic et al. (2023a), where the performance degradation after iterative removal of most important features is measured for both the regression and classification scenarios.

To measure the (in)stability of assigned attributions with slight modifications of the input signal, we use a Lipschit metric defined in Batic et al. (2023a). Given an explanation function $m(\cdot)$ and input aggregate signal x, we expose the signal to zero-mean Gaussian noise with standard deviation σ to create modified aggregate signal, $\hat{x}$. We define local Lipschitz constant estimate as Alvarez-Melis and Jaakkola (2018):

$$\hat{L} = \frac{\|m(\theta, x) - m(\theta, \hat{x})\|}{\|x - \hat{x}\| + \mu}, \tag{9}$$

where μ represents a small value added for numerical stability ($\mu = 1e^{-6}$). For validity, the procedure is repeated n times. Methods with low Lipschitz value scores display a characteristic of being stable under the presence of noise and should be favored.

Lastly, to measure the overall ease of understanding the produced explanation, an effective complexity measure is used (Batic et al., 2023a). To quantify the complexity of explanation in the context of NILM, we define the "effective complexity" measure as a combination of the attribution conciseness measure – Gini index, and the dataset complexity measure – NAR (Makonin and Popowich, 2015):

$$EC^{(i)} = \frac{Gini}{1 - NAR^{(i)}}. \tag{10}$$

4.5. Experimental results and discussion

4.5.1. Does training for better explanations lead to improved predictive performance?

The first experimental analysis is designed to examine if explain ability-informed training can lead to improved model performance, instead of the often argued conjecture of trading-off between explainablity and accuracy (Commission and Directorate-General for Communications Networks, and Technology, 2019). As can be seen in Tables 1 and 2, training with attribution priors can generally lead to significant regression and classification performance improvement compared to the case when no priors are used. Note that, explainability-informed training leads to varying degrees of improvement across different architectures and appliances. To better illustrate this, Fig. 3 showcases relative change in F1 and MAE score after training with the proposed method. For the UK-DALE scenario, applying an attribution prior to a GRU architecture leads to a slight regression performance improvement for Microwave appliance. However, regression performance improvement in appliances with long and sparse activations (Washing Machine and Dishwasher) is significant, reaching over 15%. On the other hand CNN, whilst significantly improving results for the Microwave, underperformed for the case of Washing Machine, where the MAE value increased, suggesting a nuanced relationship between model architecture, attribution priors, and appliance characteristics.

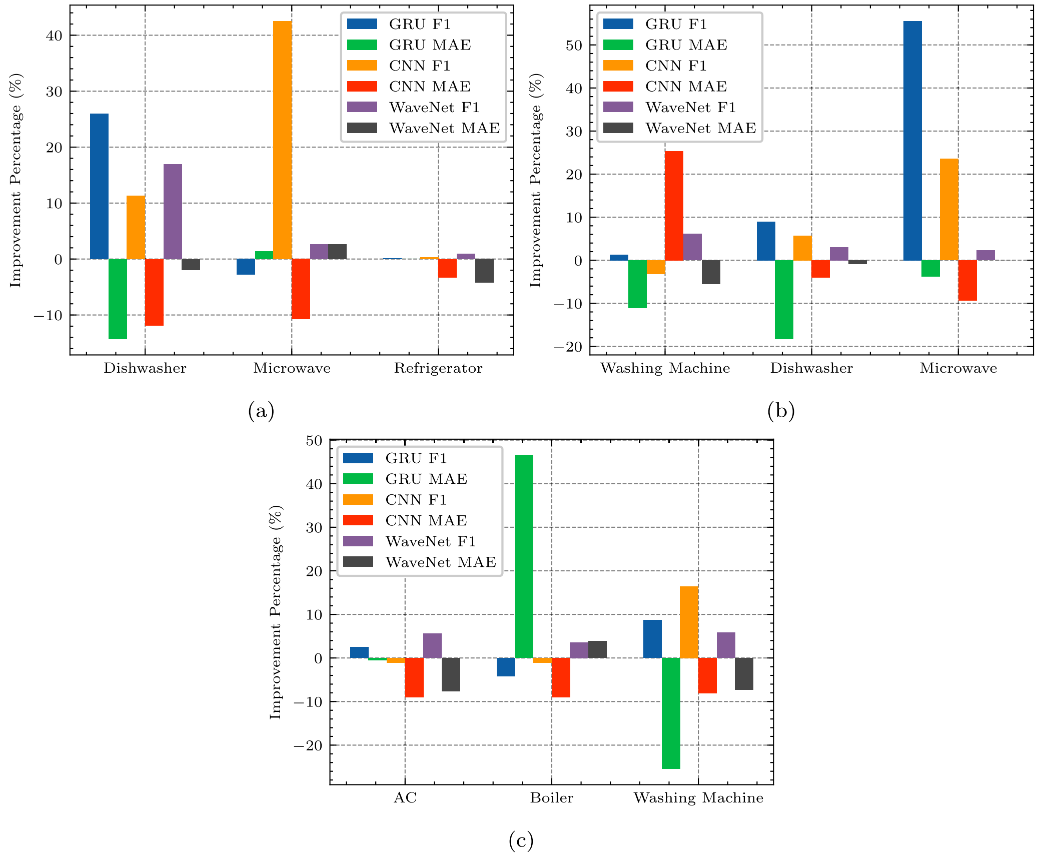


Fig. 3. Comparison of relative F1 and MAE performance improvement after explainability-informed training for GRU, CNN, and WaveNet architectures for (a) REDD dataset (b) UK-DALE dataset, and (c) Plegma dataset.

Generally, we observe that the improvement in one predictive metric follows the improvement in other, indicating that the trained models produce more robust predictions in both classification and regression domain. However, there are also cases where F1 improvement is drastically higher than MAE improvement, as is the case for Microwave trained with GRU model. This phenomenon is probably due to poor initial classification performance of the Microwave GRU model, leading to a higher relative increase. The effects of explainability-informed training are very similar for the REDD dataset. We note great improvement for the case of Dishwasher appliance, where F1 improvement surpassed 25%. On the other hand, Refrigerator appliance showed minimal relative improvement over all models, which can be explained by excellent initial predictive performance of the baseline models. Important fining is that WaveNet architecture only led to slight improvements in F1 and MAE scores, except for the case of Dishwasher appliance in REDD dataset. Possible cause for such behavior is added complexity of introducing explainability due to large number of dilated causal convolutions. Analyzing the Plegma dataset results (Table 3), we observe trends in performance improvement with explainability-informed training similar to REDD and UK-DALE, but with some notable differences. For the AC appliance, all models show improvements with attribution priors, with WaveNet demonstrating the largest relative gains. The Boiler appliance presents mixed results — GRU and CNN models without priors perform better in terms of MAE, though CNN+Prior achieves the best overall MAE while maintaining a high F1-Score. WaveNet shows significant improvement with priors for the Boiler. For the Washing Machine, all models consistently benefit from attribution priors in both MAE and F1-Score. Notably, WaveNet models show consistent improvement with attribution priors across all appliances in the Plegma dataset, contrasting with the minimal improvements observed in REDD and UK-DALE.

The impact of attribution priors on model training can be attributed to multiple interconnected mechanisms. When attribution priors are introduced, they appear to synergize differently with various model architectures (GRU, CNN, WaveNet), potentially enhancing the inherent ability of each model to capture appliance-specific behavioral patterns. Attribution priors serve a dual purpose: they act as an effective regularization mechanism that guards against overfitting, while simultaneously strengthening the model's capacity to generalize from training data. This relationship is particularly evident in the WaveNet architecture, where the inherent complexity-performance trade-off suggests that attribution priors help strike an optimal balance, resulting in more robust performance on new data, while improving explainability. Since WaveNet processes complete sequences rather than individual samples, this behavior could indicate that, when the proposed approach is utilized, optimal trade-off might be achieved with either a larger model input window or lower sampling rates. Each appliance exhibits distinctive operational signatures and power consumption patterns, which fundamentally affect how much improvement can be achieved across different devices. Thus, the varying degrees of improvement might be influenced by the baseline performance of each model-appliance combination, with initially poor-performing models showing more dramatic improvements. For example, for the REDD dataset, largest improvements in F1 score are observed for the GRU-Dishwasher pair (26%) and CNN-Microwave pair (42.5%). However, they also hold the lowest baseline F1 scores — 0.427 and 0.362, respectively. A similar trend is seen in the case of UKDALE and Plegma datasets, where the highest improvement in F1 performance is held by UKDALE-GRU-Microwave (55%) and Plegma-CNN-Washing Machine (16.1%) — where both cases correspond to poor performing baseline models which were improved.

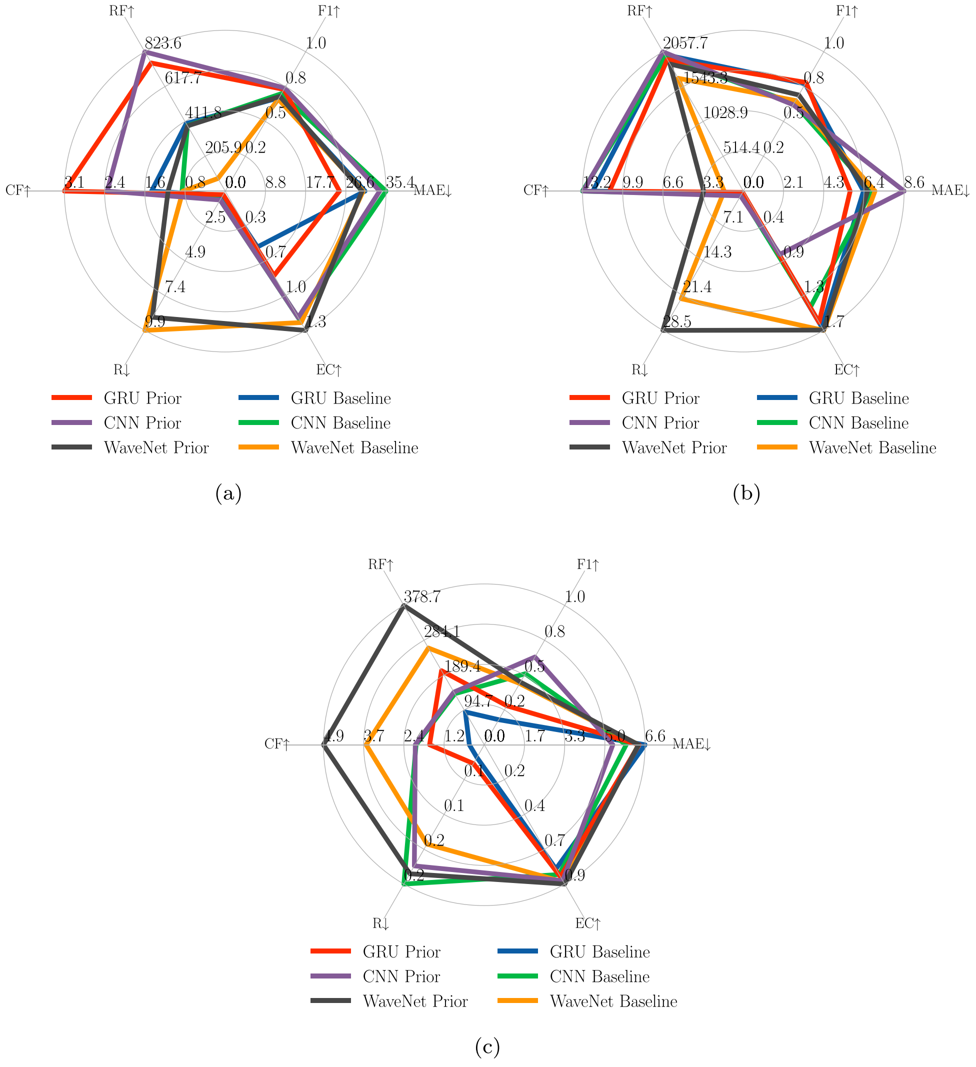


Fig. 4. Performance evaluation of the proposed XNILMBoost method for training of (a) UK-DALE Dishwasher, (b) UK-DALE Washing Machine, (c) UK-DALE Microwave. The radar plot axes are scaled based on the maximum values of the respective category. The arrows indicate if higher or lower value is better.

Table 4
Appliance characteristics for UK-DALE and REDD datasets.

Dataset	Appliance	Training houses	On threshold [W]	Min On [s]	Min Off [s]
UK-DALE	Washing Machine	1, 3, 4, 5	20	1800	150
	Dishwasher	1, 3, 4, 5	10	1800	1500
	Microwave	2, 3, 5	200	12	30
REDD	Dishwasher	2, 3, 4, 5, 6	10	1800	1500
	Microwave	2, 3, 5	200	12	30
	Refrigerator	2, 3, 5, 6	50	60	15
Plegma	AC	1, 3, 4, 5, 7, 8, 11, 12, 13	50	100	2100
	Boiler	1, 3, 4, 5, 6, 7, 9, 11, 12, 13	50	30	300
	Washing Machine	1, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13	50	30	112

4.5.2. Does training for better explanations lead to improved explainability performance?

Next, we evaluate how the models are affected by measuring their explainability performance. Applying the proposed explainability-informed training algorithm, we report several findings averaged across the selected explainability methods. Tables 5–13 showcase the performance across various NILM-specific explainability metrics. Focusing on the results of IntegratedGradients (IG) method, the average C. Faithfulness can be increased by 25.89% in REDD and by 80.61% in UK-DALE dataset. By comparing the obtained results, we observe that higher improvement in the UK-DALE dataset is largely due to poor baseline performance, i.e., in cases where the baseline metric indicates sub-optimal Faithfulness, the proposed explainability-informed training leads to largest improvements, suggesting that our training method particularly benefits models struggling in explainability. Notably, improvements in C. Faithfulness often mirrored those in R. Faithfulness,

Table 5
Comparison of XNILMBoost explainability performance improvement for CNN trained on REDD dataset.

Appliance	Model	R. Faithf.↑	C. Faith.↑	Robustness ↓	Eff. Complexity↑
Dishwasher	GradCAM (Baseline)	**2370.588**	3.122	**4.838** ± 0.782	0.900
	GradCAM (Prior)	2331.422	**3.441**	5.489 ± 1.407	**1.105**
	GradCAM++ (Baseline)	1143.810	1.565	**3.722** ± 0.662	0.570
	GradCAM++ (Prior)	**2154.310**	**2.614**	4.483 ± 0.881	**0.741**
	IG (Baseline)	2367.930	4.594	**1.267** ± 0.284	1.039
	IG (Prior)	**3231.460**	**4.877**	1.287 ± 0.319	**1.067**
	SG (Baseline)	2166.520	**3.830**	2.235 ± 0.442	0.745
	SG (Prior)	**2343.960**	2.531	**1.852** ± 0.287	**0.869**
Microwave	GradCAM (Baseline)	**97.923**	0.474	**0.190** ± 0.285	**0.477**
	GradCAM (Prior)	81.905	**0.761**	0.203 ± **0.181**	0.405
	GradCAM++ (Baseline)	**134.440**	**0.689**	**0.192** ± 0.123	**0.502**
	GradCAM++ (Prior)	74.320	0.605	0.279 ± **0.147**	0.429
	IG (Baseline)	86.190	**1.468**	**0.190** ± 0.106	**0.739**
	IG (Prior)	**238.280**	1.409	0.253 ± 0.131	0.723
	SG (Baseline)	106.320	**1.534**	**0.224** ± 0.167	**0.735**
	SG (Prior)	**242.720**	1.263	0.298 ± 0.172	0.687
Refrigerator	GradCAM (Baseline)	45.915	0.350	**1.559** ± 0.886	0.558
	GradCAM (Prior)	**155.915**	**0.684**	1.684 ± 0.939	**0.999**
	GradCAM++ (Baseline)	30.081	0.281	**1.418** ± 0.749	0.521
	GradCAM++ (Prior)	**140.081**	**0.572**	1.761 ± 1.302	**0.616**
	IG (Baseline)	147.179	4.111	**1.147** ± 0.275	**1.210**
	IG (Prior)	**386.144**	**4.445**	1.400 ± 0.275	1.206
	SG (Baseline)	173.086	2.454	**1.105** ± 0.377	0.920
	SG (Prior)	**283.086**	**2.788**	1.330 ± 0.721	**1.373**

Table 6
Comparison of explainability performance for WaveNet trained on REDD dataset.

Appliance	Model	R. Faithf. ↑	C. Faith. ↑	Robustness ↓	Eff. Complexity ↑
Dishwasher	GradCAM (Baseline)	1251.456	1.636	4.417 ± 2.830	0.161
	GradCAM (Prior)	**1285.666**	**2.056**	**3.297** ± 2.707	**0.411**
	GradCAM++ (Baseline)	1644.69	**3.185**	15.091 ± **3.709**	0.687
	GradCAM++ (Prior)	**1695.69**	2.445	**14.851** ± 3.709	**0.872**
	IG (Baseline)	403.440	4.311	**9.188** ± 2.567	0.982
	IG (Prior)	**638.560**	**4.401**	9.508 ± 2.027	**1.322**
	SG (Baseline)	1724.060	3.068	1.040 ± 0.037	1.574
	SG (Prior)	**1856.620**	**3.189**	**1.030** ± 0.061	**1.804**
Microwave	GradCAM (Baseline)	340.729	0.646	3.203 ± **2.830**	1.391
	GradCAM (Prior)	**572.829**	**1.066**	**2.976** ± 2.707	**2.056**
	GradCAM++ (Baseline)	599.720	**2.195**	**13.876** ± **1.709**	1.117
	GradCAM++ (Prior)	**982.850**	1.455	14.728 ± 3.709	**2.092**
	IG (Baseline)	280.440	3.321	8.860 ± **2.567**	1.512
	IG (Prior)	**638.560**	**3.411**	**8.278** ± 2.027	**2.362**
	SG (Baseline)	850.790	**2.078**	**6.205** ± **0.037**	1.814
	SG (Prior)	**1143.780**	2.199	6.030 ± 0.061	**1.912**
Refrigerator	GradCAM (Baseline)	52.900	0.835	5.632 ± 1.600	1.843
	GradCAM (Prior)	**75.828**	**1.496**	**2.057** ± 1.277	**1.951**
	GradCAM++ (Baseline)	446.130	**2.384**	**6.966** ± 2.839	**2.447**
	GradCAM++ (Prior)	**485.850**	1.885	13.31 ± 1.586	2.092
	IG (Baseline)	403.440	3.510	**9.066** ± 0.607	2.082
	IG (Prior)	**638.560**	**3.841**	9.496 ± 1.273	**2.552**
	SG (Baseline)	525.500	2.267	5.164 ± 0.507	1.927
	SG (Prior)	**646.780**	**2.529**	**4.275** ± 0.291	**2.827**

which can be explained by the fact that artifacts in the predicted appliance signature are no longer being produced due to improved gradient smoothness and explanation complexity after explainability-informed training. Observing the results, we corroborate previous findings that some explainability methods lead to unstable performance (Batic et al., 2023b; Alvarez-Melis and Jaakkola, 2018; Ancona et al., 2018). This is particularly evident in the case of GradCAM, while other methods provide more stable results. Furthermore, IG provides an overall satisfactory faithfulness performance across most appliances and architectures, reaffirming the previous hypothesis that that a zero signal is an appropriate choice for the baseline value for NILM data (Batic et al., 2023b). In terms of Robustness metric, we observe

Table 7
Comparison of explainability performance for GRU trained on REDD dataset.

Appliance	Model	R. Faithf. ↑	C. Faith. ↑	Robustness ↓	Eff. Complexity ↑
Dishwasher	GradCAM (Baseline)	223.006	0.883	0.39 ± 0.174	**1.348**
	GradCAM (Prior)	**696.664**	**1.753**	**0.048** ± 0.024	1.025
	GradCAM++ (Baseline)	220.897	**3.745**	0.669 ± 0.758	0.956
	GradCAM++ (Prior)	**646.830**	1.812	**0.131** ± 0.041	**1.321**
	IG (Baseline)	139.350	**4.884**	0.544 ± 0.283	1.311
	IG (Prior)	**762.740**	1.633	**0.132** ± 0.163	**1.365**
	SG (Baseline)	63.089	**4.012**	0.381 ± 0.231	**1.294**
	SG (Prior)	**727.400**	1.818	**0.062** ± 0.058	1.043
Microwave	GradCAM (Baseline)	33.042	0.372	0.095 ± 0.071	**0.570**
	GradCAM (Prior)	**60.477**	**0.622**	**0.077** ± 0.372	0.513
	GradCAM++ (Baseline)	**136.350**	**1.276**	**0.076** ± 0.068	**1.032**
	GradCAM++ (Prior)	73.660	0.366	0.126 ± 0.624	0.861
	IG (Baseline)	74.100	0.884	**0.011** ± 0.660	0.910
	IG (Prior)	**215.690**	**2.280**	0.064 ± 0.057	**1.112**
	SG (Baseline)	**211.100**	**2.156**	**0.054** ± 0.035	**1.127**
	SG (Prior)	166.050	1.814	0.033 ± 0.802	1.106
Refrigerator	GradCAM (Baseline)	14.373	**0.243**	**0.402** ± 0.216	**0.995**
	GradCAM (Prior)	**18.811**	0.040	0.478 ± 0.295	0.896
	GradCAM++ (Baseline)	**38.778**	**0.455**	1.356 ± 0.85	0.695
	GradCAM++ (Prior)	24.392	0.191	**1.351** ± 0.73	**0.825**
	IG (Baseline)	12.732	0.137	**1.895** ± 1.113	**1.039**
	IG (Prior)	**29.447**	**0.250**	1.918 ± 1.108	0.843
	SG (Baseline)	20.765	0.240	**0.386** ± 0.193	**0.777**
	SG (Prior)	**60.503**	**0.474**	0.414 ± 0.177	0.759

Table 8
Comparison of XNILMBoost explainability performance improvement for CNN trained on UK-DALE dataset.

Appliance	Model	R. Faithf. ↑	C. Faith. ↑	Robustness ↓	Eff. Complexity ↑
Dishwasher	GradCAM (Baseline)	**122.465**	0.301	1.547 ± 0.825	1.080
	GradCAM (Prior)	38.162	**0.399**	**0.931** ± 0.255	**1.353**
	GradCAM++ (Baseline)	62.629	0.102	1.740 ± 0.800	0.556
	GradCAM++ (Prior)	**96.980**	**0.799**	**1.223** ± 0.474	**0.871**
	IG (Baseline)	386.797	0.845	**0.623** ± 0.238	**1.200**
	IG (Prior)	**823.590**	**2.309**	0.627 ± 0.190	1.191
	SG (Baseline)	425.304	0.783	**0.364** ± 0.154	**1.082**
	SG (Prior)	**672.290**	**1.754**	0.441 ± 0.141	1.074
Washing Machine	GradCAM (Baseline)	1969.986	13.165	**1.734** ± 0.822	**1.616**
	GradCAM (Prior)	**1987.535**	**13.191**	3.046 ± 1.076	1.066
	GradCAM++ (Baseline)	1971.088	13.231	**4.067** ± 1.740	**0.954**
	GradCAM++ (Prior)	**2095.426**	**13.284**	4.211 ± 1.348	0.824
	IG (Baseline)	1987.030	**13.236**	**0.811** ± 0.271	**1.428**
	IG (Prior)	**2057.740**	13.174	0.978 ± 0.320	0.778
	SG (Baseline)	1943.557	**13.167**	**0.580** ± 0.305	**1.034**
	SG (Prior)	**1992.919**	13.117	0.934 ± 0.513	0.820
Microwave	GradCAM (Baseline)	134.827	2.021	0.223 ± 0.165	0.401
	GradCAM (Prior)	**142.935**	**2.058**	**0.193** ± 0.158	**0.507**
	GradCAM++ (Baseline)	138.070	2.044	**0.352** ± 0.180	0.355
	GradCAM++ (Prior)	**146.050**	**2.071**	0.396 ± 0.265	**0.383**
	IG (Baseline)	138.700	2.069	0.230 ± 0.119	0.831
	IG (Prior)	**143.090**	**2.094**	**0.200** ± 0.109	**0.870**
	SG (Baseline)	129.650	2.004	0.193 ± 0.080	0.839
	SG (Prior)	**143.460**	**2.087**	**0.176** ± 0.105	**0.866**

that WaveNet models lead to highest relative decrease of 16.64%. However, even with a significant improvement, WaveNet models still exhibit poor robustness performance, possibly due to their architectural design that is based on causal, dilated convolutional layers, which prevents robust explanations. In the case of CNNs, we observe that Robustness improvements correspond to lower MAE and increased F1 scores, as shown in Microwave model for UK-DALE dataset. Eff. Complexity has achieved highest improvement for the REDD dataset, where the relative increase achieves 89.26%, with WaveNet showing the highest relative and absolute increases. Additionally, we observe a link between Faithfulness improvement and Eff. Complexity improvement, in particular in cases of long running appliances such as Diswhasher trained on GRU with UK-DALE data. This finding suggests that the explainability metrics are interdependent, and that improved

Table 9
Comparison of explainability performance for WaveNet trained on UK-DALE dataset.

Appliance	Model	R. Faithf. ↑	C. Faith. ↑	Robustness ↓	Eff. Complexity ↑
Dishwasher	GradCAM (Baseline)	1044.074	2.314	88.782 ± 47.135	0.897
	GradCAM (Prior)	**1158.106**	**3.406**	**44.120** ± 12.400	**0.913**
	GradCAM++ (Baseline)	77.794	0.270	12.967 ± 8.725	1.237
	GradCAM++ (Prior)	**490.15**	**1.102**	**3.769** ± 1.823	**1.435**
	IG (Baseline)	74.801	0.829	9.854 ± 3.316	1.238
	IG (Prior)	**385.56**	**1.118**	**8.925** ± 2.350	**1.312**
	SG (Baseline)	661.983	1.652	**0.179** ± 0.213	1.472
	SG (Prior)	**1279.98**	**3.314**	0.260 ± 0.194	**1.513**
Washing Machine	GradCAM (Baseline)	974.101	0.835	204.055 ± 57.858	1.234
	GradCAM (Prior)	**1946.377**	**2.367**	**118.583** ± 27.907	**1.300**
	GradCAM++ (Baseline)	**1212.77**	1.242	**22.100** ± 5.100	1.330
	GradCAM++ (Prior)	1042.89	**1.529**	28.685 ± 7.333	**1.429**
	IG (Baseline)	1666.70	1.577	**22.030** ± 6.618	**1.713**
	IG (Prior)	**1878.02**	**3.261**	28.524 ± 11.33	1.709
	SG (Baseline)	**1108.20**	**1.200**	**0.029** ± 0.050	1.913
	SG (Prior)	482.810	0.740	0.277 ± 0.122	**1.941**
Microwave	GradCAM (Baseline)	56.567	**0.504**	21.416 ± 6.017	**0.110**
	GradCAM (Prior)	**109.918**	0.486	**19.519** ± 6.415	**0.110**
	GradCAM++ (Baseline)	68.626	**0.567**	**0.256** ± 0.644	0.068
	GradCAM++ (Prior)	**75.018**	0.553	0.651 ± 1.465	**0.087**
	IG (Baseline)	263.044	3.584	**0.164** ± 0.744	0.878
	IG (Prior)	**378.743**	**4.880**	0.213 ± 0.905	**0.892**
	SG (Baseline)	83.429	0.666	0.241 ± 0.100	**0.940**
	SG (Prior)	**86.459**	**0.678**	**0.144** ± 0.144	0.542

Table 10
Comparison of explainability performance for GRU trained on UK-DALE dataset.

Appliance	Model	R. Faithf. ↑	C. Faith. ↑	Robustness ↓	Eff. Complexity ↑
Dishwasher	GradCAM (Baseline)	300.885	1.250	**0.346** ± 0.233	0.625
	GradCAM (Prior)	**354.444**	**1.273**	0.414 ± 0.329	**0.901**
	GradCAM++ (Baseline)	172.887	0.458	0.395 ± 0.662	0.449
	GradCAM++ (Prior)	**487.953**	**2.190**	**0.248** ± 0.631	**0.567**
	IG (Baseline)	399.699	1.441	0.298 ± 0.446	0.526
	IG (Prior)	**757.573**	**3.146**	**0.249** ± 0.304	**0.796**
	SG (Baseline)	436.021	2.005	**0.185** ± 0.198	1.090
	SG (Prior)	**788.257**	**2.980**	**0.185** ± 0.153	**1.162**
Washing Machine	GradCAM (Baseline)	2004.603	11.255	**0.487** ± 0.300	1.663
	GradCAM (Prior)	**2140.319**	**11.440**	0.53 ± 0.316	**1.669**
	GradCAM++ (Baseline)	**2362.02**	**12.391**	**0.96** ± 1.105	**1.642**
	GradCAM++ (Prior)	1960.83	10.782	1.036 ± 0.557	1.514
	IG (Baseline)	**2017.31**	**12.384**	0.426 ± 0.314	**1.674**
	IG (Prior)	1944.02	11.014	**0.256** ± 0.211	1.614
	SG (Baseline)	1080.61	5.342	**0.361** ± 0.233	**0.772**
	SG (Prior)	**1486.52**	**6.482**	0.466 ± 0.335	0.600
Microwave	GradCAM (Baseline)	**65.804**	**0.388**	0.115 ± 0.168	0.738
	GradCAM (Prior)	41.809	0.176	**0.082** ± 0.055	**0.761**
	GradCAM++ (Baseline)	20.458	0.018	**0.200** ± 0.171	0.618
	GradCAM++ (Prior)	**85.312**	**0.358**	0.382 ± 0.335	**0.767**
	IG (Baseline)	89.247	0.453	**0.021** ± 0.012	0.795
	IG (Prior)	**201.395**	**1.660**	0.031 ± 0.035	**0.845**
	SG (Baseline)	149.567	0.759	**0.018** ± 0.010	0.779
	SG (Prior)	**170.971**	**1.153**	0.025 ± 0.044	**0.794**

gradient smoothness and complexity leads to better overall explainability of the NILM system. The Plegma dataset results, as shown in Tables 11–13, further corroborate and extend the findings observed in the REDD and UK-DALE datasets, while also revealing some unique patterns. Across CNN, WaveNet, and GRU models, we see substantial improvements in both R. Faithfulness and C. Faithfulness for many appliances when using priors, particularly for the AC appliance. For instance, CNN models show significant gains in R. Faithfulness for AC and Washing Machine, while WaveNet models demonstrate even more pronounced improvements across all appliances. GRU models present a more mixed picture, with some appliances showing improvements and others slight decreases. Robustness generally improves with the use of priors across all architectures, although the magnitude of improvement varies. Overall, it can be concluded that the utilization of explainability-informed NILM mode training can lead to explainability improvement across various architectural approaches, which is validated through

Table 11
Comparison of explainability performance for CNN trained on Plegma dataset.

Appliance	Model	R. Faithf. ↑	C. Faith. ↑	Robustness ↓	Eff. Complexity ↑
AC	GradCAM (Baseline)	508.908	0.120	0.752 ± 0.402	**0.438**
	GradCAM (Prior)	**883.792**	**1.141**	**0.269** ± 0.210	0.370
	GradCAM++ (Baseline)	153.922	0.359	**0.937** ± 0.593	**0.836**
	GradCAM++ (Prior)	**540.328**	**0.816**	0.976 ± 0.800	0.791
	IG (Baseline)	1530.439	2.657	0.984 ± 0.460	0.936
	IG (Prior)	**2147.258**	**3.737**	**0.812** ± 0.327	**1.027**
	SG (Baseline)	918.024	0.766	**1.321** ± 0.638	**0.794**
	SG (Prior)	**1042.599**	**0.867**	1.685 ± 1.310	0.757
Boiler	GradCAM (Baseline)	**3197.939**	**0.695**	0.068 ± 0.125	**0.614**
	GradCAM (Prior)	956.829	0.103	**0.056** ± 0.035	0.613
	GradCAM++ (Baseline)	408.272	**0.097**	0.420 ± 0.354	0.321
	GradCAM++ (Prior)	**524.275**	0.090	**0.188** ± 0.276	**0.540**
	IG (Baseline)	3920.85	0.275	**0.098** ± 0.085	**0.970**
	IG (Prior)	**4404.764**	**0.640**	0.128 ± 0.101	0.931
	SG (Baseline)	**3561.111**	**0.608**	**0.079** ± 0.046	**0.905**
	SG (Prior)	3110.081	0.336	**0.079** ± 0.038	0.841
Washing Machine	GradCAM (Baseline)	58.216	0.180	1.240 ± 0.701	0.441
	GradCAM (Prior)	**152.773**	**0.301**	**0.971** ± 0.611	**0.469**
	GradCAM++ (Baseline)	84.957	0.139	1.627 ± 1.233	0.268
	GradCAM++ (Prior)	**313.451**	**0.312**	**1.430** ± 0.814	**0.356**
	IG (Baseline)	263.044	**0.307**	1.240 ± 0.489	0.727
	IG (Prior)	**282.065**	0.237	**1.218** ± 0.549	**0.824**
	SG (Baseline)	83.429	**0.666**	1.229 ± 0.692	0.385
	SG (Prior)	**226.415**	0.292	**1.021** ± 0.645	**0.405**

Table 12
Comparison of explainability performance for WaveNet trained on Plegma dataset.

Appliance	Model	R. Faithf. ↑	C. Faith. ↑	Robustness ↓	Eff. Complexity ↑
AC	GradCAM (Baseline)	385.709	0.923	133.234 ± 60.912	0.503
	GradCAM (Prior)	**823.906**	**1.535**	**130.22** ± 51.595	**0.541**
	GradCAM++ (Baseline)	463.939	2.003	47.312 ± 40.933	0.798
	GradCAM++ (Prior)	**955.981**	**2.493**	**20.821** ± 12.211	**1.170**
	IG (Baseline)	**1842.737**	3.145	26.331 ± 23.487	0.854
	IG (Prior)	1671.350	**3.941**	**13.091** ± 15.238	**0.991**
	SG (Baseline)	1115.153	**1.006**	**0.205** ± 0.127	**1.077**
	SG (Prior)	**1295.212**	0.707	0.222 ± 0.121	**1.077**
Boiler	GradCAM (Baseline)	3189.995	**2.384**	97.028 ± 46.832	**0.460**
	GradCAM (Prior)	**3243.791**	1.466	**58.910** ± 48.569	0.202
	GradCAM++ (Baseline)	1607.980	0.768	1.058 ± 3.527	0.200
	GradCAM++ (Prior)	**2405.809**	**0.907**	**1.019** ± 2.507	**0.414**
	IG (Baseline)	3329.980	3.209	6.900 ± 4.990	0.900
	IG (Prior)	**3348.476**	**3.304**	**4.607** ± 3.718	**0.914**
	SG (Baseline)	985.985	0.194	0.247 ± 0.090	0.976
	SG (Prior)	**1053.160**	**0.347**	**0.217** ± 0.100	**1.012**
Washing Machine	GradCAM (Baseline)	108.316	1.048	**0.680** ± 0.472	0.868
	GradCAM (Prior)	**375.831**	**1.392**	1.093 ± 0.714	**1.187**
	GradCAM++ (Baseline)	12.015	**0.652**	**0.618** ± 0.495	**1.125**
	GradCAM++ (Prior)	**89.671**	1.255	0.462 ± 0.268	1.297
	IG (Baseline)	127.221	0.575	**1.051** ± 0.451	1.210
	IG (Prior)	**393.303**	**1.005**	0.955 ± 0.345	**1.212**
	SG (Baseline)	**185.971**	1.117	**0.446** ± 0.234	**0.863**
	IG (Prior)	454.246	**1.482**	0.478 ± 0.421	0.703

Table 13
Comparison of explainability performance for GRU trained on Plegma dataset.

Appliance	Model	R. Faithf. ↑	C. Faith. ↑	Robustness ↓	Eff. Complexity ↑
AC	GradCAM (Baseline)	**2160.41**	**5.461**	**1.572** ± 1.238	0.633
	GradCAM (Prior)	1486.443	3.010	1.596 ± 1.142	**0.686**
	GradCAM++ (Baseline)	**2517.897**	**6.393**	1.139 ± 0.814	0.637
	GradCAM++ (Prior)	1925.591	5.707	**1.041** ± 0.640	**0.751**
	IG (Baseline)	**2046.526**	4.840	0.681 ± 0.520	0.730
	IG (Prior)	1818.143	**5.049**	**0.563** ± 0.403	**0.781**
	SG (Baseline)	1244.548	**4.323**	**0.454** ± 0.324	0.694
	SG (Prior)	**1489.35**	3.631	0.501 ± 0.496	**0.734**
Boiler	GradCAM (Baseline)	**3197.939**	0.695	7.028 ± 6.832	**0.614**
	GradCAM (Prior)	3189.995	**2.384**	**0.068** ± 0.125	0.460
	GradCAM++ (Baseline)	408.272	0.097	**0.420** ± 0.354	**0.321**
	GradCAM++ (Prior)	**1607.98**	**0.768**	1.058 ± 3.527	0.200
	IG (Baseline)	**3329.98**	**3.209**	6.900 ± 4.990	0.900
	IG (Prior)	3038.328	2.847	**4.607** ± 3.718	**0.914**
	SG (Baseline)	**1053.16**	**0.347**	0.247 ± 0.090	**1.012**
	SG (Prior)	985.985	0.194	**0.217** ± 0.100	0.976
Washing Machine	GradCAM (Baseline)	348.845	0.685	74.287 ± 26.595	0.733
	GradCAM (Prior)	**588.488**	**1.392**	**41.83** ± 25.994	**1.187**
	GradCAM++ (Baseline)	**532.696**	**1.163**	**4.799** ± 5.113	**1.628**
	GradCAM++ (Prior)	107.717	0.522	14.809 ± 9.998	1.484
	IG (Baseline)	183.145	1.551	**9.248** ± 7.377	1.365
	IG (Prior)	**473.062**	**1.762**	12.146 ± 10.751	**1.481**
	SG (Baseline)	85.301	0.423	**0.201** ± 0.091	**1.757**
	SG (Prior)	**142.333**	**0.823**	0.245 ± 0.091	1.694

relative improvement in individual explainability metrics, as can be further seen in Figs. 5, 4 and 6. However, while some models exhibit significant gains in both explainability and predictive performance, others show marginal improvements, underscoring the need for a more tailored approach in explainability-informed model training.

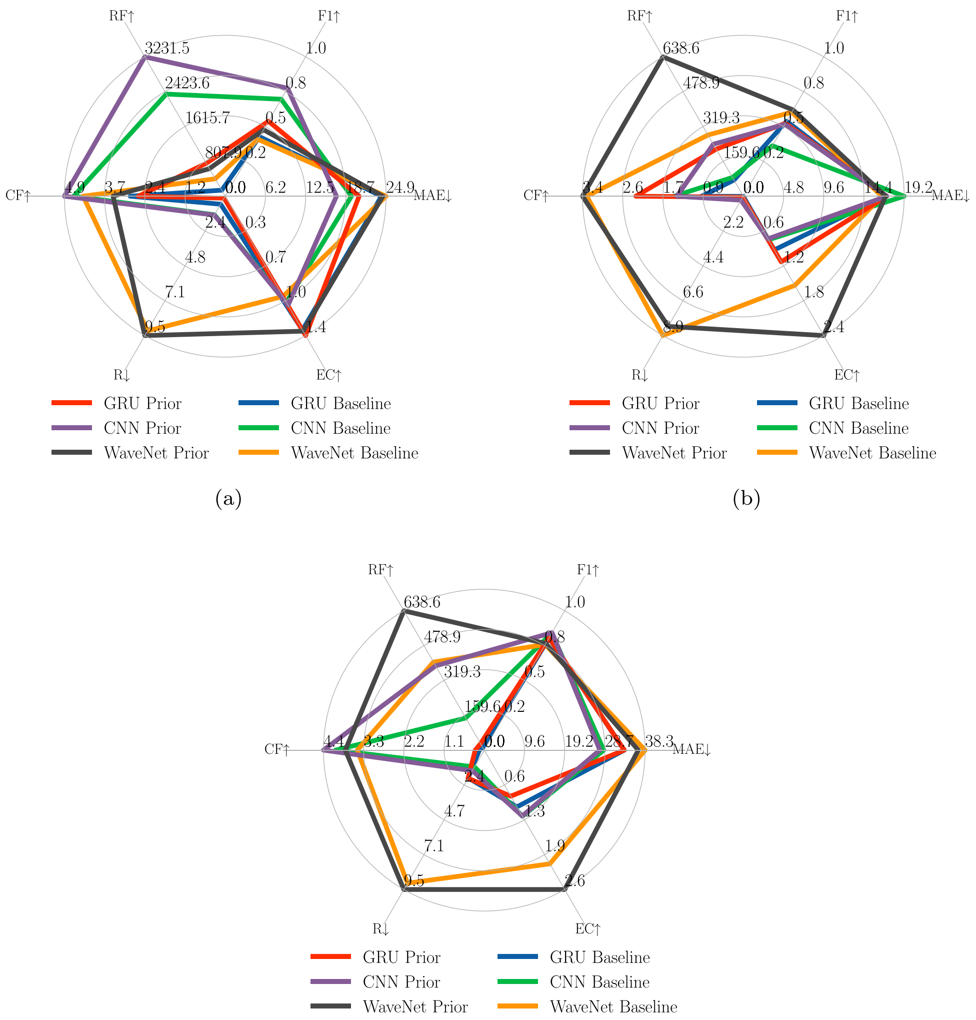
4.5.3. What is the relationship between the improved predictive performance and explainability?

Finally, it becomes evident that the trade-off between explainability and predictive performance, particularly within the context of attribution priors, presents a opportunity for evaluation of overall explainability-informed NILM system performance. To best illustrate the trade-off, we jointly visualize the explainability and predictive performance metrics in Figs. 5, 4 and 6. Figures are organized as radio plots where each axis represents one of the core metrics on the explainability-informed NILM system, while the arrows indicate if lower or higher values are favored. We observe that in the case of UK-DALE, GRU models that achieve higher C. and R. Faithfulness, generally lead to lower MAE values, as shown in the case of Dishwasher appliance, where 89.54% R. Faithfulness improvement corresponded with 18.29% decrease in MAE score. Similarly, GRU model trained on Dishwasher in REDD dataset when improved on the R. Faithfulness lead to improved F1 and lower MAE value. However, in the case of Microwave, improvement in R. and C. Faithfulness did not lead to improvement in predictive performance, albeit it did improve the Eff. Complexity result. This indicates that appliances with longer and sparser activations might benefit more from explainability-informed training. CNN model has also showed positive correlation between explainability improvement and predictive performance improvement. In cases of increased R. Faithfulness, CNNs tend to obtain better F1 and MAE score in both datasets, as shown in the case of Microwave for REDD dataset where 176.7% increase in R. Faithfulness corresponded with 42.54% increase in F1 score. In the case of WaveNet, we observe that increases in R. and C. Faithfulness, despite improving MAE and F1 scores, do not lead to dramatic improvements, suggesting that the complexity introduced through causal convolutions might be a limiting factor. However, for traditionally challenging-to-disaggregate appliances, such as the Washing Machine in the Plegma dataset (Fig. 6(c)), our proposed approach demonstrates simultaneous improvements in both explainability and predictive performance. The GRU model's results are particularly noteworthy, showing a significant decrease in MAE that correlates strongly with enhanced faithfulness metrics. This suggests that improvements in regression performance (MAE) may have a more substantial impact on explainability compared to classification performance gains (F1 score).

These findings emphasize the importance of carefully tailored approaches in machine learning applications, where model architectures and additional model inputs, such as priors, must be thoughtfully matched to specific tasks and datasets. The observed improvement in predictive and explainability performance validates our initial hypothesis that training explicitly for explainability can produce more robust and transparent NILM models. Furthermore, our proposed training procedure effectively quantifies the trade-off between model performance and explainability. More broadly, these results reveal a symbiotic relationship: more robust models naturally lead to better explainability, and conversely, enhanced explainability can contribute to increased model robustness (see Table 10).

5. Conclusions and future work

In this work, we proposed a framework for enhancement of state-of-the-art NILM models that takes into account characteristics of Trustworthy AI systems. The experimental results from our study highlight the significant impact of explainability-informed training on the performance of energy disaggregation models. This approach, which integrates attribution priors into the training process, demonstrates substantial improvements in both regression and classification performance. Additionally, we proposed an iterative optimization procedure that along with a novel explainability metric enables explainability-informed training of NILM models. Experimental results validate that our approach binds improved predictive performance with improved explainability results across various architectures and appliances. Three different research questions were addressed — First, we show that training for better explanations can lead to improved predictive performance of a NILM system and provide increased robustness; second,



(a)

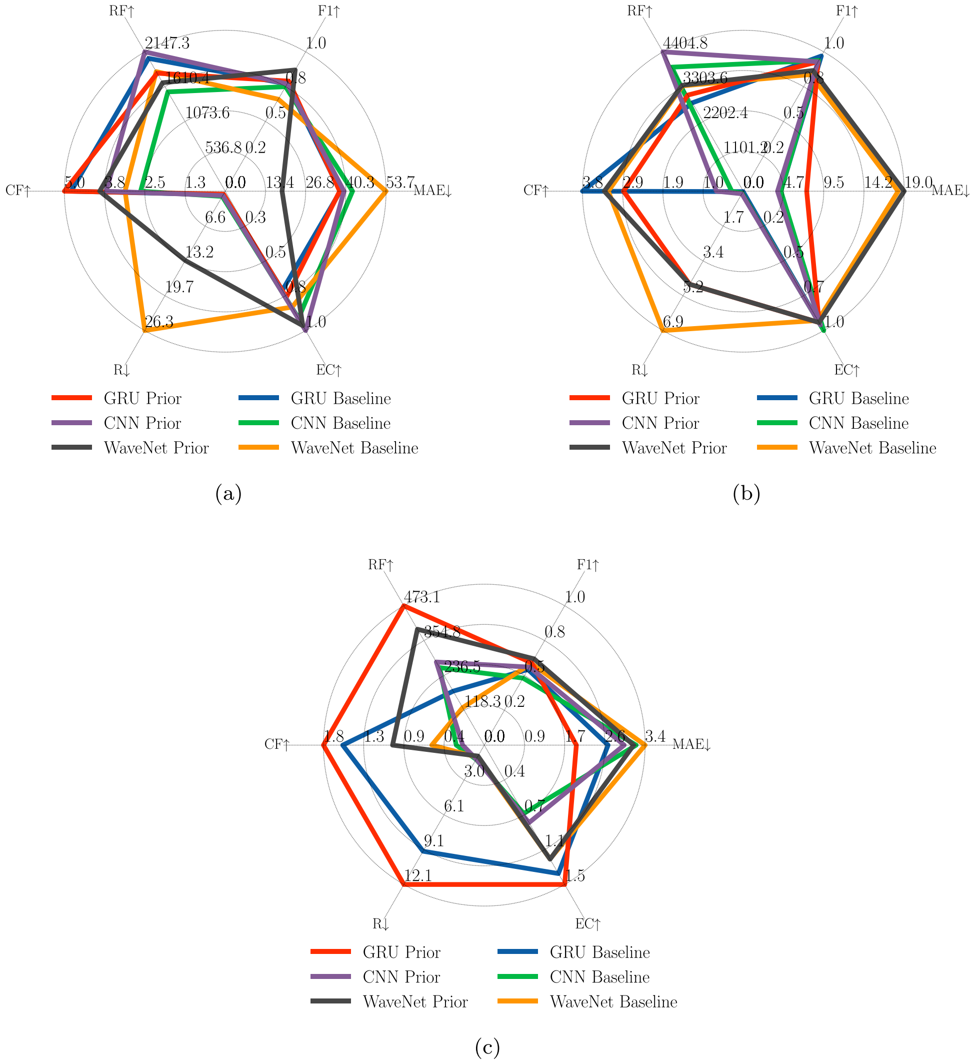
(b)

(c)

Fig. 5. Performance evaluation of the proposed XNILMBoost method for training of (a) REDD Microwave, (b) REDD Washing Machine, (c) REDD Refrigerator. The radar plot axes are scaled based on the maximum values of the respective category. The arrows indicate if higher or lower value is better.

we show that the proposed explainability-informed training can enhance the explainability performance of various state-of-the-art architectures across multiple explainability metrics; and third, we provide new insights into the relationship between the improved predictive performance and explainability for various NILM architectures. The proposed framework was applied across different architectural approaches, including convolutional (CNN), recurrent (GRU), and dilated causal (WaveNet) architectures. Worth noting is that although WaveNet models have achieved enhanced performance, the relative improvement achieved is much greater for CNN and GRU, suggesting that such architectures can benefit more from explainability-informed training. Various explainability methods were explored, including GradCAM, GradCAM++, IntegratedGradients, and SmoothGrad. Experimental results suggest that in the context of NILM, explainability methods that are design to deal with noise, such as IntegratedGradients and SmoothGrad, can generally obtain better ability to produce explanations that are faithful to the performance of the model, robust to slight changes of input, and more easily interpretable due to low complexity of outputs. Overall, the proposed methodology suggests that the incorporation of explainability considerations into the training process can substantially enhance the transparency of a model, as well as the ability to more accurately predict energy consumption of high-consumption appliances.

In future work, it is worth investigating how different emerging architectures, such as Transformer models or hybrid models, respond to the proposed training approach, and provide deeper insights into the generalizability and scalability of these techniques. Conducting user studies to understand how non-experts interpret the explanations provided by these models could also be beneficial in making NILM technologies more accessible and trustworthy. Future work could investigate methods to incorporate direct human feedback or domain knowledge into the attribution prior formulation through active learning and similar approaches, further strengthening the human agency and oversight principle. Integrating advanced NILM models into smart energy management systems could lead to more efficient and user-friendly energy consumption monitoring and management. Research in this direction could focus on creating holistic systems that leverage the strengths of explainable AI for better energy optimization and user engagement. Additionally, integration of XAI-informed NILM with other emerging technologies such as demand response programs and digital twin technology represents a promising opportunity in industrial energy management. Detailed and trustworthy load disaggregation could enhance the effectiveness of demand response strategies by identifying flexible loads that can be adjusted during peak demand periods without compromising critical operations. Simultaneously, digital twin technologies, which have showed success in healthcare (Feng et al., 2023a) and transportation sector (Feng et al., 2023b), could be integrated with



(a)

(b)

(c)

Fig. 6. Performance evaluation of the proposed XNILMBoost method for training of (a) Plegma AC, (b) Plegma Boiler, (c) Plegma Washing Machine. The radar plot axes are scaled based on the maximum values of the respective category. The arrows indicate if higher or lower value is better.

NILM outputs into digital twin models of industrial facilities to create a dynamic, real-time representation of energy usage and equipment performance, reducing operational costs, and improving equipment longevity.

CRediT authorship contribution statement

Djordje Batic: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Vladimir Stankovic:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Conceptualization. **Lina Stankovic:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955422.

Data availability

Data will be made available on request.

References

Alvarez-Melis, D., Jaakkola, T.S., 2018. On the robustness of interpretability methods. arXiv:1806.08049.

Alvarez-Melis, D., Jaakkola, T.S., 2018. Towards robust interpretability with self-explaining neural networks.

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

Ancona, M., Ceolini, E., Öztireli, C., Gross, M., 2018. Towards better understanding of gradient-based attribution methods for deep neural networks.

Angelis, G.F., et al., 2022. NILM applications: Literature review of learning approaches, recent developments and challenges. Energy Build. 261, 111951.

Armel, K.C., Gupta, A., Shrimali, G., Albert, A., 2013. Is disaggregation the holy grail of energy efficiency? The case of electricity. Energy Policy 52, 213–234.

Athanasoulias, S., Guasselli, F., Doulamis, N., Doulamis, A., Ipiotis, N., Katsari, A., Stankovic, L., Stankovic, V., 2024. The plegma dataset: Domestic appliance-level and aggregate electricity demand with metadata from Greece. Sci. Data 11 (1), 376.

Batic, D., Stankovic, V., Stankovic, L., 2023a. Towards transparent load disaggregation– a framework for quantitative evaluation of explainability using explainable AI. IEEE Trans. Consum. Electron..

Batic, D., Tanoni, G., Stankovic, L., Stankovic, V., Principi, E., 2023b. Improving knowledge distillation for non-intrusive load monitoring through explainability guided learning. In: 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2023, IEEE.

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

Chalasani, P., et al., 2020. Concise explanations of neural networks using adversarial training. In: Int. Conf. Machine Learn.. PMLR.

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 839–847.

Chen, K., Zhang, Y., Wang, Q., Hu, J., Fan, H., He, J., 2019. Scale-and context-aware convolutional non-intrusive load monitoring. IEEE Trans. Power Syst. 35 (3), 2362–2373.

Commission, E., Directorate-General for Communications Networks, and Technology, C., 2019. Ethics guidelines for trustworthy AI. Publications Office.

Feng, K., Ji, J., Zhang, Y., Ni, Q., Liu, Z., Beer, M., 2023a. Digital twin-driven intelligent assessment of gear surface degradation. Mech. Syst. Signal Process. 186, 109896.

Feng, K., Xu, Y., Wang, Y., Li, S., Jiang, Q., Sun, B., Zheng, J., Ni, Q., 2023b. Digital twin enabled domain adversarial graph networks for bearing fault diagnosis. IEEE Trans. Ind. Cyber-Physical Syst..

Harell, A., Makonin, S., Bajić, I.V., 2019. Wavenilm: A causal neural network for power disaggregation from the complex power signal. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 8335–8339.

Hossain, M., Madlool, N., Rahim, N., Selvaraj, J., Pandey, A., Khan, A.F., 2016. Role of smart grid in renewable energy: An overview. Renew. Sustain. Energy Rev. 60, 1168–1184.

Huber, P., Calatroni, A., Rumsch, A., Paice, A., 2021. Review on Deep Neural Networks Applied to Low-Frequency NILM. Energies 14 (9), 2390.

Huchtkoetter, J., Reinhardt, A., 2020. On the impact of temporal data resolution on the accuracy of non-intrusive load monitoring. In: Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. pp. 270–273.

Jiang, J., Kong, Q., Plumbley, M.D., Gilbert, N., Hoogendoorn, M., Roijers, D.M., 2021. Deep learning-based energy disaggregation and on/off detection of household appliances. ACM Trans. Knowl. Discov. Data (TKDD) 15 (3), 1–21.

Kabalci, Y., 2016. A survey on smart metering and smart grid communication. Renew. Sustain. Energy Rev. 57, 302–318.

Kaselimi, M., et al., 2022. Towards trustworthy energy disaggregation: A review of challenges, methods, and perspectives for non-intrusive load monitoring. Sensors 22 (15), 5872.

Kelly, J., Knottenbelt, W., 2015. The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. Sci. Data 2 (1), 1–14.

Kolter, J.Z., Johnson, M.J., 2011. REDD: A public data set for energy disaggregation research. In: Workshop on Data Mining Applications in Sustainability (SIGKDD). vol. 25, Citeseer, San Diego, CA, pp. 59–62.

Krystalakos, O., Nalmpantis, C., Vrakas, D., 2018. Sliding window approach for online energy disaggregation using artificial neural networks. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence. pp. 1–6.

Machlev, R., Malka, A., Perl, M., Levron, Y., Belikov, J., 2022. Explaining the decisions of deep learning models for load disaggregation (NILM) based on XAI. In: 2022 IEEE Power & Energy Society General Meeting. PESGM, IEEE, pp. 1–5.

Makonin, S., Popowich, F., 2015. Nonintrusive load monitoring (NILM) performance evaluation. Energy Eff. 8 (4), 809–814.

Massidda, L., Marrocu, M., Manca, S., 2020. Non-intrusive load disaggregation by convolutional neural network and multilabel classification. Appl. Sci. 10 (4), 1454.

Mollel, R.S., Stankovic, L., Stankovic, V., 2023. Explainability-informed feature selection and performance prediction for nonintrusive load monitoring. Sensors 23 (10), 4845.

Murray, D., Stankovic, L., Stankovic, V., 2021. Transparent AI: explainability of deep learning based load disaggregation. In: Proc. the 8th ACM Int. Conf. Sys. Energy-Eff. Buildings, Cities, and Transp..

Murray, D., Stankovic, L., Stankovic, V., Lulic, S., Sladojevic, S., 2019. Transferability of neural network approaches for low-rate energy disaggregation. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 8330–8334.

Pan, Y., et al., 2020. Sequence-to-subsequence learning with conditional gan for power disaggregation. In: Proc. - ICASSP IEEE Int. Conf. Acoust. Speech Sig. Process..

Rafiq, H., Shi, X., Zhang, H., Li, H., Ochani, M.K., Shah, A.A., 2021. Generalizability improvement of deep learning-based non-intrusive load monitoring system using data augmentation. IEEE Trans. Smart Grid 12 (4), 3265–3277.

Selvaraju, R.R., et al., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proc. IEEE Int. Conf. Comput. Vis..

Siano, P., 2014. Demand response and smart grids—A survey. Renew. Sustain. Energy Rev. 30, 461–478.

Smilkov, D., et al., 2017. Smoothgrad: removing noise by adding noise. arXiv:1706.03825.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. In: Int. Conf. Machine Learn.. PMLR.

Tanoni, G., Principi, E., Squartini, S., 2022. Multilabel appliance classification with weakly labeled data for non-intrusive load monitoring. IEEE Trans. Smart Grid 14 (1), 440–452.

Todic, T., Stankovic, V., Stankovic, L., 2023. An active learning framework for the low-frequency non-intrusive load monitoring problem. Appl. Energy 341, 121078.

Yue, Z., Witzig, C.R., Jorde, D., Jacobsen, H.A., 2020. Bert4nilm: A bidirectional transformer model for non-intrusive load monitoring. In: Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring. pp. 89–93.

Zhang, X.Y., Watkins, C., Kuenzel, S., 2022. Multi-quantile recurrent neural network for feeder-level probabilistic energy disaggregation considering roof-top solar energy. Eng. Appl. Artif. Intell. 110, 104707.

Zhang, C., et al., 2018. Sequence-to-point learning with neural networks for non-intrusive load monitoring. In: Proc. AAAI Conf. Artif. Intell.. AAAI, 32, (1).