Towards Trustworthy AI Systems for Smart Grid Management: Facilitating Robustness, Transparency and Fairness in the Energy Transition

PhD Thesis

Djordje Batic

A thesis presented for the degree of
Doctor of Philosophy

Department of Electronic and Electrical Engineering
University of Strathclyde
United Kingdom
03/04/2025

Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree. The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Acknowledgements

I would first and foremost like to express my sincere gratitude to my academic supervisors, Prof. Vladimir Stankovic and Prof. Lina Stankovic. Their support, guidance, and expertise have been instrumental throughout my doctoral journey. I have been incredibly fortunate to have supervisors who consistently encouraged me to explore new ideas and provided patience and wisdom in navigating challenges. I would also like to extend my sincere thanks to my examiners, Dr. Nikos Deligiannis and Dr. Johannes Norheim, whose constructive feedback, discussions, and valuable insights have greatly refined this thesis. I deeply appreciate the time and dedication they have invested in the examination process. Finally, I wish to acknowledge the entire GECKO Marie Curie ITN consortium. Being part of this network has provided a unique and enriching environment for research and collaboration. The opportunities for training, networking, and interdisciplinary collaboration with fellow researchers, leading academics, and industrial partners across Europe have been immensely beneficial and have significantly contributed to the broader context of this work.

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.

Abstract

The integration of Artificial Intelligence (AI) into Smart Grid (SG) management presents significant opportunities for enhancing energy efficiency, reliability, and sustainability. However, deploying AI in critical energy infrastructure raises challenges related to robustness, transparency, and fairness, essential components of trustworthy AI. This thesis addresses these challenges by proposing novel methodologies and frameworks to enhance trustworthiness in AI-driven SG management. First, it introduces a quantitative framework for evaluating and visualizing explainability in deep learning-based Non-intrusive Load Monitoring (NILM) systems. Next, it presents a new training enhancement approach that incorporates explainability principles directly into training of NILM models, achieving improvements in interpretability and predictive performance. Recognizing the constraints of deploying complex AI models on edge devices, the thesis proposes an explainability guided knowledge distillation framework that balances model efficiency with interpretability and reliability, facilitating robust edge deployment without compromising performance. Finally, it addresses equity concerns in Electric Vehicle Charging Station (EVCS) infrastructure placement by developing a geodemographic-aware placement strategy using Graph Neural Networks (GNNs), ensuring equitable access across diverse socioeconomic groups. Collectively, these contributions establish a comprehensive approach to embedding robustness, transparency, and fairness into AI applications within the SG context, aligning technical innovation with ethical

and societal imperatives. This work supports broader adoption of trustworthy AI, contributing significantly to sustainable development and equitable energy transition objectives.

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence

BCE Binary Cross-Entropy Loss

CNN Convolutional Neural Network

CRNN Convolutional Recurrent Neural Network

DL Deep Learning

DNN Deep Neural Network

EU European Union

EV Electric Vehicle

EVCS Electric Vehicle Charging Station

GAN Generative Adversarial Network

GAT Graph Attention Network

GHG Greenhouse gas

GNN Graph Neural Network

GRU Gated Recurrent Unit

ICD Incremental Coverage Difference metric

KD Knowledge Distillation

LSTM Long Short Term Memory

MAE Mean Absolute Error

NILM Non-intrusive Load Monitoring

POI Point of Interest

PVs Solar photovoltaics

RNN Recurrent Neural Network

ROTR Robustness-Trust metric

SG Smart Grid

SGDs Sustainable Development Goals

SIMD Scottish Index of Multiple Deprivation

UK United Kingdom

UN United Nations

XAI Explainable AI

Chapter 1

Introduction

1.1 Trustworthy Artificial Intelligence in Smart Grid Management

The integration of AI in SG management systems marks a critical juncture where technological advancement intersects with ethical imperatives. As AI systems increasingly control critical infrastructure decisions - from power distribution to resource allocation and demand forecasting - the need for responsible implementation has become paramount. This urgency was highlighted by documented cases where AI systems exhibited algorithmic bias in various domains [24, 44, 86], prompting concerns about similar risks in SG applications [4, 74, 87]. In response, major organizations developed comprehensive frameworks for responsible AI implementation, such as the European Commission's Ethics Guidelines [36], OECD Principles [93], Toronto Declaration [58], The Bletchley Declaration [127], Seoul Declaration [94] and others, which established foundational standards positive impact of AI, emphasizing human rights and well-being, democratic values, fundamental freedoms, and privacy. Specifically, at the core of Trustworthy AI framework [36] lie three fundamental pillars: lawfulness, ethics, and robustness. These pillars

give rise to seven essential requirements that directly shape SG implementations: human oversight, technical safety, data privacy, transparency, non-discrimination, societal well-being, and accountability.

In the SG context, each requirement presents distinct challenges with far-reaching implications. For instance, human oversight must balance automated efficiency with human control in time-critical grid operations, while data privacy requirements must protect sensitive consumption patterns without compromising system optimization capabilities. The translation of these theoretical frameworks into practical SG applications reveals a critical gap in current research. While extensive work has focused on optimizing technical performance metrics such as prediction accuracy and computational efficiency, less attention has been paid to implementing and measuring trustworthy AI properties like fairness, privacy, and transparency. This implementation challenge is compounded by the inherent tensions between different trustworthy AI requirements. For example, enhancing privacy protection through data encryption or anonymization can reduce system transparency and interpretability, while increasing human oversight might compromise real-time performance in critical grid operations.

These challenges motivate four key research directions in AI-powered SG systems. First, we need new methods to introduce explainability and transparency into deep learning-based systems, supported by mathematical guarantees that ensure the quality and reliability of explanations. Second, we must develop training processes that incorporate explainability objectives alongside traditional performance metrics, optimizing both transparency and robustness simultaneously rather than treating them as separate concerns. Third, we need innovative approaches to maintain interpretability and reliability when deploying complex systems on resource-constrained edge devices, where traditional explainability methods may be computationally prohibitive and relationship between transparency and privacy is often seen as a trade-off. Finally, we must design decision-making frameworks that holistically

integrate equity considerations with technical constraints, ensuring that SG advancements benefit all stakeholders fairly.

1.2 Research Motivation and Aims

The global transition to renewable energy sources represents a fundamental transformation of energy infrastructure, demanding sophisticated SG technologies to manage increasingly complex and distributed energy systems. This transition presents three interconnected challenges: the need for granular energy monitoring and optimization, the requirement for efficient edge computing solutions, and the imperative to ensure equitable access to emerging energy technologies. These challenges must be addressed while adhering to trustworthy AI principles, creating a multi-dimensional challenge at the intersection of technical performance and societal impact. At the monitoring and optimization level, NILM has emerged as a crucial technology for understanding and managing energy consumption patterns. Deep Learning (DL)-based NILM systems have demonstrated exceptional accuracy in disaggregating energy data, providing valuable insights for both consumers and utilities. However, their "black-box" nature presents significant challenges for deployment in critical infrastructure. The lack of transparency in these systems raises concerns about reliability, interpretability, and accountability - key requirements for trustworthy AI in practice. The practical deployment of NILM systems introduces additional complexities, particularly in edge computing environments. While edge deployment offers advantages in privacy protection and reduced latency, it creates tension between computational efficiency, model performance, and interpretability. Traditional explainability methods, often computationally intensive, may not be feasible on resource-constrained devices. This necessitates novel approaches that can maintain both performance and trustworthiness under practical constraints. Lastly, EVCS infrastructure placement decisions often favour affluent ar-

eas, potentially reinforcing socio-economic disparities, necessitating a need for more equitable infrastructure planning. These challenges together motivate the research agenda of this thesis, which aims to develop methodologies and frameworks that advance both technical capabilities and trustworthy AI principles in SG applications. Thus, this thesis investigates four fundamental research questions:

RQ1: Quantifying and Evaluating NILM Explainability How can we effectively evaluate and quantify the explainability of deep learning-based NILM systems? This question encompasses:

- Development of rigorous metrics for assessing explanation quality
- Validation of explanation faithfulness
- Evaluation of explanation robustness across different scenarios
- Assessment of explanation utility for different stakeholders

RQ2: Explainability-Aware NILM Training How can explainability be incorporated into the NILM training process to improve both transparency and performance? Key aspects include:

- Investigation of relationship between explainability and model robustness
- Analysis of the relationship between explainability and predictive performance
- Development of multi-objective training strategies

RQ3: Edge-Deployed Interpretable NILM How can we maintain interpretability and reliability when deploying NILM systems on edge devices? Critical considerations include:

- Analysis of trade-offs between model compression and explainability

- Optimization of knowledge distillation for both efficiency and interpretability
- Development of reliable performance guarantees under resource constraints

RQ4: Equitable EVCS Infrastructure Planning How can we design equitable Electric Vehicle (EV) charging infrastructure placement strategies that balance technical and social factors? This encompasses:

- Development of methods for modeling urban dynamics and social factors
- Identification of factors influencing utilization patterns and charging behavior
- Creation of metrics and algorithms for ensuring equitable access

These research questions form an integrated framework for advancing trustworthy AI principles in SG applications, addressing both technical challenges and societal implications of the energy transition.

1.3 Contribution of Thesis

The contributions of this thesis are organized around four critical aspects of the energy transition:

1. **Transparent Load Disaggregation**: Development of a comprehensive framework for evaluating and quantifying the explainability of deep learning-based NILM systems, and introduction of novel visualization techniques and rigorous evaluation metrics for assessing explanation quality.

2. **Explainability-informed AI Training**: Demonstration of methods to improve both transparency and performance through explainability-informed training. Introduction of novel learning mechanisms and loss techniques to enhance model performance and robustness.

3. **Interpretability and Robustness in Knowledge Distillation on the Edge**: Design of a novel knowledge distillation framework that improves knowledge distillation by providing interpretability and reliability when deploying NILM systems on edge devices. Development of methods to balance computational efficiency with model performance and interpretability.

4. **Equitable Infrastructure Planning**: Creation of a geodemographic-aware methodology for optimal placement of EVCS. Development of graph neural network approaches that consider both technical and social factors in infrastructure planning. Implementation of novel metrics and algorithms to ensure fair access across different socio-economic groups.

This research advances the field in several ways: i) Provides novel frameworks for evaluating and improving the explainability of AI systems in energy applications, ii) Demonstrates practical methods for deploying efficient and interpretable AI models on edge devices, and iii) Introduces innovative approaches to ensure equitable access to emerging energy technologies. By addressing these challenges, this thesis contributes to the development of more transparent, efficient, and equitable SG systems that can support the ongoing energy transition while ensuring that algorithmic decisions adhere to principles of Trustworthy AI are, namely transparency, robustness, and fairness. As a result, this thesis is in line with United Nations (UN) Sustainable Development Goals (SGDs) [13], specifically, SGD 7.3 regarding energy efficiency, SGD 16.6 regarding development of effective, accountable and transparent systems, SGD 9.1 regarding equitable access to infrastruc-

ture and SDG 11.2 regarding sustainable and accessible transport systems for all. While the Trustworthy AI principles of transparency, robustness, and fairness were primary motivators of this thesis, it is important to note that other principles were indirectly addressed. For example, privacy principle has a direct link to the edge deployed algorithms for NILM, while metrics developed for evaluation of interpretability of XAI methods for NILM can be interpreted as approaches that facilitate human oversight by providing tools and metrics to understand and evaluate the decisions made by AI models. Consequently, these advancements in transparency and robustness lay the groundwork for improved accountability and technical safety. Finally, the focus on equitable EVCS placement directly promotes societal well-being and non-discrimination.

1.4 Thesis Chapters Overview

The thesis is organized as follows: Chapter 2 provides a literature review of key challenges in implementing Trustworthy AI systems for SG management. Chapters 3 and 4 focus on explainability evaluation and improvement in NILM systems. Chapter 5 addresses the challenges of edge deployment while improving interpretability. Chapter 6 presents methods for equitable EVCS infrastructure planning. Finally, Chapter 7 concludes with a summary of contributions and future research directions. This introduction provides a clearer structure and better highlights the key contributions and significance of the research. It emphasizes how the work addresses important technical and societal challenges in the energy transition while maintaining focus on trustworthy AI principles.

1.5 Publications

1. Djordje Batic, Vladimir Stankovic, and Lina Stankovic. Toward transparent load disaggregation—a framework for quantitative evaluation of explainability using explainable ai. IEEE Transactions on Consumer Electronics, 2024 (Chapter 3). *Author's contribution: Conceptualization, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis.*

2. Djordje Batic, Vladimir Stankovic, and Lina Stankovic. XNILMBoost: Explainability-Informed Load Disaggregation Training Enhancement using Attribution Priors. Engineering Applications of Artificial Intelligence, 2024 (Chapter 4) *Author's contribution: Conceptualization, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis.*

3. Djordje Batic, Vladimir Stankovic, and Lina Stankovic. Geodemographic Aware Electric Vehicle Charging Location Planning for Equitable Placement using Graph Neural Networks: Case Study of Scotland Metropolitan Areas. Energy, 2025 (Chapter 6) *Author's contribution: Conceptualization, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis.*

4. Djordje Batic, Giulia Tanoni, Emanuele Principi, Lina Stankovic, Vladimir Stankovic, Stefano Squartini. Interpretability and Reliability-driven Knowledge Distillation for Non-intrusive Load Monitoring on the Edge. Under Review, 2025 (Chapter 5) *Author's contribution: Conceptualization, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis.*

5. Djordje Batic, Giulia Tanoni, Lina Stankovic, Vladimir Stankovic, and Emanuele Principi. 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, 2023. (Chapter 5) *Author's contribution: Conceptualization, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis.*

6. Djordje Batic, Vladimir Stankovic, and Lina Stankovic. ChargeDEM: geodemographic awareEVcharging infrastructure placement for enhanced site selection using graph neural networks. In 12th International Conference on Energy Efficiency in Domestic Appliances and Lighting (EEDAL'24). 2024. (Chapter 6) *Author's contribution: Conceptualization, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis.*

7. Djordje Batic, Lina Stankovic, and Vladimir Stankovic. Analysis of Smart Meter Data for Energy Waste Management. In Artificial Intelligence for Sustainability: Innovations in Business and Financial Services, pp. 153-173. Cham: Springer Nature Switzerland, 2024. (Chapter 2) *Author's contribution: Writing – original draft, Investigation.*

Chapter 2

Literature Review

2.1 Background - Smart Grid and Energy Transition

Modern energy systems face unprecedented challenges related to integration of renewable energy sources, distributed generation, and SG technologies. As the world moves to more renewable energy sources, the impending energy transition requires significant transformations in the energy sector. Not only will renewables cause major increases in electricity demand, but they will also cause substantial supply fluctuations [59]. Solar photovoltaics (PVs) and wind turbines, the main drivers of renewable technologies, only operate when sunlight or wind is abundant. Because of this, the electricity grid is subjected to massive injections of PV or wind power during sunny or windy days and low inputs during cloudy or calm days, which poses a grid stability and renewable energy integration challenge. EVs can absorb excess energy that would otherwise be wasted, improving the economic benefit of wind and solar power generation [104]. However, this can introduce major changes in the electricity demand, especially in locations with a large number of charging stations, where EVs may be charged at the same time during the day,

leading to a very high load on the distribution grid and uncertainty in electricity supply meeting demand [96]. Another important element of modern grids is that an increasing number of energy consumers are becoming producers of energy, i.e., prosumers. As a result, the generation, transmission, distribution, and control operations of the traditional grid need to be able to accommodate the dramatic changes caused by the transition towards renewable energy and the electrification of transport [98]. This need has led to the emergence of SG systems that aim to facilitate operational efficiency and reliability of the grid using advances in infrastructure, intelligent information collection, automation and knowledge extraction [43].

SGs enable the decentralization of distribution and communication and are at the forefront of efforts geared towards addressing operational complexity introduced by the increased utilization of renewable energy. It aims to deliver power more efficiently and automatically address any events that may impact the quality or reliability of the power supply and generation [46]. However, despite the apparent benefits, it is important to note that in some instances, energy efficiency efforts lead to rebound effects (a phenomenon where energy efficiency can lead to an increase in the consumption of energy services). Consequently, rebound effects could offset the expected benefits of SGs, where lower energy costs induced by them may stimulate additional energy use. In the context of rebound mitigation, one promising initiative to boost environmental consciousness is raising awareness of households' energy consumption behavior [133]. Knowledge extracted from electricity use behavior can be used to better understand the changing needs of consumers and prosumers alike and make sense of complex consumption patterns, and the core technology that enables this is smart metering [42]. Successful operation of SG infrastructure depends on the ability to collect and extract knowledge from live load measurements that will enable better monitoring of supply and demand, reduce operational costs, and optimize energy efficiency. Smart meters enable real-time consumption feedback that is communicated remotely

to utilities, consumers, prosumers, and other stakeholders with an interest in meeting energy efficiency targets. Unlike traditional meters, which usually provide consumption feedback only once a month, smart meters can provide feedback in real-time. Smart meters represent the core pillar of the SG and are the key to the successful realization of future energy management systems that make informed decisions for consumers, electricity producers, and network operators alike. SG technology is essential in lowering carbon emission goals, as it facilitates the broader incorporation of renewable energy sources like solar and wind power. This incorporation enables small-scale electricity production, enhances supply and demand flexibility, assures accurate customer billing, and promotes the decentralization of power generation [116].

2.2 Trustworthy AI

The integration of AI in SG management has emerged as a critical challenge at the intersection of technological innovation and social responsibility [11]. While AI presents unprecedented opportunities for improving energy systems, recent developments in trustworthy AI frameworks highlight the need for solutions that not only deliver technical performance but also ensure fairness, accountability, and transparency in critical infrastructure decisions. As emphasized by the previous work [48], AI applications must align with sustainable development goals while adhering to ethical principles that protect human rights and promote inclusive growth. The initial push for responsible AI emerged from early warning signs of AI systems perpetuating or amplifying existing societal inequities. High-profile cases of algorithmic bias in recruitment, facial recognition, and criminal justice systems demonstrated how AI trained on historical data could systematically disadvantage certain populations. Simultaneously, the successful application of AI in projects like automated monitoring of crop diseases, predictive modeling for poverty reduction, and climate informatics demonstrated the technology's potential

for advancing sustainable development goals and delivering positive social impact. These contrasting outcomes prompted leading organizations and institutions to develop comprehensive frameworks for responsible and trustworthy AI. The European Commission's Ethics Guidelines emphasized that AI systems must be lawful, ethical, and robust to avoid unintended harm. The Montreal Declaration for Responsible AI and Toronto Declaration specifically addressed the intersection of AI development with human rights and discrimination concerns. As a result of rapid expansion of generative AI technologies, many of the frameworks have been expanded and updated, such as OECD Principles framework [93] which established that AI should drive inclusive growth while respecting human rights, democratic values, and diversity. These frameworks collectively stress that trustworthy AI requires not just technical excellence, but also careful consideration of ethics, fairness, transparency, and accountability. Correspondingly, Trustworthy AI for SG management systems would imply that the algorithms trained on user data obtain good predictive performance but are also designed to emphasize other important properties such as robustness, fairness, transparency, and privacy. Accounting for these properties is an essential step towards wider adoption of AI and an opening for a new, mature era of AI design and deployment.

Building on the established frameworks for trustworthy and responsible AI [36, 58, 93, 94, 127], we can identify three fundamental pillars that must be addressed in SG applications: lawfulness, ethics, and robustness. The lawful dimension ensures compliance with energy sector regulations and data protection laws; the ethical dimension addresses fair access to energy resources and environmental sustainability; and the robustness dimension guarantees reliable operation even under adverse conditions or other threats. These pillars support seven key requirements that must be fulfilled: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and non-discrimination, societal and environmental well-being, and accountability. In the context of SGs, these requirements

take on specific meanings and challenges. For instance, human oversight must balance automated grid management with human control over critical infrastructure decisions. Privacy concerns extend beyond individual consumer data to include industrial energy usage patterns that could reveal trade secrets. Fairness must consider equitable access to energy resources while maintaining grid stability, and transparency must be balanced against potential security vulnerabilities that could be exploited by malicious actors.

However, what is not yet clear is how to effectively implement trustworthy AI principles in practical SG applications while maintaining high performance standards. Most studies in the field of AI-powered grid management have focused solely on technical metrics such as prediction accuracy, response time, and optimization efficiency, while the crucial aspects of fairness, accessibility, and transparency have received limited attention. This technical-centric approach has created a significant gap between theoretical frameworks for trustworthy AI and their practical implementation in SG systems. Furthermore, it is vital to understand that the pursuit of trustworthiness should include awareness of the interactions between the properties of the trustworthy system, which is particularly important when requirements interfere with each other. For example, by ensuring strong data privacy guarantees, the predictive performance of the system may suffer. Additionally, if input data is significantly altered, it might negatively affect the system's transparency. Thus, even though there are significant possibilities for progress in individual instances of trustworthiness, it is equally important to understand the requirements and interactions of each aspect of the trustworthy system.

2.3 Towards Trustworthy Artificial Intelligence in Smart Grid Management

2.3.1 Explainable Non-Intrusive Load Monitoring

NILM Problem Statement and Low-frequency NILM Algorithms

Given a sequence of aggregated power consumption $\mathbf{y} = (y_1, y_2, ..., y_T)$, captured at time $t = \{1, 2, ..., T\}$, the goal of a NILM algorithm is to determine the individual power contribution x_t^i of appliance $i \in \{1, 2, ..., M\}$, 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:

$$y_{t=1...T} = \sum_{i=1}^{M} x_t^i + \epsilon_t \tag{2.1}$$

To extract the power consumption of a selected appliance $i \in \{1, 2, ..., M\}$, 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 sec to 30 minutes. 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) [12, 57]. 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 [10].

In order to infer individual appliance consumption, various machine learning approaches have been proposed in the recent literature, where Deep Neural Network (DNN) approaches form the basis of state-of-the-art implementations according to the recent review of [10]. Convolutional Neural Networks (CNNs) form the majority of implementations - [138] propose a sequence-to-point (seq2point) learning approach using a CNN; [97] address the high computational complexity of seq2point and propose a CNN architecture for sequence-to-subsequence learning; [35] use a two sub-networks that are connected in order to infer both regression and classification outputs; [92] propose a CNN model that provides generalisability to new domains; [85] 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); [139] use a multi-quantile RNN to disaggregate the loads and improve the demand side management of solar energy; [71] propose a Gated Recurrent Unit (GRU) approach that reduces memory usage and computational complexity while achieving good disaggregation performance, while [122] proposes a Convolutional Recurrent Neural Network (CRNN) approach for multi-label classification of appliances. Lastly, other literature attempts to introduce new learning mechanisms include Generative Adversarial Networks (GANs) [97], temporal-causal networks [51] and atten-

tion mechanisms [136]. A DNN-focused review for low-frequency NILM [56] provides a detailed review of current DNN NILM approaches, where GRU and CNN architectures and their variants, including WaveNet with dilated convolutions [51], 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.

Explainable AI for Low-frequency NILM

The increasing integration of AI into SG management, particularly for tasks like NILM, promises significant benefits in energy efficiency and grid stability. However, many high-performing AI models, especially DL architectures, operate as 'black boxes.' Due to their complexity, the internal decision-making processes are not well understood, making it difficult for human operators, developers, and even end-users to understand how a particular prediction or decision was reached. This lack of transparency poses significant challenges. First, if stakeholders cannot understand or trust the AI's reasoning, adoption of these powerful tools in critical infrastructure will be challenged with skepticism. This is particularly true for consumer-facing applications like NILM, where understanding energy usage breakdowns is key. Second, when a black-box model makes an error, diagnosing the cause is incredibly difficult. Explainability can help reveal flawed logic or biases in the model, guiding developers towards more robust and accurate systems. Third, in critical systems like SG management, unexpected or erroneous AI behavior can have severe consequences. Understanding the 'why' behind AI decisions is crucial for ensuring safety, identifying potential failure modes, and building resilient systems. Lastly, without transparency, it is hard to assess if an AI system is making fair decisions or if it has learned unintended biases from the data, potentially leading to inequitable outcomes (e.g., in demand-response programs or tariff assignments).

The wider problem of explainability of DL models has recently gained

traction, leading to the emergence of the field of Explainable AI (XAI). Recent literature [9,15,47,109,111,113,118] suggest that XAI can facilitate trust by providing algorithmic transparency, support assessment of levels of bias, and improve the overall understanding of the inner workings of DL models. The majority of XAI work, predominantly tackling computer vision tasks, primarily centers around the integration and development of techniques that analyse the outputs of the model and visualise the importance of the input features. These include approaches such as GradCAM [109], IntegratedGradients (IG) [118], LRP [15], and SmoothGrad [113]. These methods, also known as feature-attribution methods, are effective in revealing problems in a model, understanding the model decisions, or revealing dataset bias. Given a trained DL-based model and an input, the goal is to provide a prediction and attribute a score in a way that high scores correspond to important features of the input. For image modality, this could be an area of the image associated with the predicted class, while for textual data this could be a set of important words or tokens. In the context of NILM, an overview of a XAI-supported system can be seen in Fig. 2.1. The system provides the prediction, as well as the associated heatmap that shows areas of the signal that are highly important. It is important to note that such a system doesn't provide guarantees of robustness or transparency. For this to be the case, both the DL model as well as XAI method need to be evaluated.

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 [89]. 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 [90] 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 analysing 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 [80] compares the success of using the GradCAM XAI technique against occlusion sensitivity for visualising 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.

2.3.2 Knowledge Distillation and Edge Deployment for Privacy Preservation

The deployment of NILM systems on edge devices has emerged as a promising direction for practical energy monitoring applications. Edge-based approaches address privacy concerns and reduce latency by processing data locally rather than transmitting sensitive consumption data to external servers. While various implementations have been proposed, ranging from Raspberry Pi [134] to more resource-constrained micro-controllers [120], FPGA [52] and

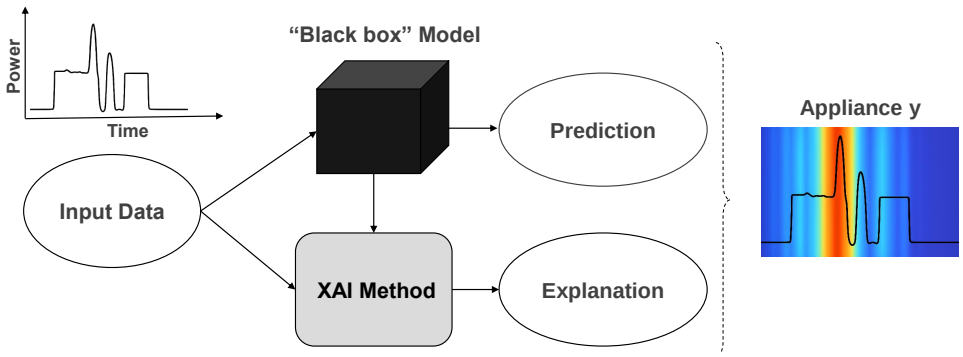


Fig. 2.1. Overview of XAI-assisted decision making for NILM. The trained DL (black-box) model performs predictions, while the XAI method (e.g. GradCAM [109]) provides explanations for a particular prediction. The resulting visualization contains the predicted appliance (y), as well as a heatmap that indicates the important areas of the signal that correspond to the predicted appliance y. The red-hued areas show the most important area of a signal that lead to the prediction of appliance y.

ARM [84] platforms, significant challenges remain in deploying deep learning models on these platforms due to their computational limitations. DNNs, despite their success in NILM applications, typically contain hundred thousands or millions of parameters, making their deployment on resource-constrained devices particularly challenging. Traditional approaches to NILM model compression for edge deployment have explored various techniques, including quantization [5], model design optimization [78], parameter pruning [72], feature fusion [49] and Knowledge Distillation (KD) [123], as well as a combination of different approaches [119]. However, while these methods have shown promising in reducing model size and computational requirements, they often address only the computational aspects without considering the interpretability and reliability of the compressed models. The challenges of deploying NILM systems on edge devices, particularly the trade-offs between model complexity, performance, and interpretability, highlight a critical need for approaches that can maintain model transparency even after compression. This has led to an increased focus on XAI techniques specifically designed

for NILM applications, as described in the previous section. KD has emerged as a particularly promising approach for edge NILM deployment, as it enables the transfer of knowledge from complex teacher models to more efficient student models while maintaining or improving predictive performance. Previous work has explored KD-based weakly supervised learning strategies to train less complex networks maintaining the classification performance [123]. Starting from a CRNN model considered as the Teacher, several lower complexity CRNNs have been distilled to evaluate how the performance is influenced by removing convolutional layers or recurrent units, showing promising performance in resource constrained environments. However, existing research has revealed critical limitations in the knowledge transfer process. Previous studies, notably [123], overlooked critical aspects of how incorrect Teacher knowledge impacts Student learning outcomes. Namely, while conventional approaches to distillation focus primarily on optimizing Student learning processes, they fail to address a fundamental challenge: the presence of corrupted knowledge in the Teacher network. In the context of NILM, this corruption manifests in two critical dimensions: state classification accuracy and explainability. From a state classification perspective, the Teacher network's imperfect nature [123] introduces inherent uncertainties that fundamentally compromise the NILM system's reliability. Simultaneously, from an explainability standpoint, the absence of rigorous validation of Teacher explainability prior to and during the distillation raises serious concerns about the interpretability of the designed system. This indicates that deployment of NILM systems on edge devices introduces additional constraints beyond mere model compression. Edge devices must operate effectively across diverse environmental conditions and usage patterns while maintaining reliable performance, which necessitates not only efficient model architectures but also robust and interpretable decision-making processes that can be validated and trusted by end-users. The intersection of these requirements - computational efficiency, reliability, and interpretability - presents a complex challenge that

current edge NILM approaches have yet to fully address.

2.3.3 Equitable EV Charging Infrastructure Placement

EVs have emerged as a cornerstone in the strategy towards decarbonisation of the transport sector and the wider transition towards net-zero. A comprehensive range of solutions and policy interventions have been proposed, aimed at promoting the ownership of EVs and reducing economic costs for end-users [22]. Globally, EV ownership reached 26 million by the end of 2022 and is expected to rise to over 240 million by 2030 [1]. Notably, China, as the world's leading EV market, accounted for 14.1 million of these vehicles [1]. Meanwhile, in the United Kingdom (UK), more than 950,000 EVs were registered by the end of 2022, with numbers predicted to escalate quickly in response to increasing demand [114]. Recent advancements in EV technology and the gradual shift towards price parity with conventional vehicles are lowering barriers to entry, making EVs increasingly accessible and attractive to consumers. In addition, a suite of tax benefits and financial incentives are facilitating this trend in all leading EV markets, including China, the European Union (EU), the United States, and the UK. The UK had approximately 37,000 public EV charging devices at the end of 2022, equivalent to approximately 26 EVs to one charging point [1], though according to recent studies [77], the optimal ratio is 12 to 1. Although home charging currently meets a large portion of charging demand, publicly accessible charging is increasingly needed to provide accessibility, comfort, and facilitate long-distance driving akin to refueling a fossil fuel vehicle. This is particularly important in dense urban areas where access to home charging is more limited and public charging infrastructure is a key enabler for EV adoption. To this end, several leading economies have developed national EV charging infrastructure strategies: China has announced plans to accommodate charging infrastructure for more than 20 million EVs by 2025 [101]; the United States announced plans to invest up to $5 billion to promote the

penetration of EVs through introduction of 500,000 public chargers by 2030, fiscal incentives and subsidies [101]; Japan pushed forward the national target of 150,000 charging points by 2030, including 30,000 fast chargers [1]; the European Parliament announced the alternative fuel infrastructure regulation aimed at delivering charging infrastructure with a particular focus on fast charging stations and charging for heavy-duty vehicles [1]; the UK has allocated £1.3 billion in government funding aimed to support the roll-out of the charging infrastructure, with a particular focus on local on-street residential charging and targeted plug-in vehicle grants [101].

While numerous studies have explored optimal placement strategies for EVCS, the socio-economic dimensions of these placements have often been overlooked. Existing research has predominantly focused on technical and operational optimization criteria, such as accessibility, speed, and cost of charging, without adequately addressing the disparities in EVCS distribution across different locations and socio-economic groups [54]. This has resulted in EVCS infrastructure that is often dense in high-income neighborhoods while being sparse and/or underutilized in socially disadvantaged communities [76, 106, 115]. As a result, despite the advancements in EVCS placement strategies, there is a pressing need to address the inequities in EVCS distribution. Equity in EVCS placement ensures that charging infrastructure is accessible to all community segments. Additionally, some critical gaps remain in addressing spatial interdependencies, land use-specific deployment, and equity integration. Existing approaches predominantly employ spatial regression [79], clustering [39], or multi-objective optimization [77], treating urban areas as static grids rather than dynamic networks. For instance, [79] utilizes multi-scale geographically weighted regression to analyse the spatial heterogeneity in intra-city public EVCS distribution but neglects the interconnected nature of urban zones, overlooking how charger placement in one area influences demand in adjacent regions. Similarly, [77] applies multi-objective optimization and TOPSIS optimization to propose equitable place-

ment by balancing site development costs, equity access, and demand fulfillment. However, the decisions are made without modeling the spatial propagation of charging needs across a city's graph structure. While [75] employs agent-based modeling to simulate charging behavior, it treats urban networks as homogeneous grids rather than interconnected graphs. Similarly, [28] applies linear regression to correlate charger density with deprivation indices but ignores demand spillover effects between adjacent zones. Secondly, while land use categorization is acknowledged in studies like [27], which manually labels zones, prior works mostly fail to provide granular, land-use-specific policy recommendations. For example, [100] evaluates accessibility across broad census tracts and evaluates horizontal and vertical equity using spatial autocorrelation but does not differentiate optimal charger types (e.g., 22kW vs. fast chargers) for residential versus industrial zones. Similarly, [66] and [28] identifies correlations between EVCS distribution and income levels but provides no framework for zone-specific (residential/commercial) deployment. Third, equity considerations in existing frameworks are often reductionist. Studies like [55] correlate EVCS distribution with income levels but omit multi-dimensional deprivation indices and real-time utilization patterns. [75] models income-based charging access but omits multi-dimensional deprivation metrics (e.g., health, education). [66] uses census-based approach but fails to integrate real-time utilization data, masking disparities in deprived areas.

Above recent research highlights work emerging to cater for inclusive EVCS placement that considers various socio-economical factors. However, these prior studies generally contain the following limitations: (i) EVCS placement decisions are often made without understanding the spatial dynamics of the urban network, a crucial aspect of interconnected systems such as urban EVCS infrastructure, (ii) most works fail to address the challenge of targeted infrastructure deployment within specific urban land uses, which is becoming increasingly important as government funding is often targeted

within certain urban areas of the city, for example, residential, industrial, or other, and (iii) insufficient consideration of equity, and a lack of integration of multiple factors influencing equity and local contexts.

Chapter 3

A Framework for Quantitative Evaluation of Explainability in Load Disaggregation

Recently, DL approaches have seen increased adoption in NILM community. However, DL-NILM models are often treated as black-box algorithms, which introduces algorithmic transparency and explainability concerns, hindering wider adoption. In this chapter, we present a methodology for explainability of regression-based DL-NILM with visual explanations, using XAI. Two explainability levels are provided. Sequence-level explanations highlight important features of predicted time-series sequence of interest, while point-level explanations enable visualising explanations at a point in time. To facilitate wider adoption of XAI, we define desirable properties of NILM explanations - faithfulness, robustness and effective complexity. Addressing the limitation of existing XAI -NILM approaches that do not assess the quality of explanations, desirable properties of explanations are used for quantitative evaluation of explainability. We show that the proposed framework enables better understanding of NILM outputs and helps improve design by providing a visualization strategy and rigorous evaluation of quality of XAI methods,

leading to transparency of outcomes. The content of this chapter is taken from the work reported in "Towards Trustworthy Load Disaggregation - A Framework for Quantitative Evaluation of Explainability using XAI" [16].

DL based implementations for NILM have grown sharply over the past few years with very good performance demonstrated via domain-agnostic accuracy metrics, such as the popular Mean Absolute Error, across a wide range of real-world datasets [56]. However, using accuracy metrics as a standalone determinant for selection of an AI technology is inadequate for wider consumer adoption, as put forth in [63] and [62]. The latter recommends that, in order to ensure Trustworthy AI, robustness, fairness, transparency, and privacy need to be addressed. Indeed, the European Commission has recently published seven principles of Trustworthy AI [36], which include transparency as one of the key elements of trustworthy AI systems. Transparency is closely linked to traceability of the datasets, as well as explainability of the technical processes of the AI system and the related AI decisions, and finally communication of level of accuracy of an AI system and limitations to the end-users and system developers. For AI-based NILM, the majority of work has focused on addressing technical robustness in the form of accuracy, reliability and reproducibility across different datasets [56, 69, 130] and data transparency through the use of public, peer-reviewed and well-documented datasets [64, 91], with limited research in the area of privacy protection [26, 125, 140] and technical explainability [18, 80, 90]. The majority of DL-based NILM approaches are designed as “black-box“ systems due to their inherent algorithmic complexity and absence of explainability. Since the underlying mechanics resulting in NILM predictions are not interpretable or explainable, DL based NILM cannot be fully trusted, which somewhat hinders wider deployment of NILM systems [63]. As the adoption of smart home devices and energy management systems continues to grow, the necessity to ensure these technologies are both transparent and understandable to consumers grows concurrently. By developing and evaluating

XAI methods for NILM, the research community can contribute to design of AI solutions that adhere to consumer standards such as the EU's vision of ethical and responsible AI [36] and foster consumer trust in these emerging technologies, empowering users to make informed decisions about their energy consumption. Furthermore, understanding the produced outputs can help improve the design, provide a better overview of the model accuracy, and facilitate better understanding of failure scenarios. Thus, the role of explainability is to ensure a transparent inference process of the AI system by providing decisions that are understood and traceable. As a result, algorithmic transparency facilitated by explainability has been identified as a paramount challenge in the present landscape of NILM research [63].

The wider problem of explainability of DL models has recently gained traction, leading to the emergence of the field of XAI. Recent literature [9,15,47,109,111,113,118] suggest that XAI can facilitate trust by providing algorithmic transparency, support assessment of levels of bias, and improve the overall understanding of the inner workings of DL models. The majority of XAI work, predominantly tackling computer vision tasks, primarily centers around the integration and development of techniques that analyse the outputs of the model and visualise the importance of the input features. Such work frequently illustrates that explainability can enhance the understanding of the model and foster trust in the AI systems [47]. However, many existing XAI techniques can lead to unstable explanations in real-world scenarios due to limited, qualitative evaluation [20,31,110,112]. As a result, subjective evaluation of their quality could lead to false sense of trust in the XAI-supported system. Addressing such issues is particularly important for systems that can reveal personal information, such as temporal appliance patterns of use, generated by NILM. XAI approaches for NILM are still in their infancy, with limited literature available [18, 80, 89, 90]. While previous works have shown that XAI can be utilized as an important tool for contexts such as model debugging [18, 90], as XAI-based solutions for NILM continue to grow, it

is of vital importance to properly evaluate their explainability components. This assessment can serve as a way to assert that the used explainability techniques are truly able to be deployed in the real-world scenarios and help with understanding of model outputs. Therefore, XAI system design that incorporates robust qualitative and quantitative evaluation procedures for explainability techniques used in the real-world environment is of crucial importance for the successful adoption in NILM. The main contributions of this work are summarized as follows:

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

3.1 Proposed Explainability Evaluation Framework

The backbone of our proposed XAI framework for NILM is the proposed visualization procedure, illustrated in Fig. 3.1, that facilitates the generation of human-interpretable explanations of NILM model outputs. Since the desired granularity of explanations can vary, the visualization procedure offers an ability to generate explanations for both sequential-level, as well as

point-level predictions. The sequence-level explanations highlight the areas of the signal most responsible for the prediction, while the point-level explanations display the reasoning behind a prediction of a particular point in time. These two layers of explainability can be used interchangeably as they offer varying degrees of specificity. In the visualization procedure, we utilize five distinct XAI techniques to formulate explanations. Subsequently, the created explanations are subjected to a quantitative evaluation of quality. Taking into consideration a diverse set of needs and possible deployment scenarios, the quality of an explanation is defined as alignment with three desirable properties of explanations, specifically: faithfulness, robustness, and low complexity.

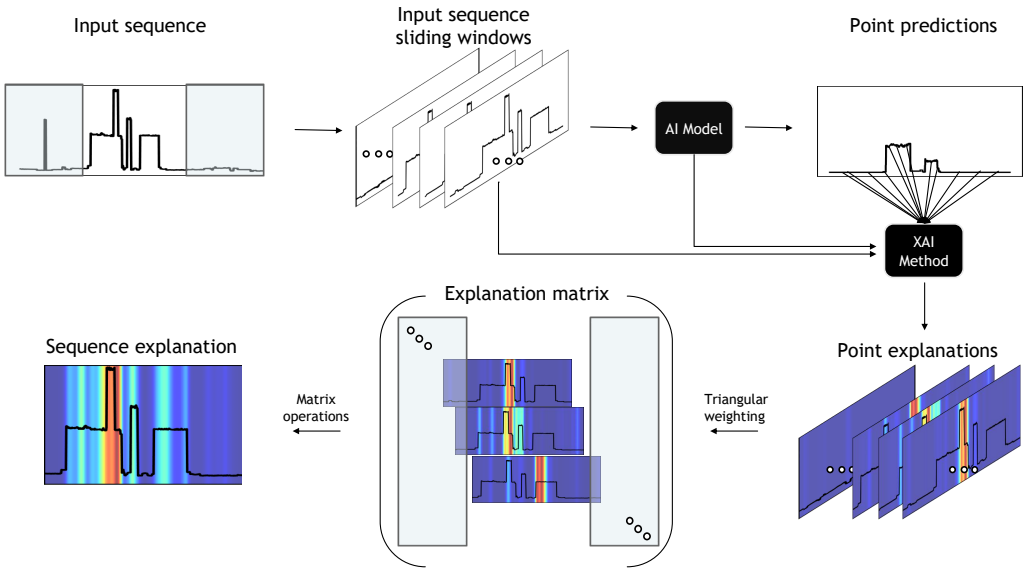


Fig. 3.1. Visual outline of the proposed approach showcasing the mechanism for visualization of importance at two levels of specificity, leading to point-level and sequence-level explanations for an input sequence of interest.

3.1.1 Visualization via heatmaps

We demonstrate how to integrate XAI in the popular seq2point DL-NILM implementation of [138] trained for load disaggregation of various appliances, via regression, on two popular datasets: UK-DALE [64] and REFIT [91]. The full procedure is illustrated in Fig. 3.1. First, to account for the nature of the seq2point algorithm, sliding windows are used to split the input signal into small, overlapping segments, and generate the point output predictions. Then, for a seq2point model with input size δ, for each generated point along the sliding window, a point explanation heatmap of size δ is created via existing XAI methods such as GradCAM [109], GradCAM++ [34], IntegratedGradients (IG) [118], LRP [15], and SmoothGrad [113]. If a step size of 1 is used, and the length of activation window of interest is ω, the total number of generated heatmaps is:

$$N = \omega - \delta + 1. \tag{3.1}$$

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

$$\omega' = \omega + 2 * (\delta - 1). \tag{3.2}$$

Given that the size of activation window of interest, ω, is larger than the model input size, δ, to map the N resulting heatmaps to a single, sequence-level heatmap of size ω, which corresponds to the activation of interest, we

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

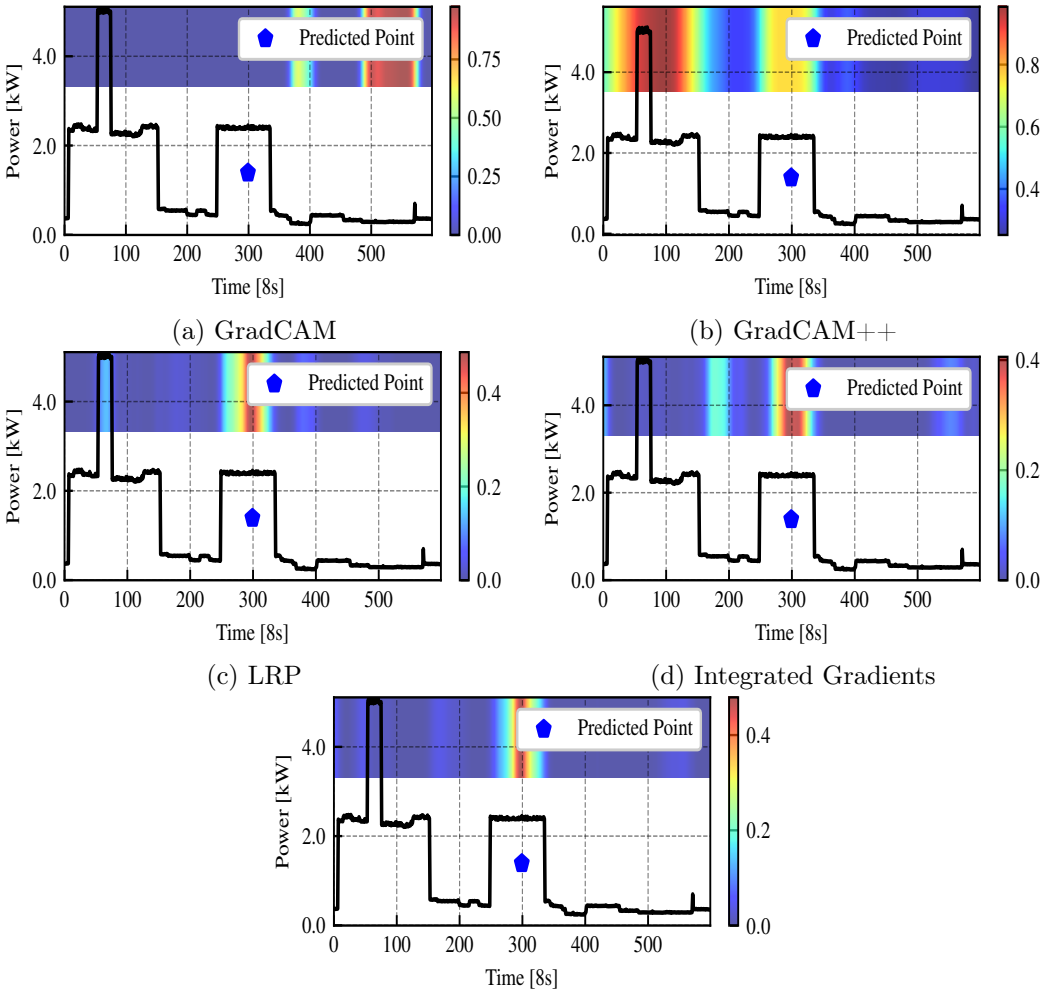
Before populating the matrix, we apply a weight function to mitigate the presence of noise and promote smoothness of heatmaps. Given that the temporal dimension of the middle point of the input corresponds to the output point of prediction, and is highly influential to the prediction, we apply a triangular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

$$\psi(x) = \begin{cases} \frac{x}{m}(p_{max} - p_{min}) + p_{min} & \text{if } 0 \leq x \leq m \\ \frac{x-m}{m}(p_{min} - p_{max}) + p_{max} & \text{if } m < x \leq 2m \end{cases} \tag{3.3}$$

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

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

the cumulative importance for each of the predicted points of the signal.



(a) GradCAM

(b) GradCAM++

(c) LRP

(d) Integrated Gradients

(e) SmoothGrad

Fig. 3.2. Explanations generated for positive activation of dishwasher in UK-DALE dataset. We can observe unreliable results from GradCAM, while other methods offer more accurate and concise explanations.

3.1.2 Properties of Explainable NILM Systems

Property of Faithfulness

The proposed faithfulness evaluation strategy quantifies the extent to which explanations attest to the predictive performance of a model. In other words, faithfulness aims to determine if the feature importance scores, generated by the visualization procedure, are indicative of importance w.r.t. prediction. The property of faithfulness addresses a fundamental question: Does the explanation accurately reflect the decision-making process of a model for a given prediction? An explanation is considered faithful if the input features it highlights as important are genuinely the ones the model relied upon. Conversely, an unfaithful explanation might mislead users by pointing to irrelevant features or missing crucial ones. In critical applications like NILM, where decisions based on AI outputs can impact energy management and user trust, ensuring the faithfulness of explanations is paramount. Given that a ground truth explanation can rarely be known, faithfulness is often assessed indirectly. The most common approach, and the one adopted in this work, involves perturbation-based evaluation. The core idea is: if an explanation correctly identifies important input features, then altering or removing these features should significantly degrade the predictive performance or confidence of a model for that specific instance. Conversely, altering unimportant features (as per the explanation) should have minimal impact. To measure the faithfulness of an XAI-enabled NILM approach, the following steps are taken:

1. Generate a sequence-level feature importance map of an input signal of interest.

2. Partition the sequence-level maps into sorted, non-overlapping segments based on the sum of importance scores over a certain period, to determine the most important areas of the input signal.

3. Evaluate the faithfulness of the derived explanations by performing an iterative perturbation of features by changing the input signal values in the segments of interest, starting with the segments of highest relevance. The perturbation of input segment is performed by replacing the power level of the initial signal by the signature of low consuming appliances (e.g., a combination of TV, Lights and Fridge, equaling around 250W). This perturbation ensures that the activation signal is attenuated, while keeping the input data distribution within the space that the model has learned on, as opposed to setting the power level to zero, which would constitute an unfavorable case of an out-of-distribution scenario.

4. To establish whether there is a significant impact on the predictive performance, after each perturbation of features we measure the difference between the performance metrics calculated on predictions of non-perturbed and perturbed signals.

5. To convey the degradation of performance, we consider both classification and regression-based performance metrics. As a way of capturing the classification performance, we convert the regression output to a step function and calculate the F_1 score as:

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}, \tag{3.4}$$

where TP stands for True Positives, FP for False Positives, and FN for False Negatives. To quantify the disaggregation performance, we utilize mean absolute error Mean Absolute Error (MAE) between the true (E_i) and predicted ($\hat{E}_i$) consumed energy of the appliance of interest where MAE is calculated as follows:

$$MAE = \frac{1}{T} \cdot \sum_{i=1}^{T} |\hat{E}_i - E_i|. \tag{3.5}$$

6. After each perturbation step, compute the difference between performance metrics of altered and original input. The faithfulness score is the resulting area under curve (AUC) after a set number of iterations, where more faithful XAI methods correspond to a higher AUC score. The classification faithfulness showcases the difference in F_1 score values, while regression faithfulness depicts the difference in MAE values. Iterative perturbation of features that leads to a sharper increase in the difference between the F_1 and/or MAE scores (and thus higher faithfulness score) suggests that the feature importance scores generated by the XAI method successfully assign scores to the highly relevant input features and are indeed indicative of predictive performance of the model.

Property of Robustness

The growing body of literature in deep learning theory [135] suggests that robustness of neural networks is closely related to the value of its local Lipschitz constant. Intuitively, a Lipschitz constant represents the value by which neural network's output is allowed to change relative to its input. It has been used as a hard constraint to enable adversarial robustness, better generalization and training of generative adversarial networks. Moreover, it has been suggested as a technique for evaluating the robustness of explanations [7]. Given a slight modification of input, and consequently negligibly small effect on the prediction, a robust explanation should not differ drastically compared to those created from the unmodified input. We aim to investigate the (in)stability of existing XAI methods w.r.t. slight modifications of household aggregated consumption signal. Given an explanation function $h(\cdot)$ and input aggregate signal x, we expose the signal to zero-mean Gaussian noise with

controlled standard deviation σ to create modified input aggregate signal, $\hat{x}$. We define local Lipschitz constant estimate as [7]:

$$\hat{L} = \frac{\|h(x) - h(\hat{x})\|}{\|x - \hat{x}\| + \mu}, \tag{3.6}$$

where μ represents a small value added for numerical stability ($\mu = 1e^{-6}$). For validity, we repeat this procedure n times and report the averaged robustness score (RS). Methods with low Lipschitz value scores display a characteristic of being stable under the presence of noise and should be favoured. In the context of NILM-like data it is important to note that we assume bounded input space, i.e., that maximum change in the function value is finite, which can be assumed for NILM signals as the magnitude of the aggregate power signal is bounded.

Property of Complexity

One of the core principles of XAI is to provide human understandable explanations. Previous studies in the area of research focusing on applying XAI in the energy sector have reported mixed results when applying XAI tools to real-world energy data [81]. Yet, none of these studies have delved into the evaluation of explainability methods, particularly the complexity of explanations. We argue that this property is one of the most desirable ones, as it quantifies the entropy of the XAI output. If most of the input features are deemed important, it does not provide an adequate level of clarity and lowers the human interpretability of explanation. To measure the conciseness of explanation output, we measure the statistical dispersion of the output map. The output map is first sorted in ascending order, and indices of the sorted values are determined. Finally, the conciseness of explanation is formulated as a Gini index computation [31]:

$$Gini = \frac{\sum_{a=1}^{\omega}(2a - \omega - 1) \cdot h_a}{\kappa + \omega \cdot \sum_{a=1}^{\omega} h_a}, \tag{3.7}$$

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

Evaluation of explainability is in general a two-step process, where at first an explanation result is generated using an XAI method considering the input of the model and the model itself, followed by the measurement of the desirable property of explanation result. In this sense, explanation sparseness points to the dispersion of the distribution of the output of the XAI method (i.e., the complexity of explanation). However, it disregards information about the complexity of the input variable. We argue that this is highly important for systems that include time-varying data, as the presence of noise is a common phenomenon, and the system's ability to deal with it is of particular interest. Consequently, explanation sparseness in the context of NILM does not reflect one of the most common challenges of working with time-series. One of the existing measures that capture the percentage of noise in data sample, noise-aggregate measure (NAR) [82], is defined as:

$$NAR = \frac{\sum_{i=1}^{T} |y(t) - \sum_{i=1}^{N} x_i(t)|}{\sum_{t=1}^{T} y(t)}. \tag{3.8}$$

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

$$NAR^{(i)} = \sum_{j=1}^{T} \Big| 1 - \frac{x_j(t)}{y(t)} \Big|. \tag{3.9}$$

We observe that the explanation complexity is often similar for inputs with varying degrees of noise. To establish the relationship between the complexity of an input variable and the complexity of explanation, we introduce an additional term to the explanation complexity that reflects the

“noisiness" of the input. Thus, to quantify the complexity of explanation in the context of NILM, we define the “effective complexity" measure as:

$$EC^{(i)} = \frac{Gini}{1 - NAR^{(i)}}. \tag{3.10}$$

3.2 Experimental results: qualitative and quantitative evaluation of explainability

3.2.1 Experimental setup: Datasets and model training

For transparency, we used the most widely used [56] and well documented REFIT [91] and UK-DALE [64] public datasets. These datasets contain real-world active power measurements obtained from residential buildings, exhibiting a realistic spectrum of appliance ownership and usage patterns. To evaluate explainability across appliance activations with different levels of power and activation periods, we focus our attention on popular multi-state and single-state appliances, namely: Washing Machine, Dishwasher, Microwave, and Kettle. The aggregate data were pre-processed using normalization with mean and standard deviation values computed from the training set. All models were trained and evaluated by reproducing the procedure outlined in [138]. Houses were chosen based on the condition that they must contain measurements of all four aforementioned appliances. For UK-DALE, we use houses 1, 3, 4, and 5 for training, while house 2 is used for testing. In the case of REFIT, houses 2, 3, 6, 11, 13, and 15 were used for training, while the test set contains data from house 5.

The explainability dataset is created by randomly sampling 30 days when appliances of interest are running and selecting a window of size ω samples centered around the appliance activation window from each chosen day. Given a dataset with granularity of 8 seconds, ω is determined from the typical operation time of the appliance of interest. For appliances with lengthy

duration, i.e., Washing Machine (WM) and Dishwasher (DW), activation length $\omega = 1024$ is chosen, which represents roughly 2 hours and 15 minutes of measurements, in line with the average length of a duty cycle of most WM and DW devices. For the Microwave (MW), activation length $\omega = 80$ was chosen, which corresponds to around 10 minutes. Finally, Kettle (KT) activation length ω is set at 40, corresponding to around 5 minutes. If the total length of the activation length of interest is larger than ω, the first ω data samples are selected.

3.2.2 Interpretation of Faithfulness, Robustness and Complexity Scores

Faithfulness is of particular importance to an algorithm designer, as it facilitates understanding of how feature importance scores influence the prediction. Conversely, robustness provides an indication of the change in prediction if the input to the DL model changed slightly (e.g., due to appliance model fluctuations, appliance settings and influence of unknown appliances), which is a crucial indicator of scalability. Finally, complexity reflects the human comprehensibility of the visualization. The relative significance of each score is determined by the use-case, i.e., which property is most desirable to an algorithm designer, system developer, consumer or technology enthusiast. Explainability scores (see Subsec. 3.1.1, 3.1.2, 3.1.3) obtained for four different appliances are presented in Tables 3.1 and 3.2, for the UK-DALE and REFIT datasets, respectively. Regression (R) and Classification (C) scores are calculated as the AUC for MAE and F1 scores, as described in Sec. 3.1.2. For long duration appliances (WM and DW), we perform 75 perturbation steps, while for MW and KT we perform 10 and 5 steps, respectively. To calculate the sorted, non-overlapping segments of importance (as per 3.1.2), for appliances with a long activation period, we choose segments containing 40s of measurements, while other appliances contain 24s of measurement. High faithfulness score indicates that the explainability method is able to

correctly identify the important features of the input signal, thus leading to a large drop in prediction accuracy after perturbation. The Robustness score is calculated as mean and standard deviation of $n = 35$ computations of Lipschitz constant estimate, defined in Eq. 3.6, where μ and σ values of Gaussian noise are 0 and 0.1, respectively. Low robustness score indicates the ability of the explainability method to deal with noise. The Effective complexity is calculated as per Eq. 3.10. High effective complexity suggests that the explainability method is able to generate explanations that are concise and human understandable.

Table 3.1
Comparison of explainability and predictive performance of seq2point model for UK-DALE dataset.

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

* RF: Regression Faithfulness, CF: Classification Faithfulness, EC: Effective Complexity

Tables 3.1 and 3.2 suggest that LRP-ϵ achieved the most success across

Table 3.2
Comparison of explainability and predictive performance of seq2point model for REFIT dataset

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

* RF: Regression Faithfulness, CF: Classification Faithfulness, EC: Effective Complexity

the proposed properties that explainable NILM systems based on sequence-to-point learning should satisfy. This can largely be attributed to the ability to deal with gradient noise as the relevance is propagated through the layers of the network. We report a strong relationship between the choice of parameter ϵ and the results in performance metrics, where ϵ value should be guided by the noisiness of the dataset. As the REFIT dataset is known to be significantly noisier than UK-DALE, we set the parameter ϵ to be a large value ($\epsilon = 1$) compared to UK-DALE ($\epsilon = 0.1$). Contrary to previous studies in the energy sector that recommended GradCAM as the best XAI method [81], our analysis indicates that GradCAM is not the ideal XAI

approach for time-series NILM applications employing sequence-to-point architectures. Notably, GradCAM's faithfulness scores for dishwashers were significantly lower compared to other methods, implying an inability to identify crucial signal features. This observation is further supported by Fig. 3.2 and the results for the noisier REFIT dataset in Table 3.2, where faithfulness scores for both WM and DW were unsatisfactory. In an attempt to improve the score, we explored guided gradient technique used for GradCAM, but our findings point to further degradation of performance. On the other hand, our findings reveal that GradCAM++ method does outperform the original GradCAM, achieving better faithfulness and robustness. However, while the results demonstrate significant enhancements of GradCAM++ over GradCAM in these two aspects, the complexity of explanations generated by Grad-CAM++ is observed to be less than ideal. This finding suggests that the enhancements in faithfulness and robustness of GradCAM++ may come at the cost of increased complexity. Intriguingly, IG exhibited excellent performance for the complex signals (i.e., WM and DW) within the REFIT dataset. This implies that a zero signal is an appropriate choice for the baseline value of the IG algorithm for NILM-like data. Meanwhile, SmoothGrad produced robust results across most scenarios due the nature of the algorithm.

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

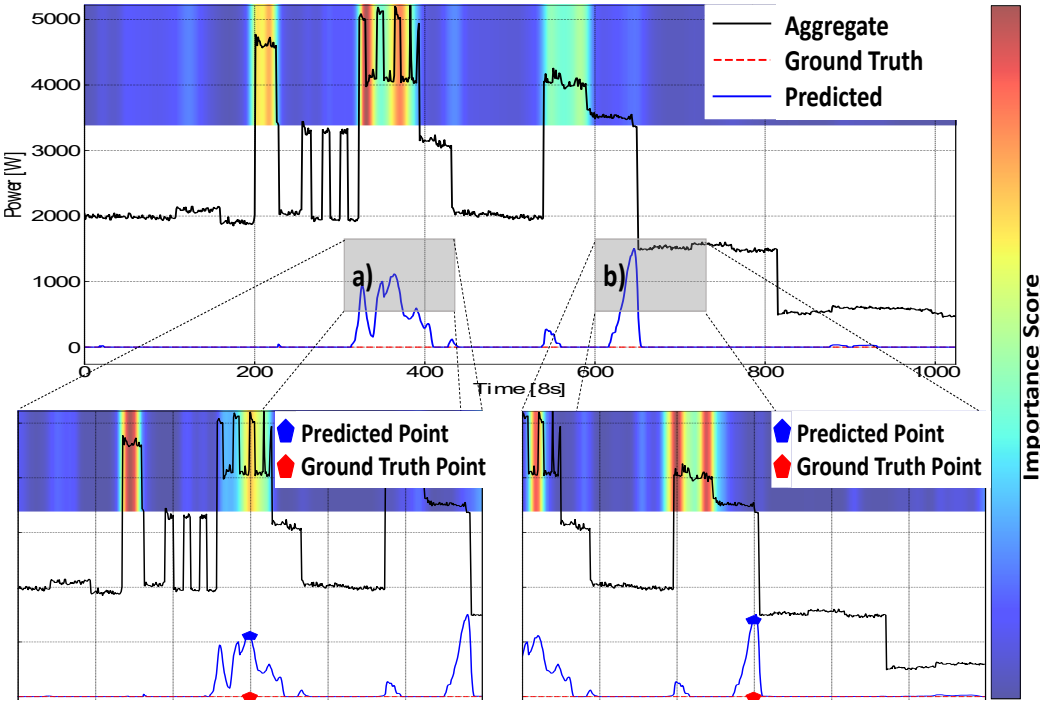


Fig. 3.3. Visual outline of the proposed approach showcasing an example of false positive prediction of washing machine for UK-DALE dataset, and the derived explanations using LRP. Two levels of explainability provide general, sequence-level (top image), and specific, point-level explanations (a and b), under a test scenario of signal incorrectly predicted as a washing machine.

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

function is bounded.

3.2.3 Visualisation via heatmaps

The proposed approach enables two levels of explainability. On one hand, point-level explainability provides visual understanding of how a prediction of a single time step was made. It is specific to a point of reference. On the other hand, the visualization algorithm generates another, sequence-level explanation, showcasing the aggregate importance of the input signal for the prediction of the output, and acting as a more general representation of the importance. Point-level explanation is preferred to illuminate the features that have contributed to an individual point of the prediction especially if that point prediction is an outlier. Sequence-level explanations are more appropriate if trying to comprehend the decision on inference of a complete appliance duty cycle, such as why a time-series sequence was predicted as a Washing Machine. A key aspect of a good explanation is the ability to align with human intuition. For NILM, an end-user or a domain expert might expect explanations to highlight periods of significant power change corresponding to appliance turn-on/off events, or stable consumption periods characteristic of a continuously running appliance (e.g., a refrigerator). The heatmaps generated (as shown in Fig. 3.2 and 3.3) offer a visual medium for such subjective assessment. For instance, an explanation that consistently highlights irrelevant noise or assigns high importance to periods where the target appliance is clearly off would be deemed counter-intuitive and unhelpful, regardless of its quantitative scores on certain metrics.

Our visualization approach offers several advantages over the previously proposed methods. We tackle the more challenging regression scenario for the NILM problem compared to earlier work, which utilized a multi-class CNN for the simpler task of detecting appliance presence without recognizing on/off states [80]. Moreover, our method has been rigorously validated on numerous real-world datasets, demonstrating its adaptability and gen-

eralizability across diverse contexts. Unlike previous work that relied on a single dataset, our approach handles varied energy consumption patterns and appliance configurations, ensuring its practicality and resilience. In comparison to the regression-based XAI visualization method in NILM [90], our approach is more computationally efficient, as gradient-based methods require fewer iterations and calculations than occlusion sensitivity, making them well-suited for real-time applications and large-scale datasets. Additionally, our approach avoids the introduction of out-of-distribution scenarios caused by setting parts of the input signal to zero, ensuring that the generated explanations are more faithful to the model's behavior. A key strength of our method lies in its ability to provide multi-temporal explanations, offering insights into both local and global patterns at various levels of granularity, such as point-level and sequence-level explanations. This enhanced interpretability facilitates a better understanding of the NILM model's decision-making process and allows users to make more informed decisions based on the model's output. Furthermore, the gradient-based XAI methods can be applied to a wider range of DL-based NILM algorithms.

Fig. 3.2 provides an example of point explanations for a Dishwasher signal prediction from the UK-DALE dataset. This is a true positive prediction where the primary features contributing to the prediction of the middle point (marked with a blue pentagon) are displayed in a form of heatmap. We observe that most XAI methods highlight the true positive part of the input signal. However, different XAI methods produce varying heatmap visualizations, underscoring the necessity for their quantitative quality evaluation. Comparing the results in Fig. 3.2 with the results displayed in Table 3.1 and Table 3.2, LRP and SmoothGrad indeed showcase the best performance. We observe that both heatmaps highlight the truly important parts of the signal, suggesting high faithfulness, and that explanations are concise, pointing to low complexity. On the other hand, GradCAM shows the lowest faithfulness score, which can be observed from Fig. 3.2 as the GradCAM visualised

explanation highlights an area that is not related to high activity of the dishwasher signal, suggesting a case of instability. To a smaller extent, this phenomenon is also observed in the case of IG. While the localization of feature importance scores in GradCAM++ improved compared to GradCAM, we observe a higher complexity of generated explanation. Comparing to LRP and SmoothGrad, we observe that the explanation heatmaps of GradCAM, GradCAM++, and IG cover a larger area of the input signal, and are of noticeably higher complexity, which is a finding that is reinforced by the complexity evaluation scores. Another scenario showcasing the mechanism behind a false positive prediction of a NILM DL model is presented in Fig. 3.3. In this example, a DW signature is incorrectly predicted as WM. We observe that the general explanation (on the top) enables us to assign the importance scores to the areas of the signal that the network deemed as indicative of a WM duty cycle. Looking further, the point-level explanations (a and b) enable us to understand that the DL model recognizes that there may be multiple cycles in a typical WM signature, which is supported by high importance score assigned to past signal spikes that look similar to a WM duty cycle. This can help the algorithm designer to improve the training and tuning process or adopt a multi-classification approach to better distinguish these multistate appliances with similar power level, duty cycle and duration.

3.3 Summary

This chapter proposes a methodology for determining the explainability of a time-series DNN regression NILM problem. Specifically, we propose visualization via heatmaps approach by integrating XAI methods into the DL NILM and quantify explainability via faithfulness, robustness and complexity scores. As a way of overcoming the problem of transparency inherent to DL algorithms, the proposed approach provides a dual mode of explainability, one at a general, sequence level, and other at a specific, point level.

Both levels of explainability can be used interchangeably based on the use case, as they provide varying degrees of specificity, i.e., they can deal with different scenarios when the decisions of NILM systems are unclear or difficult to explain. We show that this can be achieved without changing the architecture of the model. Furthermore, we define the core properties that should be considered when designing explainable NILM systems, and provide a strategy for quantitative evaluation of their explainability. We show that XAI methods, such as LRP, that have an inherent ability of dealing with noise, can lead to explanations that satisfy properties of being faithful to the performance of the model, robust to slight changes of input, and offer unambiguous interpretation of resulting heatmaps. The choice of the most appropriate methods should be guided by the target user of explanation, be it a domain expert, researcher, or customer, considering the trade-off between the aforementioned properties. By using the proposed method, the diverse set of needs of various users of the system can be satisfied, while maintaining the predictive performance and facilitating trust in the NILM system deployed in a real-world scenario.

Chapter 4

Explainability Informed Training Enhancement for Load Disaggregation

Previous studies in the field of 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 [20]. As a result, more recent literature has emphasized the need for a mathematical definition of explanation quality and evaluation of feature attribution methods [8, 9, 20]. XAI approaches for NILM are still in their infancy, with limited literature available [80, 89, 90]. 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. Additionally, explanation heatmaps show what features were important, but

they do not inherently explain why the model learned to consider those features important, or why its internal decision thresholds are set the way they are. 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 [92], embedding human oversight through inclusion of user or expert knowledge in the learning process [126], or XAI methods for transparency [16, 18, 89]. 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 chapter, we propose the first explainability-informed NILM training framework for low-frequency NILM. The content of this chapter is taken from the work reported in "XNILM-Boost: Explainability-Informed Load Disaggregation Training Enhancement using Attribution Priors" [17].

The proposed framework aims to directly mitigate shortcomings of existing NILM approaches in line with EU guidelines for Trustworthy AI, by prioritising 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 generalise 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, using quantitative metrics presented in Chapter 3.

- 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-

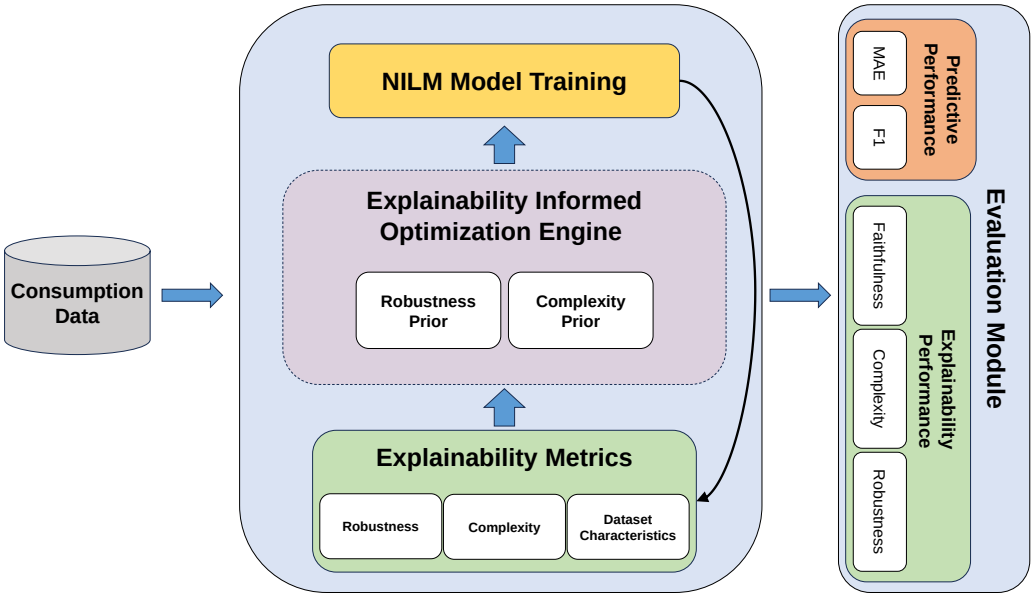


Fig. 4.1. Overview of the proposed explainability-informed NILM training framework.

world measurements from households in the UK, United States, and Greece.

4.1 Methodology

Our explainability-informed learning framework for low-frequency NILM is shown in Fig. 4.1. The backbone of our approach is the explainability-informed optimisation engine, which is responsible for the optimisation 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 generalisability 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.

4.1.1 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 seconds, ω is determined from the typical operation time of the appliance of interest. For appliances with lengthy duration, i.e., Washing Machine (WM) and Dishwasher (DW), activation length $\omega = 1024$ is chosen, which represents roughly 2 hours and 15 minutes of measurements, in line with the average length of a duty cycle of most WM and DW devices. For the Microwave (MW), activation length $\omega = 80$ samples was chosen, which corresponds to around 10 minutes. 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 second granularity measurements, activation length of WM appliance is set to $\omega = 820$, while Boiler and AC appliances have activation length set to

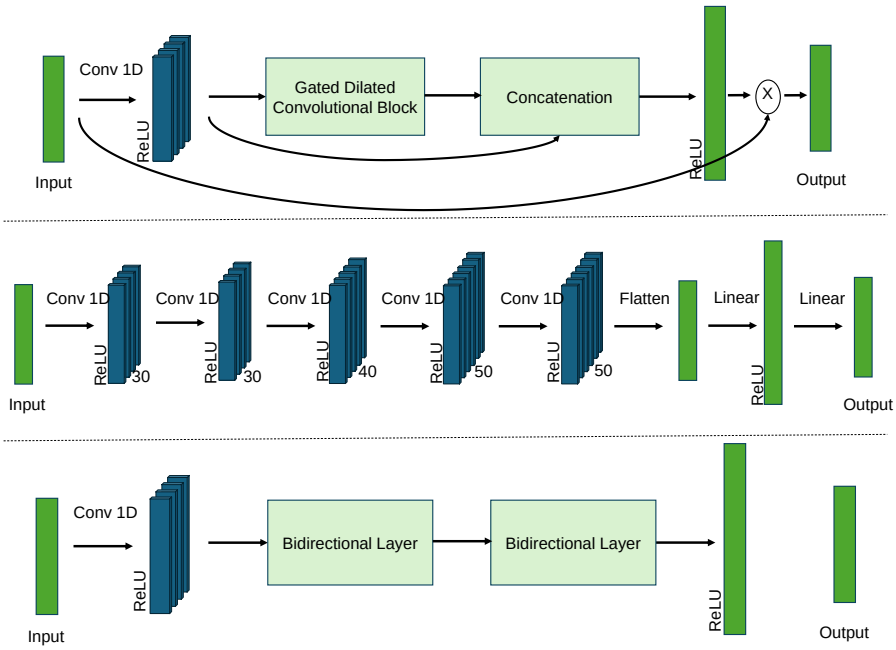


Fig. 4.2. Model architecture for the four NILM models used in this chapter. The upper subfigure describes a WaveNet network [51], whereas the middle and bottom subfigures indicate CNN [138] and GRU [103] architectures, respectively.

$\omega = 700$ and $\omega = 1000$, respectively.

4.1.2 Low-frequency NILM Algorithms

For the purpose of demonstrating the adaptability and generalisability 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 [138], GRU-based [103], and WaveNet-based [51] NILM networks. One of the most cited CNN-based approaches for NILM is seq2point architecture [138]. 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 [61]. 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 [51]. 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 [138], [103], and [51].

4.1.3 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 optimise the explainability performance of attribution methods used for visualisation 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 characterised 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:

$$f = argmin_\theta \frac{1}{n} \mathcal{L}(\theta; x, y) + \alpha \mathcal{R}(\theta), \tag{4.1}$$

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 $p : \mathcal{R} \to \mathcal{R}$. 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:

$$f = argmin_\theta \frac{1}{n} \mathcal{L}(\theta; x, y) + \alpha \mathcal{R}(\theta) + \beta p(m(\theta, x)), \tag{4.2}$$

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 [31] 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:

$$p(m(\theta, x)) = \frac{\sum_{a=1}^{\omega}(2a - \omega - 1)m(\theta, x)}{k + \sum_{a=1}^{\omega} m(\theta, x)}, \tag{4.3}$$

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:

$$p(m(\theta, x)) = \sum_{i} |m_{i+1}(\theta, x) - m_i(\theta, x)|. \tag{4.4}$$

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. Lastly, it is important to note that albeit faithfulness property

Table 4.1
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**

introduced in Chapter 3 is an important metric to measure the quality of assigned attributions, it is not a suitable attribution prior due to a high computational overhead related to iterative obfuscation of input features.

4.1.4 Explainability-informed Training

Finding the optimal attribution prior that represents the best trade-off 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 metric (ROTR). This approach enables us to iteratively determine the optimal prior for a given NILM model while considering multiple performance aspects simultaneously. Instead of evaluat-

Table 4.2
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**

ing 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:

$$ROTR = \frac{XF_{prior}}{XF_{base}} \frac{XR_{base}}{XR_{prior}} \frac{XC_{prior}}{XC_{base}}, \tag{4.5}$$

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 [18]. To thoroughly explore the trade-offs between the smoothness and low com-

Table 4.3
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**

plexity 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 Sub-

section 4.1.1 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.

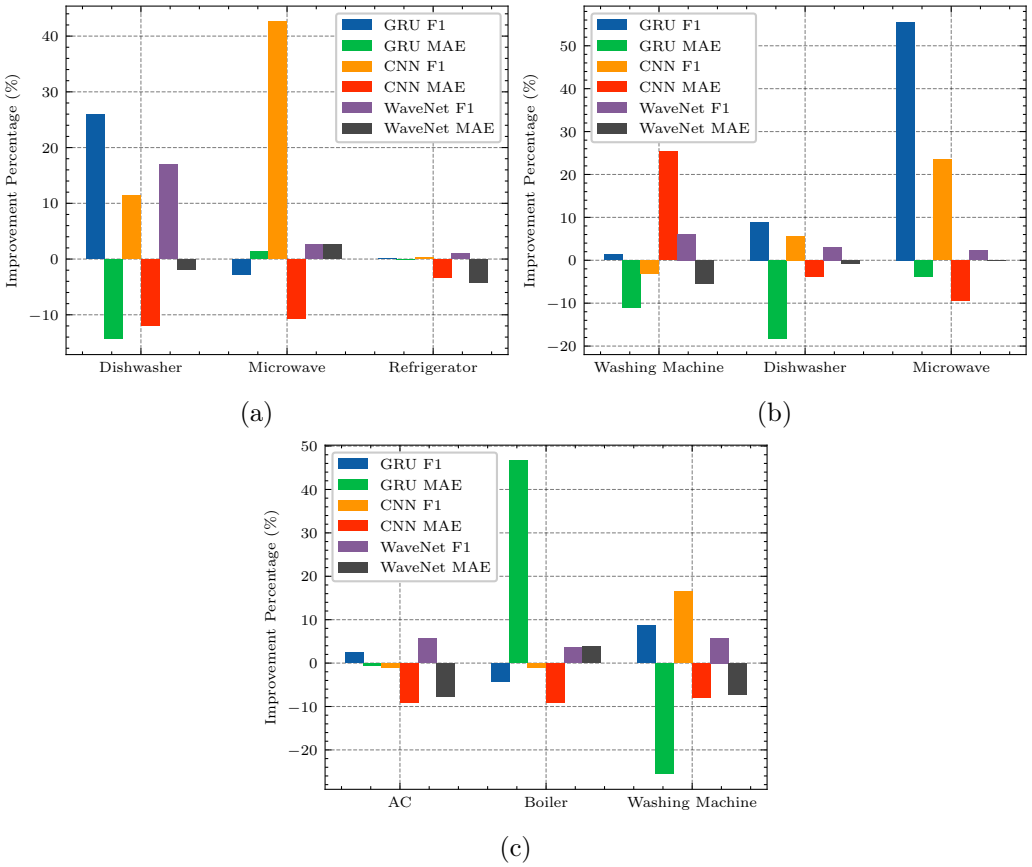


Fig. 4.3. Comparison of relative F1 and MAE performance improvement after explainability-informed training for GRU, CNN, and WaveNet architectures for (a) REDD dataset (b) UK-DALE dataset, and (c) Plegma dataset.

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.

4.1.5 Explainability Methods

To quantify the explainability performance of the networks used in this chapter, we adapt the explainability methods and evaluation methodology described in Chapter 3. To accommodate to different architectures used in this chapter, the visualization procedure is modified for GRU and WaveNet

Table 4.4
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

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 chapter, we utilize 4 popular XAI methods: GradCAM [109], GradCAM++ [34], IntegratedGradients [118], and SmoothGrad [113]. 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 Subsection 3.1.2-faithfulness, robustness, and low complexity.

4.2 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.

4.2.1 Datasets and Appliances

To evaluate our approach, we conducted experiments on appliances from UK-DALE [64], REDD [70] and Plegma [14] 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 6s, 1s, and 10s 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 8s resolution, while Plegma kept the original resolution of 10s. Detailed dataset characteristics and selection of houses for training data is described in Table 4.4. 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 normalised 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.

4.2.2 Model architectures and training

To enhance the generalisability and robustness of our proposed framework, we base our evaluation on three distinct NILM model architectures: a convolutional network [138], a recurrent network [103], and a WaveNet neural

network [51] network, as illustrated in Fig. 4.2. Recurrent 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 [138], [103], and [51].

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.

4.2.3 Computational Complexity

For the purpose of performing the experiments, a PC with the following specifications is used: Intel(R) Core(TM) i9-10980XE CPU @ 3.00GHz, 258GB RAM, and two NVIDIA GeForce RTX 3080 GPUs. When analyzing the computational complexity of our framework across various architectures, we observed that incorporating priors consistently affects training speed. Specifically, the recurrent architecture experiences the most substantial impact, with training time increasing by 50% compared to the baseline model without priors. The convolutional architecture exhibits a 39% increase in training duration, whereas the dilated causal network shows a 32% increase. These differences in computational overhead result from the additional calculations required for priors and their interaction with the distinctive structural characteristics of each architecture. The recurrent network's higher computational cost may be due to the sequential processing nature of recurrent architectures. 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.

4.2.4 Evaluation Metrics

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

$$MAE = \frac{1}{T} \cdot \sum_{i=1}^{T} |\hat{E}_i - E_i|. \tag{4.6}$$

Whilst MAE is the most common measure for evaluating regression or disaggregation performance [56], the F1-score measure is typically used in the NILM literature to evaluate the classification performance [10]. To generate events from the regression output, we apply a threshold, as explained in Subsection 4.1.1. 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:

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}, \tag{4.7}$$

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 [16], 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 Lipschitz metric defined in [16]. 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 [7]:

$$\hat{L} = \frac{\|m(\theta, x) - m(\theta, \hat{x})\|}{\|x - \hat{x}\| + \mu}, \tag{4.8}$$

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 favoured.

Lastly, to measure the overall ease of understanding the produced explanation, an effective complexity measure is used, as described in Chapter 3. 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 [82]:

$$EC^{(i)} = \frac{Gini}{1 - NAR^{(i)}}. \tag{4.9}$$

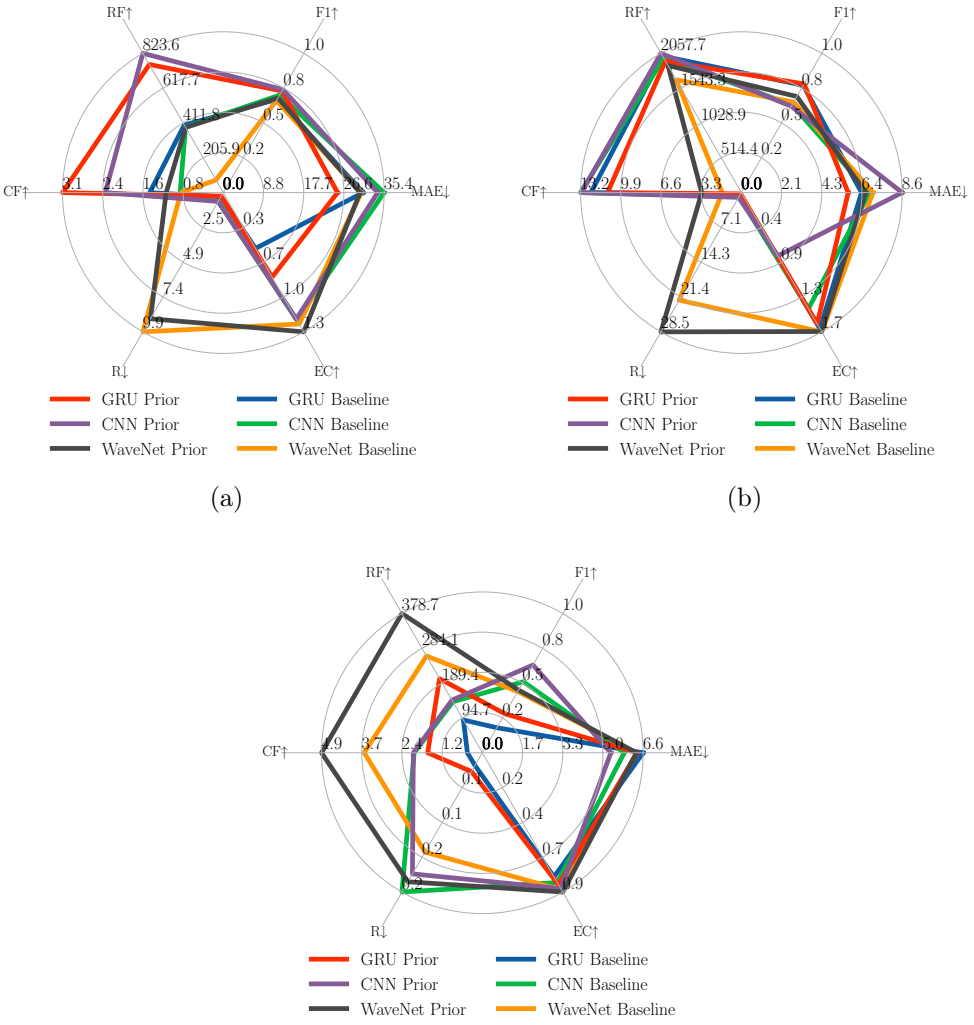


Fig. 4.4. Performance evaluation of the proposed XNILMBoost method for training of a) UK-DALE Dishwasher, b) UK-DALE Washing Machine, c) UK-DALE Microwave. The radar plot axes are scaled based on the maximum values of the respective category. The arrows indicate if higher or lower value is better.

4.2.5 Experimental Results and Discussion

Does training for better explanations lead to improved predictive performance?

The first experimental analysis is designed to examine if explainability-informed training can lead to improved model performance, instead of the often argued conjecture of trading-off between explainablity and accuracy [36]. As can be seen in Tables 4.1 and 4.2, 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. 4.3 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. 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 im-

provement 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 behaviour is added complexity of introducing explainability due to large number of dilated causal convolutions. Analyzing the Plegma dataset results (Table 4.3), 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 improv-

ing 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. Lastly, it is important to note that very strong robustness prior might lead to extremely smooth explanation heatmaps. While highly robust to noise, these smooth explanations could obscure fine-grained, sharp temporal features that are genuinely important for distinguishing certain appliance states, potentially reducing faithfulness or even slightly degrading predictive accuracy for appliances reliant on such subtle cues. Additionally, an overly aggressive low-complexity prior might force the model to rely on an extremely sparse set of features. While yielding very simple explanations, this could lead to the model ignoring less dominant but still relevant contextual information, potentially impacting its ability to handle nuanced scenarios and decreasing faithfulness to the model's complex decision process or even its predictive power.

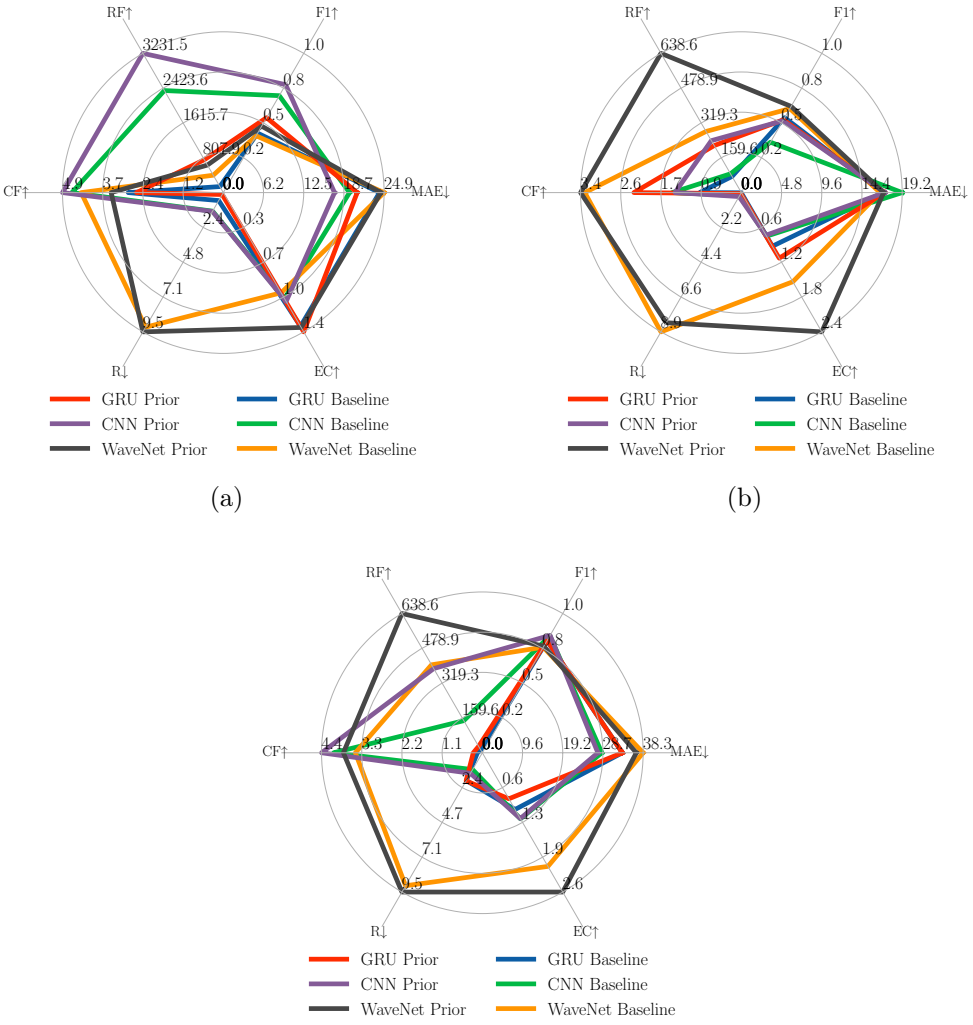


Fig. 4.5. Performance evaluation of the proposed XNILMBoost method for training of a) REDD Microwave, b) REDD Washing Machine, c) REDD Refrigerator. The radar plot axes are scaled based on the maximum values of the respective category. The arrows indicate if higher or lower value is better.

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 4.8-4.9 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, 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 [8, 9, 18]. 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 [18]. In terms of Robustness metric, we observe 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 gradient smoothness and complexity leads to better overall explainability of the NILM system. The Plegma dataset results, as shown in Tables 4.11-4.13, 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 relative improvement in individual explainability metrics, as can be further seen in Fig. 4.5, 4.4 and 4.6. 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.

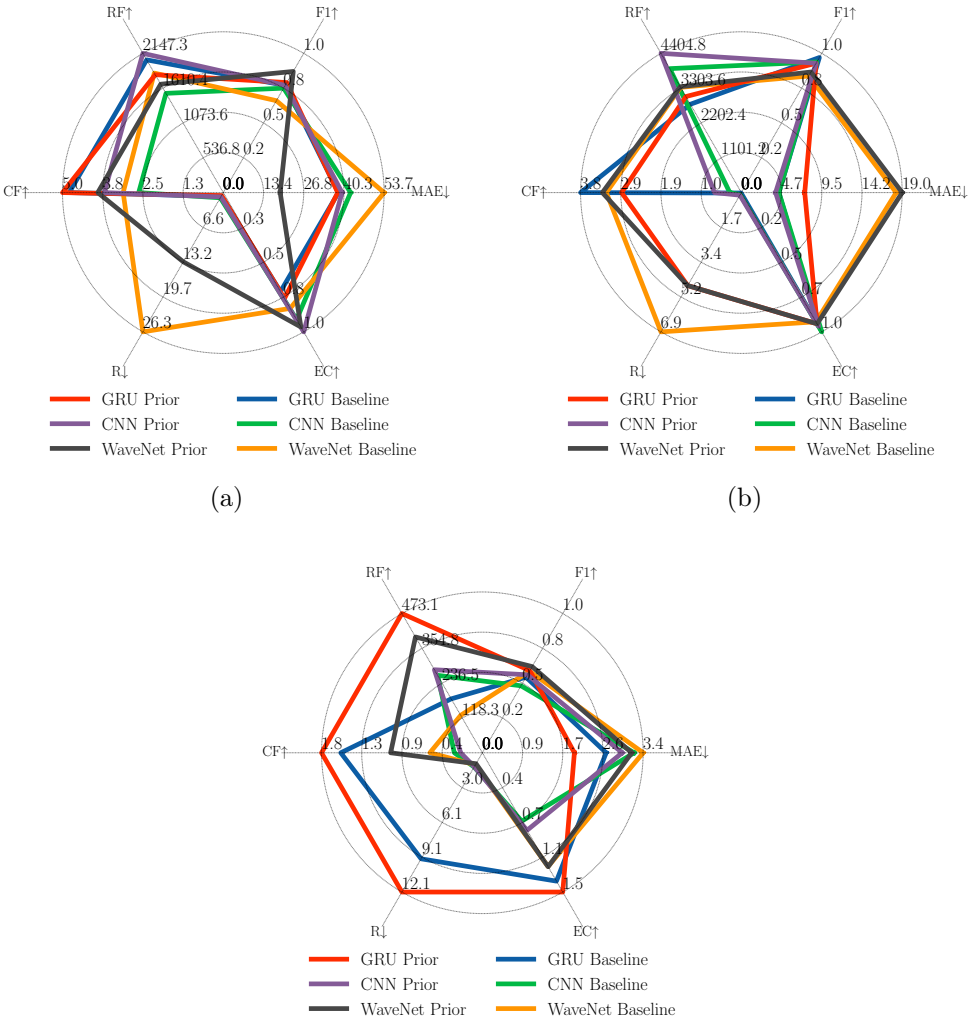


Fig. 4.6. Performance evaluation of the proposed XNILMBoost method for training of a) Plegma AC, b) Plegma Boiler, c) Plegma Washing Machine. The radar plot axes are scaled based on the maximum values of the respective category. The arrows indicate if higher or lower value is better.

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 Fig. 4.5, 4.4 and 4.6. 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 favoured. 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. 4.6 (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.

Table 4.5
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 4.6
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**

Table 4.7
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 4.8
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**

Table 4.9
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 4.10
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**

Table 4.11
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 4.12
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 4.13
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

4.3 Summary

In this chapter, 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 optimisation 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, 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++, IG, and SmoothGrad. Experimental results suggest that in the context of NILM, explainability methods that are design to deal with noise, such as IG 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.

Chapter 5

Explainability on the Edge through Explainability Guided Learning

Building upon the ideas of perception aligned gradients and evaluation of explainability, in this chapter we try to answer a question: Can edge deployed systems benefit from explanability enhancement? To achieve this, a novel idea of explainability guided knowledge distillation is proposed. The core intuition is can we condition a student model to imitate not only the predictions of a teacher, but also localization of it's explainability heatmaps. Additionally, we ask a question of can more explainable teachers lead to downstream improvement of explanability of the student model. In recent years, knowledge distillation [53] emerged as an effective approach to address concerns of resource-heavy computational complexity, generalisability, privacy and bandwidth requirements. This technique involves transferring knowledge from a large, complex model (Teacher) to a smaller, more computationally-efficient model (Student), effectively balancing computational efficiency with performance. This reduction in computational complexity directly addresses the energy efficiency conflict highlighted by [21], as the distilled models require

substantially less energy for inference tasks. The Student model, while maintaining comparable accuracy to its Teacher counterpart, requires significantly fewer parameters and computational resources, making it more suitable for deployment on edge devices. Moreover, KD enhances model generalization ability, as the distillation process often acts as a form of regularization, helping the Student models learn more robust and generalizable features from the Teacher's knowledge [60, 123]. However, current KD approaches primarily focus on optimizing Student performance metrics without adequately addressing the quality and interpretability of the transferred knowledge, which remains a crucial barrier to deploying reliable and trustworthy edge AI systems. The interpretability challenge in edge AI systems is particularly significant as these systems often operate in critical applications where understanding model decisions is essential for user trust and system reliability. When deploying compressed models through KD, maintaining interpretability becomes even more complex as the knowledge transfer process itself may introduce additional lack of transparency in decision-making mechanisms. This creates a compound challenge: not only we must ensure efficient model compression for edge deployment, but we must also maintain or enhance the interpretability and reliability of the resulting systems.

NILM serves as an exemplary case study of these challenges. Through disaggregation of aggregate metered energy consumption into individual appliance-level usage patterns, NILM provides granular insight into energy consumption without costly sub-metering, thus empowering consumers to make informed decisions about their energy usage and enabling utilities to implement more effective demand-response programs [102]. This application is particularly relevant as it encompasses the key challenges of edge AI deployment: it requires real-time processing of continuous data streams, demands high accuracy for appliance load disaggregation, and operates in diverse environmental conditions that rely on model generalisability. Furthermore, personal smart meter data never leaves the premises, thus preserving the privacy of

the household. To address the aforementioned concerns, we propose a robust methodology focused on ensuring interpretability and reliability, by jointly focusing on improvements in interpretability and reliability of KD mechanism learned from the Teacher. We evaluate our method in a domain adaptation scenario that mimics real-world NILM deployment scenario by identifying activation state of five common appliances in residential buildings and performing rigorous evaluation of predictive and explainability performance of the NILM models.

The content of this chapter is taken from the work reported in "Improving knowledge distillation for non-intrusive load monitoring through explainability guided learning" [18] and "Interpretability and Reliability-driven Knowledge Distillation for Non-intrusive Load Monitoring on the Edge" which is currently under review. Interpretability and Reliability-driven Knowledge Distillation (IR-KD) implements a structured approach for training individual student networks for each appliance, leveraging a complex multi-label Teacher classifier trained on diverse, cross-domain datasets. Our framework introduces three key novelties to enhance the knowledge distillation process. First, we identify the main type of inconsistency in transfer of explainable knowledge, and propose explainability guided learning that aims to alleviate erroneous knowledge transfer during the distillation process. Second, we address the interpretability challenge by proposing a regularization technique based on perception-aligned gradients, which ensures the model's decision-making process closely mirrors human intuition. This alignment is particularly crucial in NILM applications, where the model must identify interpretable features such as characteristic power consumption patterns and temporal usage signatures that energy experts and consumers can readily understand. Third, we address the challenge of incorrect knowledge transfer through the introduction of a specialized loss term that optimizes Teacher training and subsequent knowledge distillation. This approach leverages hidden information extracted from weak labels [123], significantly improving the

quality and reliability of knowledge transferred to Student models. By focusing on refining the Teacher's knowledge directly, our method achieves computational efficiency while avoiding the time-consuming and often ineffective process of correcting Student models post-distillation. Lastly, our evaluation methodology encompasses both traditional performance metrics and NILM-specific explainability measures. We assess model robustness through predictive performance analysis and employ specialized metrics to quantify explainability improvements. These metrics evaluate the model's capability to identify salient input features, maintain resilience against input perturbations, and generate simplified, interpretable feature attribution maps.

The key contributions of this work can be summarized as follows:

- The main type of inconsistency in the process of transfer of explanation knowledge in the KD framework for NILM is identified.
- A technique for alleviation of explanation inconsistencies in KD NILM via a new, explainability guided loss function.
- A KD framework that specifically addresses Teacher limitations to ensure both reliability and interpretability in computationally efficient Student models.
- A perception-aligned gradients approach that enhances model interpretability by aligning the decision-making process with human understanding, while simultaneously improving the capture of robust and transferable features.
- A learning strategy that exploits hidden information from weak labels, enabling both Teacher and Student models to capture more reliable and transferable features from locally collected data, thereby enhancing state classification performance across new environments.

These contributions collectively advance the field of model compression and knowledge transfer, exemplified by NILM applications, that is robust,

interpretable, and computationally efficient. Our approach not only improves model performance, but also ensures greater reliability and interpretability in practical applications, addressing the concerns of trustworthy approaches to NILM. The overall framework of the proposed approach is presented in the following section.

5.1 Methodology

This section details our proposed IR-KD framework, which incorporates weakly supervised explainability-guided learning, perception-aligned gradients, and zero-strong labels in the distillation framework, as illustrated in Fig. 5.1. We start with defining the NILM problem, followed by an in-depth discussion of the theoretical foundations and practical implementation of the newly introduced components.

5.1.1 Distillation Framework

The proposed framework employs a weakly supervised approach [121, 122] that requires only weak labels from the target environment, significantly reducing the labeling burden. Weak labels are proposed in [121, 122] as a trade-off to lighten the annotation effort but still providing supervision during training and achieving results comparable to fully supervised approaches.

The core idea of weakly supervised learning is to train the DNN by only using coarse information collected in the target environment. This approach reduces the labeling effort while retaining sufficient information to refine the model's performance for the target environment. Thus, the framework aims to estimate the precise state of each appliance $s_n(t)$ on a sample-by-sample basis, despite using coarse-grained supervision.

Raw samples of the aggregate signal $y(t)$ are referred to as instances, and

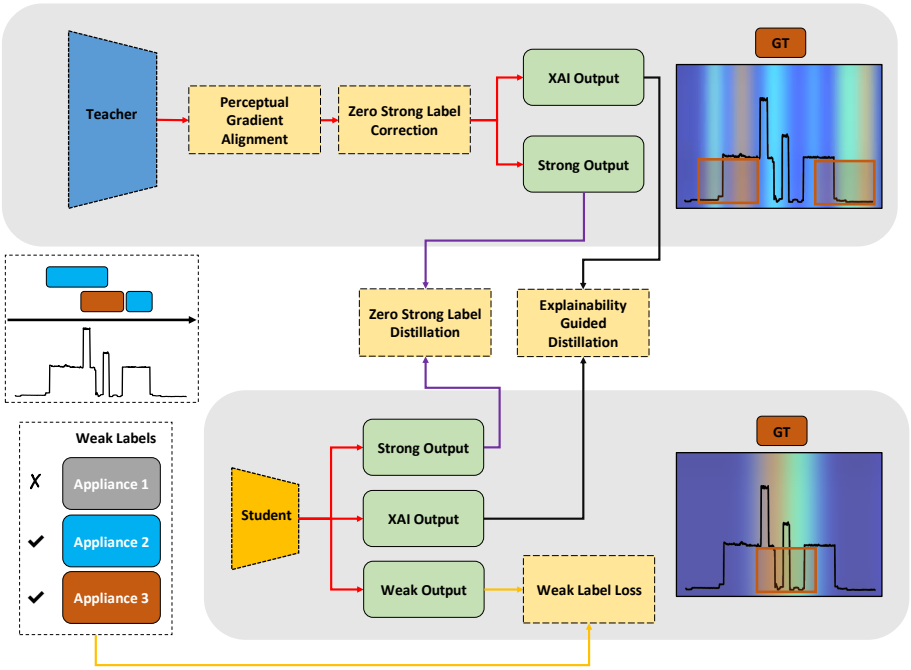


Fig. 5.1. Interpretability and Reliability-guided Knowledge Distillation (IR-KD) framework. Teacher fine-tuning and Student distillation are depicted. The data available for training are annotated with weak labels that specify only if an appliance is active or not inside a certain aggregate window. GT represents ground truth labels. The associated heatmap represents the outputs of XAI visualization, where colors closer to red represent areas of high importance for the prediction.

the related labels are represented by one-hot vectors $\mathbf{s}(t) \in \mathbb{R}^{N\times 1}$ defined as:

$$\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_N(t)]^T. \tag{5.1}$$

A segment of $y(t)$ with a length L is referred to as a window. Assuming that $y(t)$ is divided into disjointed segments, the j-th window is represented by the following vector:

$$\mathbf{y}_j = [y(jL), \ldots, y(jL + L - 1)]^T \in \mathbb{R}^{L\times 1}. \tag{5.2}$$

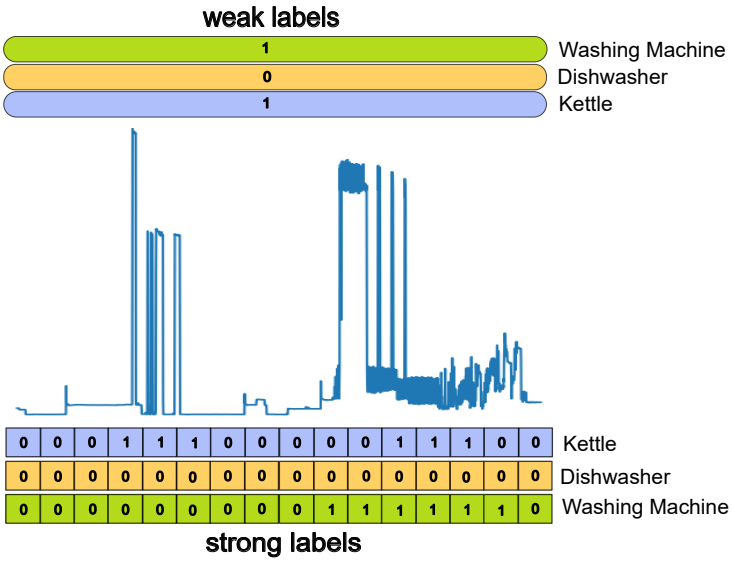


Fig. 5.2. Strong and weak labels graphical representation. When the weak label is zero (as for the dishwasher) each strong label associated is zero as well. Instead, if the weak label is equal to one, some strong labels are ones, other zeros.

The related weak label is again encoded as a one-hot vector $\mathbf{w}_j \in \mathbb{R}^{N\times 1}$. Notably, $\mathbf{w}_j$ depends on the instance labels within the segment. The assignment of the labels is based on the presence or not of an activation inside a certain aggregate window. Thus, if for each $y_j(t)$ of the j-th window the n-th appliance is never active (all $s_n(t)$), the label w_n provided by the annotator from the target environment is equal to 0. On the contrary, if the appliance has been active at least for one $s_n(t)$, the w_n provided by the annotator is equal to 1. For clarity, a real example is depicted in Fig. 5.2.

Denoting with $\mathbf{S}_j = [\mathbf{s}(jL), \mathbf{s}(jL+1), \ldots, \mathbf{s}(jL+L-1)] \in \mathbb{R}^{N\times L}$ the set of instance labels related to segment j, the mathematical relationship can be represented by a pooling function $\mathbf{b} : \mathbb{R}^{N\times L} \to \mathbb{R}^N$, such that:

$$\mathbf{w}_j = \mathbf{b}(\mathbf{S}_j). \tag{5.3}$$

The Teacher network initiates the KD process by training on extensive

data from public consumption datasets, establishing a robust foundation of transferable knowledge and generalisability. Then, due to the large differences that occur between different domains (houses and different countries in terms of energy load profiles), a fine-tuning phase is necessary to ensure good performance on the target domain. For this fine-tuning process, only weak labels are assumed to be collected from the target environment, to lighten the annotation process and reproduce the realistic scenario whereby annotated data for particular domains is not available in the absence of submetering or manual offline annotation [123]. Then, the fine-tuning loss L_{ft} has the form of Binary Cross-Entropy Loss (BCE):

$$L_w = -\frac{1}{N}\sum_{n=1}^{N}\left[w_n \log(\hat{w}_n) + (1-w_n)\log(1-\hat{w}_n)\right], \tag{5.4}$$

where the segment index j has been omitted for simplicity of notation. $\hat{w}_n$ is the weak prediction for the n-th appliance and w_n is the related weak label. Once the Teacher is ready to transfer its acquired knowledge to the Student, the following distillation loss is used to train the Student network:

$$L_{dist} = \beta L_{soft}\left(\sigma\left(\frac{\mathbf{Z}_j^{st}}{T}\right), \sigma\left(\frac{\mathbf{Z}_j^{te}}{T}\right)\right) + (1-\beta)\theta(e)L_w(\hat{\mathbf{w}}_j^{st}, \mathbf{w}_j), \tag{5.5}$$

where T is the temperature parameter always included to obtain soft Teacher and Student logits $\mathbf{Z}_j^{st}$ and $\mathbf{Z}_j^{te}$ before applying the sigmoid activation function $\sigma(\cdot)$. Both L_{soft} and L_w are expressed using the binary cross-entropy function and L_{soft} is the loss related to the soft strong-level output of the Teacher and the Student. The parameter β balances the contributions of L_{soft} and L_w during the distillation, while $\theta(e)$ adjusts the relative magnitude of the losses at each training epoch e.

5.1.2 Explainability Guided Learning

KD minimizes the divergence between the probability distributions of the Teacher and Student models [53], with the aim of aligning the logits produced by the Student with those of the Teacher. This process achieves effective transfer of knowledge by conditioning the Student model to mimic the outputs of the Teacher. However, we observe that KD might not always be successful in transferring the explainable knowledge of the Teacher. In particular, we note the main erroneous case of inconsistency in the explanation knowledge transfer, that is, given identical inputs, Teacher and Student networks produce dissimilar output explanations for a given class. This phenomenon is illustrated with an example in Fig. 5.3 a)-b) in the form of a heatmap, where the highest values correspond to input features most important for the predictive output of the Washing Machine class. We observe that the distillation process has been unsuccessful in transferring the magnitudes of most relevant importance values to the Student, possibly causing the occurrence of a false positive prediction. We hypothesize that a reduction of such inconsistencies might be a crucial step in the optimization of the distillation process, leading to a more stable predictive performance.

To prevent inconsistencies in the transfer of explainable knowledge, we introduce the loss term L_{xai}^{ϕ}, which focuses on reducing dissimilarities between the Teacher's and Student's explanation maps, which represent areas of the input sequence that are attributed by XAI method with having high levels of importance for model prediction. As explanation heatmaps are represented in vector form, the inconsistency between two explanations is quantified using a loss function based on cosine similarity, defined as:

$$L_{xai}^{\phi}(a,b) = -\frac{ab}{\|a\|\|b\|} = -\frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n}(a_i)^2}\sqrt{\sum_{i=1}^{n}(b_i)^2}}, \tag{5.6}$$

where a and b represent two generated explanations, while ϕ represents the output type to be compared (weak or strong). It is expected that two similar

vectors will have a similar angle between them, leading to the conclusion that the similarity of two vectors increases as the value of their cosine angle increases. To this end, in order to promote the minimization of the loss function, we invert the sign of the generated cosine similarity measure.

To alleviate inconsistencies w.r.t transfer of explainable knowledge in KD, we introduce a modification to the KD loss function by including the cosine similarity-based loss between the explanations produced by the Teacher and the Student networks. Thus, the explainability-guided knowledge distillation loss function is defined as:

$$L_{xai-guided} = L_{dist} + \gamma \cdot L_{xai}^{\phi}(h_t, h_s), \tag{5.7}$$

where h_t and h_s represent explanations generated by Teacher and Student networks, respectively, while γ is a parameter that adjusts the influence of the cosine similarity loss component L_{xai}^{ϕ}.

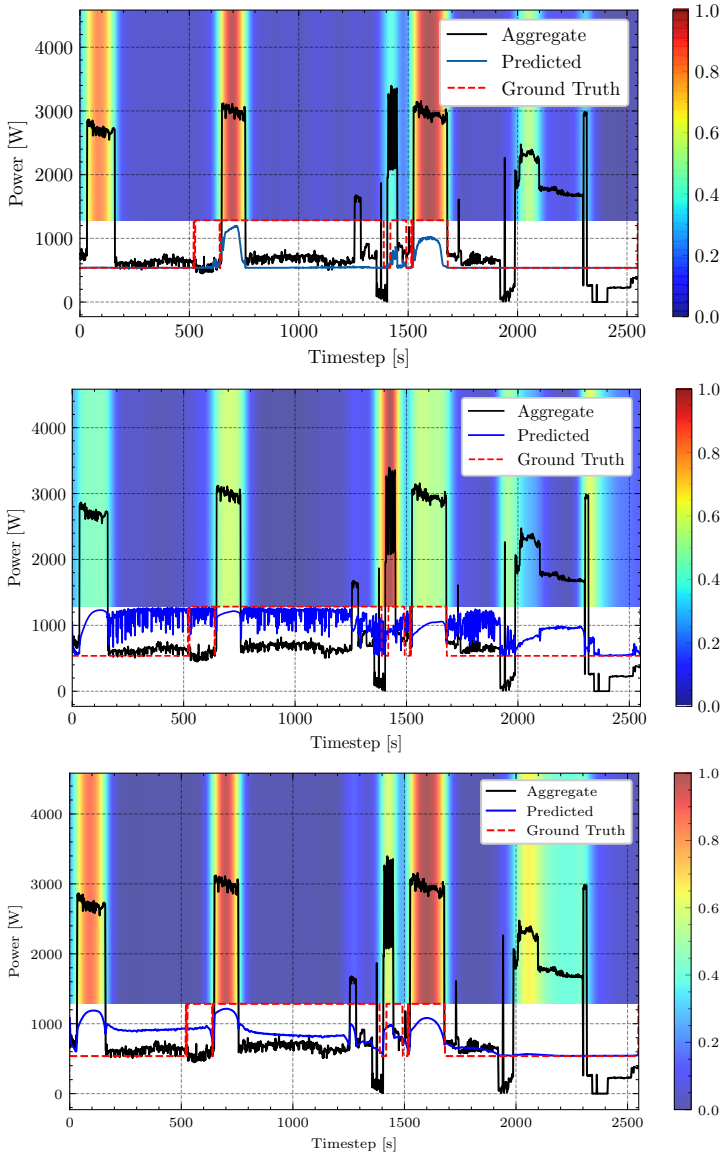


Fig. 5.3. Explanations for prediction of Washing Machine on a sample from the test set in the REFIT-to-REFIT domain adaptation scenario. a) Teacher explanation b) baseline Student explanation, displaying the inconsistent transfer of explanation knowledge c) Corrected Student explanation and prediction after explainability guided learning. Strong predictions are displayed before quantization.

5.1.3 Interpretability and Reliability-driven Learning

Despite being larger and more complex, the Teacher model cannot be assumed to be inherently trustworthy, particularly when processing signals from novel target domains. This fundamental limitation can lead to the propagation of errors through the distillation process, affecting both the Student model's predictions and the quality of generated explainability maps. These issues manifest in two primary ways: through incorrect predictions that get transferred to the Student, and through inconsistent or misleading explainability maps that fail to properly highlight relevant features in the input signal (see Fig. 5.4).

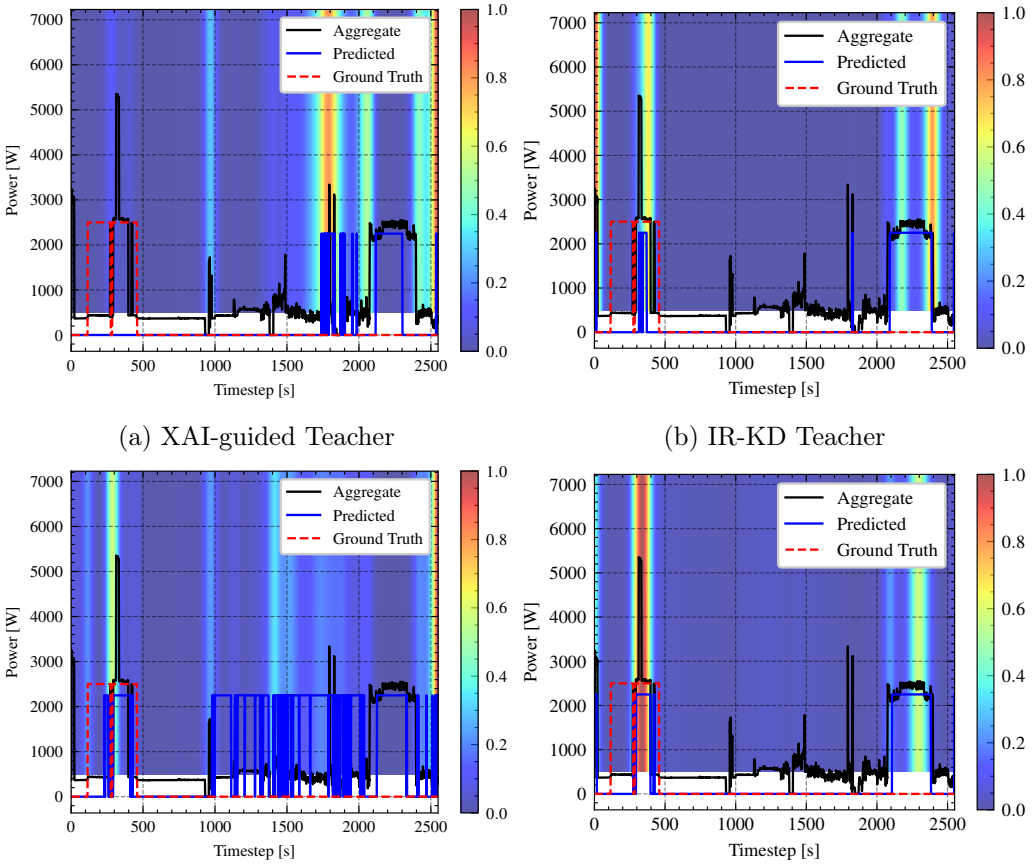
To address these Teacher-induced distillation artifacts, we propose a novel dual-component methodology that enhances the quality of knowledge transfer:

- *Zero-strong Label Loss.* We introduce a Teacher enhancement framework that optimizes the model's learning process during target domain fine-tuning. This component specifically addresses the critical need to minimize the propagation of corrupted knowledge to the Student model by maximizing information extraction from weak labels, thereby improving classification performance through systematic reduction of false positive predictions. The same concept is then adopted during the distillation, to additionally filter uncertainties transferred to the Student.

- *Perception Aligned Gradient Learning.* We implement an explainability-guided learning optimization that systematically improves the quality of Teacher-generated gradients. This enhancement ensures that the Student model mimics more accurately behavioral patterns during the distillation process. The optimization procedure incorporates techniques aimed at generating gradients that are in line with human perception to generate more precise and reliable explainability outputs.

Zero-Strong Label Loss

We propose the so-called *zero-strong label loss* to exploit as much information as possible from weak labels collected in the target domain. Weak labels, by definition, contain window-level information about the appliance usage. If a window is assigned a positive weak label ($w_n = 1$), determining the exact activation localization becomes an ill-posed problem, as it is not possible to derive where the appliance is active within that window and for how many time instants. On the contrary, if the appliance is never active in the window ($w_n = 0$), no temporal mapping is needed, and the appliance can be assumed inactive at every time instant within that window. In this case, all the strong labels can be set equal to the corresponding weak label.



(a) XAI-guided Teacher

(b) IR-KD Teacher

(c) XAI-guided Student

(d) IR-KD Student

Fig. 5.4. Comparison of predictions for Dishwasher appliance and their corresponding XAI attribution maps.

This assumption not only provides useful strong labels for windows without activations but it also reduces the quantity of false positives produced by the Teacher without the necessity for collecting additional data from the target environment. For this reason, the loss function used for Teacher fine-tuning is as follow:

$$L_{ft} = L_w + L_{z-s}, \tag{5.8}$$

where the L_{z-s} loss function is defined as the BCE computed only for the windows where $w_n = 0$. This loss considers the strong output of the Teacher,

strengthening the training process by leveraging strong information derived from weak labels. As a consequence, it ensures consistency between strong and weak predictions, at least for inactive periods, effectively preventing the presence of improper false positives.

During the distillation, the same principle is applied but we restructure the loss in (5.5) by applying the weak labels to suppress incorrect Teacher predictions generating false positives. Thus, the loss becomes:

$$L_{IR-KD} = \beta L_{soft}\left(\sigma\left(\frac{\mathbf{Z}_j^{st}}{T}\right), \sigma\left(\frac{\mathbf{Z}_j^{te}}{T}\right) \cdot \mathbf{w}_j\right) + (1-\beta)\theta(e) L_w(\hat{\mathbf{w}}_j^{st}, \mathbf{w}_j). \quad (5.9)$$

In this way, we directly correct the information transferred by multiplying by $w_j = 0$ or $w_j = 1$ the soft labels $\sigma\left(\frac{\mathbf{Z}_j^{te}}{T}\right)$ generated by the Teacher. This distillation contribution aims to additionally filter other mistakes of the Teacher predictions. The soft labels become reliable at least for the windows with $w_j = 0$. For the other windows, they are unchanged due to multiplication by $w_j = 1$.

Perception Aligned Gradient Learning

Since during distillation the Student is forced to mimic the Teacher to generate the same explainability maps (related to correct outcomes), we want to ensure the Teacher produces high-quality XAI maps. Extending the concept of explainability-guided KD, as shown in Fig. 5.1, we incorporate perception-aligned gradients to create explainable and robust distillation models. This approach aims to improve both the explainability and predictive performance of a model by ensuring that the model's decision-making process aligns more closely with human perception. The key idea behind perception-aligned gradients is that models with gradients that better align with human perception tend to be more robust and interpretable [40,105]. This alignment is achieved through gradient regularization during training. Specifically, we introduce a gradient norm penalty to the knowledge distillation loss function to enforce

this alignment.

We first train a large Teacher network using regularized gradient training to promote the development of perception-aligned gradients. Thus, referring to Equation (5.8), the fine-tuning loss becomes:

$$L_{ft} = L_w + L_{z-s} + L_{reg}, \tag{5.10}$$

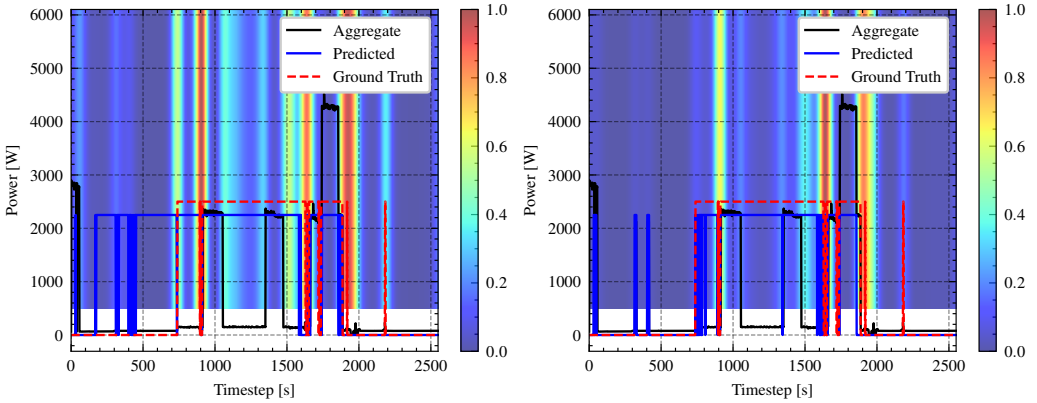
where the term L_{reg} ensures meaningful gradient alignment via gradient normalization, and is defined as:

$$L_{reg} = \lambda \cdot \frac{1}{N} \sum_{i=1}^{N} ||\nabla_x L_\omega(x_i)||_2^2. \tag{5.11}$$

This approach encourages the models to be sensitive to perceptually relevant changes in the aggregate power signal, potentially leading to more robust disaggregation decisions. For NILM, these explainable features should highlight temporal regions corresponding to actual appliance state transitions and characteristic power consumption patterns. We validate this assumption of transfer of perceptually meaningful features by examining the gradients with respect to the input power measurements, as seen in Fig. 5.5. We observe that incorrect feature attribution of teacher has been transferred to the student, coinciding with poor predictive performance. On the other hand, IR-KD approach led to improved teacher gradients, which correspond with high Dishwasher activation cycle, which was correctly transferred to the Student.

5.2 Experimental setting

To assess the validity of our method, we used the datasets, training procedure, and evaluation metrics described in the following paragraphs. Moreover, we provide a description of the benchmark methods.



(a) XAI-guided Teacher (b) IR-KD Teacher

Fig. 5.5. Comparison of Teacher XAI and predictive performance improvement using the proposed loss function.

5.2.1 Datasets

To evaluate the generalization effectiveness of the proposed distillation methodology, we used the widely adopted datasets UK-DALE [65] and REFIT [91], collected from different regions in the UK. For testing, 70% of the aggregate measurements from houses 2, 4, 8, 9, and 15 in the REFIT dataset (spanning 2013 to 2015) were used, while the remaining data were used for fine-tuning and distillation. Training data, used as the source for knowledge transfer, was derived from UK-DALE houses 1, 3, 4, and 5 over a two-year period. To ensure consistency between the datasets, REFIT data was up-sampled to match the 6-second granularity of UK-DALE. To address differences in the number of active samples among the monitored appliances, the datasets were balanced as in [122,123]. To evaluate the generalization effectiveness of the proposed distillation methodology, we used the widely adopted datasets UK-DALE [65] and REFIT [91], collected from different regions in the UK. For testing, 70% of the aggregate measurements from houses 2, 4, 8, 9, and 15 in the REFIT dataset (spanning 2013 to 2015) were used, while the remaining data were used for fine-tuning and distillation. Training data, used

as the source for knowledge transfer, was derived from UK-DALE houses 1, 3, 4, and 5 over a two-year period. To ensure consistency between the datasets, REFIT data was up-sampled to match the 6-second granularity of UK-DALE. To address differences in the number of active samples among the monitored appliances, the datasets were balanced as in [122, 123].

To validate our proposed approach, we use real-world UK-DALE [64] and REFIT [91] datasets. UK-DALE contains aggregate and appliance-level power measurements from 5 buildings acquired at a granularity of 1 s and 6 s, respectively, while REFIT contains power measurements collected from 20 houses at 8 second intervals. To account for different sampling rates in the two datasets, we resample the UK-DALE aggregate and REFIT measurements to 6 s, using back-filling to upsample REFIT data. To account for class imbalance, the datasets have been balanced as in [122]. Houses 2, 4, 8, 9, and 15 in REFIT have been used for testing. We extract a portion of this data for fine-tuning and distillation (30% of the total number of windows). To evaluate the success of our approach in performing domain adaptation, two different scenarios are used to pre-train the Teacher network, where training data are taken from UK-DALE houses 1, 3, 4, and 5 (UK-DALE-to-REFIT scenario). The UK-DALE-to-REFIT scenario is used to evaluate the performance of the proposed method when pre-training and target environment domains are different. The selected appliances (kettle, microwave, washing machine, dishwasher and washer dryer) are the most common devices monitored based on the NILM literature and allows for easier comparison with prior studies. A similar rationale guided the selection of houses from the UK-DALE and REFIT datasets. We do not distinguish between the multiple states assumed by some devices during an activation because the aim is to identify the entire period of activation. The validation set is extracted from the pre-training set, as well as the mean and standard deviation values used to normalize the input signals.

Appliance	γ	μ
Washing Machine	0.50	*weak*
Dishwasher	0.85	*strong*
Washer-Dryer	0.60	*weak*
Kettle	0.30	*weak*
Microwave	0.70	*weak*

Table 5.1
Training hyperparameters used for training of XAI-guided Student models.

5.2.2 Benchmarks

We compare our proposed perception guided method with [72,78], which were adapted for this task, and perform ablation study using only explainability guided learning [18].

EdgeNILM [72] uses pruning and tensor decomposition. For our experiments, we used the source code made available by the authors to ensure reproducibility. To adapt the EdgeNILM for multi-label appliance classification, we modified the last layer of the Sequence-to-Point CNN by replacing it with a sigmoid function to output the state probability and used the BCE loss function during training. As in [72], we trained a separate network for each appliance, and applied the 60% iterative pruning method for complexity reduction, as this approach demonstrated the lowest average disaggregation error [72]. A window size of 99 samples was adopted for all the appliances in EdgeNILM, based on the results presented in [72].

The LightweightCNN proposed in [78] adopts a model design approach and consists of only two convolutional layers and one dense layer. To ensure a fair comparison, the lightweight network was implemented and trained within the same framework as EdgeNILM, using a window size of 199 samples as specified in [78]. Similar to EdgeNILM, a separate network was trained for each appliance in this approach. Finally, we compare the perception aligned approach with the results of XAI-guided NILM [18], to highlight the improvements and benefits.

Model	Size (MB)	FLOPS (M)
EdgeNILM Unpruned [72]	69.1	32.11
EdgeNILM Pruned 60% [72]	11.15	5.23
LightweightCNN [78]	10.65	5.01
XAI-guided NILM Student [18]	0.44	7.5
IR-KD	0.44	7.5

Table 5.2
Model Size (MB) and FLOPS (M) for the benchmark methods and the proposed approach. The model size and the number of FLOPs are calculated on all the networks used to classify $N = 5$ appliances.

In Table 5.2 a computational complexity comparison between our approach and the benchmarks is reported. The analysis is conducted based on memory occupancy (MB) and FLOPs as in [72] to ensure an appropriate comparison independently from the selected hardware. The size of our networks (IR-KD), as for XAI-guided NILM, is widely reduced compared to LighweightCNN and EdgeNILM Pruned 60%. In terms of FLOPs, our method requires a larger number of operations that impacts on the inference time. For consistency, the same post-processing steps applied for our method were applied to the raw predictions of the benchmark methods to obtain the state of appliance. Seeds have been fixed to reproduce the experiments and the best models have been chosen based on the performance on the validation set. Lastly, we optimize the hyperparameters using a grid search approach, and the best values found for each network related to each appliance are reported in Table 5.1 for the Student models.

5.2.3 Classification and energy-based metrics

Our experiments aim to demonstrate that the proposed method effectively reduces false positive predictions, first in the Teacher model and then in the Student model during the distillation process. Four metrics commonly used in the NILM literature have been considered to evaluate our method. Defining True Positives (TP) as the number of correctly classified active samples, False Positives (FP) as the number of inactive samples incorrectly

classified as active, and False Negatives (FN) as the number of active samples incorrectly classified as inactive, we used the Recall (RE) and Precision (PR) defined as RE $= TP/(TP + FN)$, PR $= TP/(TP + FP)$ to evaluate the percentage of active samples that are not detected and percentage of inactive samples predicted as active, respectively. The F_1-score is the harmonic mean between Precision and Recall and is formulated as $F_1 = 2 \cdot P \cdot R/(P + R)$.

In order to evaluate energy efficiency of a household, it is useful to report consumption and duration. For this reason, it is important to evaluate our method based on the energy correctly assigned. For each activation, the related energy consumption can be calculated, and the accuracy of this assignment can be assessed using the Total Energy Correctly Assigned (TECA) metric [70], defined as follows:

$$\text{TECA} = 1 - \frac{\sum_n \sum_t |\hat{p}_n(t) - \bar{p}_n(t)|}{2 \sum_t \bar{y}(t)}, \tag{5.12}$$

with $\bar{y}(t) = \sum_n \bar{p}_n(t)$. The terms $\hat{p}_n(t)$ and $\bar{p}_n(t)$ denote, the product of the average power consumed by appliance n at the time instant t, and respectively, estimated states $\hat{s}_n(t)$ and the ground-truth states $s_n(t)$. The average power consumed by each appliance is determined based on the average power consumed by the appliances in the training set.

5.3 Results and Discussion

In this section, we first discuss the state identification performance in Section 5.3.1, followed by an analysis of interpretability performance in Section 5.3.2, using metrics introduced in Chapter 3. Lastly, in Section 5.3.3, we jointly discuss the performance on both interpretability and state identification.

5.3.1 State identification performance

We begin by describing the performance of the Teacher model using our proposed fine-tuning strategy, which leverages strong information for windows with weak labels $w_j = 0$.

Table 5.3 shows that both PR and RE for the IR-KD Teacher are improved compared to the XAI-guided Teacher. As a consequence, the F_1-score also improves, with an average increase of 4.7%. In terms of energy correctly assigned, TECA shows an even greater improvement, with an increase of 10.16%. For almost all appliances the occurrence of false positives is reduced, based on the PR scores. For the washing machine, false negatives are reduced based on the RE scores.

We observe that the explainability guided learning led to an increase in performance compared to the baseline model for all appliances. In the context of low-granularity NILM time series (e.g., sub-30 second data), appliance activations are often sparse and short-lived relative to the entire observation window. This means that the corresponding explanation heatmaps are also likely to be sparse, with high importance scores concentrated in relatively small regions corresponding to these activations, and near-zero or very low scores elsewhere. This could be one of the reasons for good performance of cosine similarity measure, as if an appliance is off (and thus has near-zero true importance), and both models correctly attribute low importance to

Appliances	**IR-KD Teacher**			**XAI-guided Teacher [18]**			**IR-KD Student**			**XAI-guided Student [18]**		
Metrics	PR ↑	RE ↑	F_1 ↑	PR ↑	RE↑	F_1 ↑	PR ↑	RE↑	F_1 ↑	PR↑	RE↑	F_1 ↑
Kettle	0.78	0.45	0.57	0.77	0.42	0.55	0.73	0.60	0.66	0.31	0.97	0.47
Microwave	0.49	0.97	0.66	0.43	0.98	0.60	0.75	0.93	0.83	0.69	0.96	0.80
Washing machine	0.53	0.77	0.63	0.56	0.69	0.62	0.56	0.89	0.69	0.55	0.81	0.65
Dish washer	0.58	0.75	0.65	0.49	0.84	0.62	0.51	0.90	0.65	0.52	0.83	0.64
Washer Dryer	0.82	0.80	0.81	0.79	0.77	0.78	0.78	0.76	0.77	0.75	0.81	0.78
AVG.	**0.64**	**0.75**	**0.66**	0.61	0.74	0.63	**<u>0.67</u>**	0.82	**<u>0.72</u>**	0.56	<u>0.88</u>	0.67
TECA	**69.25**			62.86			**<u>72.65</u>**			61.39		

Table 5.3
Comparison between the XAI-guided Teacher and Student [18] with the IR-KD Teacher and Student. Bold represents the best scores when Teachers and Students are compared, respectively. Underlined are the best scores among all the networks.

	Appliance					
Model	**Kettle**	**Microwave**	**Washing Machine**	**Dishwasher**	**Washer Dryer**	**Average**
EdgeNILM Unpruned [72]	0.64	0.01	0.43	0.19	0.23	0.25
EdgeNILM Pruned 60% [72]	0.68	0.03	-	0.07	-	0.13
LightweightCNN [78]	**0.75**	0.33	0.51	0.53	0.42	0.43
XAI-guided NILM Teacher [18]	0.55	0.60	0.62	0.62	**0.78**	0.63
XAI-guided NILM Student [18]	0.47	0.80	0.65	0.64	**0.78**	0.67
IR-KD Student	0.66	**0.83**	**0.69**	**0.65**	0.77	**0.72**

Table 5.4
Comparison in terms of F_1-score between the proposed approach and the benchmarks. For the Edge-NILM Pruned 60% approach, Washing Machine and Washer Dryer are not reported because the model was not able to learn with a high pruning percentage.

that region, these areas contribute little to the dot product but still factor into the magnitude calculations. The measure becomes more sensitive to whether both models agree on the few important, active regions. When comparing with the Teacher model, we note improvements for all appliances, except for WD, where the F-score remains unchanged, and KT, where the F-score decreased, but still remained significantly higher than the baseline model. A possible reason for the poor performance for KT is the fact that in this case, the Teacher model might not be ideal for knowledge distillation, as its low recall value suggests that it exhibits a high number of false negative predictions.

The IR-KD Student, distilled from the refined IR-KD Teacher, achieves notable improvements: a 14.28% increase compared to the XAI-guided Teacher, a 9.09% increase compared to the IR-KD Teacher, and a 7.46% increase compared to the XAI-guided Student. In particular, the results for kettle and washing machine show even greater enhancements due to inclusion of the proposed IR distillation loss contribution. Kettle signatures can be confusing in presence of spikes that last seconds or minutes and that easily occur in a common day consumption signal. The benefits on washing machine predictions, on the other hand, depend on its complex activations. This demonstrates that exploiting more information from weak labels enhances both Teacher and Student knowledge, specifically reducing the number of false positives.

Table 5.4 evidences that our approach outperforms also benchmark com-

plexity reduction strategies. Specifically, it achieves an average F_1-score that is four times higher than EdgeNILM Pruned 60% and improves by 67.4% compared to LightweightCNN. It is important to clarify that for the EdgeNILM Pruned 60% approach, Washing Machine and Washer Dryer are not reported because the model was not able to learn with a high pruning percentage. This suggests the underpinning importance of the fine-tuning phase while moving into a real-world scenario while reducing complexity that benchmark methods did not consider. The same holds for the LightweightCNN. Despite this, it exhibits the best score for kettle. For our approach, this incorrect behavior on the kettle could be justified by the lower score of the Teacher, from which our Student will learn and that can not overcome completely Teacher's mistakes. For microwave, washing machine and dishwasher the proposed approach matches the desired result of having lower complexity with improved performance.

5.3.2 Explainability performance

The experimental results, shown in Table 5.5 and Fig. 5.6, provide compelling insights into the explainability characteristics of both teacher and student models across different NILM KD approaches. Our analysis focuses on three key metrics: faithfulness (F), robustness (R), and complexity (C), each offering unique perspectives on model interpretability. Faithfulness, which measures how accurately explanations reflect model behavior, is par-

Appliances	**IR-KD Teacher**			**XAI-guided Teacher [18]**			**IR-KD Student**			**XAI-guided Student [18]**		
Metrics	F↑	R↓	C↑	F↑	R↓	C↑	F↑	R↓	C↑	F↑	R↓	C↑
Kettle	18.731	19.620	0.927	17.257	19.683	0.925	23.055	20.025	0.982	16.821	20.073	0.982
Microwave	6.662	19.916	0.918	4.863	19.953	0.920	15.344	20.146	0.932	11.227	20.154	0.98
Washing machine	12.701	19.620	0.921	10.383	19.682	0.919	18.731	19.62	0.927	17.257	19.682	0.925
Dish washer	7.810	19.724	0.920	4.778	19.720	0.907	24.33	19.854	0.908	22.87	20.18	0.931
Washer Dryer	14.027	19.833	0.940	12.093	19.880	0.937	15.129	19.294	0.911	12.953	19.671	0.872
AVG.	**11.986**	**19.743**	**0.925**	9.875	19.784	0.922	**19.318**	**19.788**	0.932	16.226	19.952	**0.938**

Table 5.5
Comparison of explainability performance between the XAI-guided NILM Teacher and Student [18] with the IR-KD Teacher and Student.

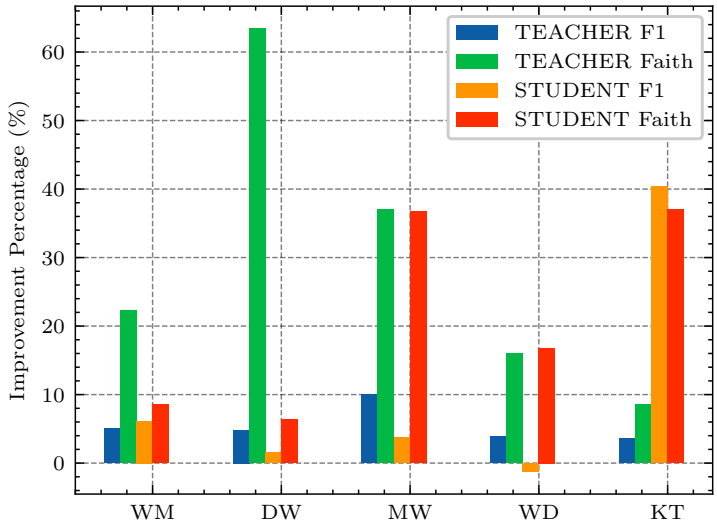


Fig. 5.6. Relative improvement over the XAI-guided NILM Teacher and Student [18] of explainability (Faithfulness - "Faith") and predictive performance metrics ($F_1 - score$)

ticularly crucial in NILM contexts where understanding feature importance directly impacts the reliability of appliance detection. Our implementation quantifies faithfulness through feature perturbation analysis, where higher scores indicate stronger correlation between identified important features and model predictions. The results demonstrate distinct patterns between IR-KD and XAI-guided approaches, with IR-KD Teacher models exhibiting consistently higher faithfulness scores (mean $F = 11.986$) while maintaining stable robustness (average $R = 19.743$) and strong complexity scores (mean $C = 0.925$). In contrast, baseline XAI-guided Teacher model shows lower faithfulness scores (mean $F = 9.875$) with comparable robustness scores (average $R = 19.784$). The KD process yields particularly interesting results in model behavior, with IR-KD students showing great faithfulness improvements across all appliances (average increase over the teacher of 89.4%). This improvement is particularly pronounced for high-power appliances, with the dishwasher showing a 211.5% increase in faithfulness scores compared to

the teacher. Cross-appliance analysis reveals important patterns in model behavior. High-power appliances (kettle, dishwasher) consistently demonstrate exceptional faithfulness improvements under IR-KD distillation while maintaining high complexity scores (> 0.90) with minimal robustness degradation ($< 2\%$ increase). Appliances with variable operating modes (washing machine, washer dryer) show more moderate faithfulness improvements (47-68%) but exhibit greater variations in complexity scores. These patterns suggest that appliance characteristics significantly influence the effectiveness of different distillation approaches. Overall, the results indicate that gradient normalization is most effective for appliances with distinct, high-power signature spikes, where the clear separation between operational and idle states allows for more precise feature attribution localization. This indicates that the proposed method helps improve the localization of the attribution maps, while keeping the robustness and complexity relatively stable.

The architectural implications of our results suggest that IR-KD is particularly effective at transferring feature importance recognition while maintaining the overall explanation stability of XAI-guided NILM. Notably, student models consistently show improved faithfulness despite their reduced capacity, indicating that model compression through knowledge distillation can enhance, rather than compromise, interpretability. This finding has significant implications for edge computing deployments, where both model efficiency and interpretability are crucial considerations. Additionally, our results reveal interesting trade-offs between different explainability metrics. The consistent improvement in student model faithfulness, coupled with maintained complexity scores, suggests that smaller, more efficient models can provide more reliable and interpretable results than their larger counterparts. However, the lack of improvement in robustness metrics indicates a potential for further work on jointly improving explanation accuracy and stability.

5.3.3 Interpretability and Reliability joint discussion

Interestingly, the performance-explainability relationship in IR-driven KD shows varying patterns across different appliance types. For complex appliances like the washing machine, modest improvements in F_1-score (0.63 to 0.69, 9.5% increase) are accompanied by larger gains in faithfulness (12.701 to 18.731, 47.5% increase). This suggests that enhanced interpretability does not always translate to proportional improvements in predictive performance, particularly for appliances with more complex operation patterns.

The recall metric (RE) (Table 5.3) shows consistent improvements across all appliances in the student model, with an average increase from 0.75 to 0.82. This improvement aligns with higher faithfulness scores, suggesting that better feature identification leads to more reliable appliance state detection. However, we observe a trade-off in precision (PR), where some appliances show slight decreases (e.g., kettle: 0.78 to 0.73), despite improved faithfulness. This indicates that while enhanced interpretability generally supports better recall, it may not always contribute to improved precision.

A notable observation is that appliances with the highest relative improvements in faithfulness (microwave and dishwasher) also show the most substantial gains in recall, suggesting that better feature attribution particularly benefits the model's ability to identify true positive cases, possibly due to more stable model gradients. The washer dryer, which shows the smallest improvement in faithfulness scores (7.9% increase), also demonstrates the least improvement in predictive metrics, further supporting the correlation between explainability and performance. The preservation of complexity scores (C) in the student model (maintaining values above 0.9) alongside improved F_1-scores suggests that the IR-KD approach successfully transfers both performance capabilities and interpretability characteristics. This is particularly important as it indicates that the compression of model size does not compromise either predictive performance or explanation quality. These findings have important implications, suggesting that optimization for

explainability through IR-KD can lead to improved predictive performance. The strong correlation between faithfulness improvements and F_1-score gains indicates that focusing on interpretability during the knowledge distillation process can enhance overall model effectiveness, rather than presenting a trade-off between explainability and performance.

5.4 Summary

Energy-efficient and low-complexity algorithms are essential for deploying DNNs on resource-constrained edge devices. While KD has emerged as a prominent technique for model compression, our work addresses critical gaps in existing approaches that primarily focus on performance metrics while overlooking interpretability and reliability challenges. Additionally, previous studies overlooked critical aspects of how incorrect Teacher knowledge impacts Student learning outcomes, negatively influencing the aforementioned challenge. Through our proposed IR-KD framework, we demonstrate that incorporating perception-aligned gradients and weak label information during knowledge transfer can significantly enhance both model interpretability and reliability. Our evaluation in the context of NILM applications reveals that the IR-KD approach not only maintains computational efficiency, but also enhances decision-making transparency across diverse deployment scenarios. Quantitative explainability metrics confirm that the proposed method leads to more faithful explanations, while keeping the explanation visualizations robust and low in complexity. On the other hand, the improved learning strategy demonstrates enhanced predictive performance. These results indicate that training explicitly for transparency and reliability can substantially improve the knowledge distillation process.

Chapter 6

Towards Equitable EV Charging Station Placement Using Graph Neural Networks

While major world economies have announced ambitious charging infrastructure plans, current deployment patterns risk reinforcing socio-economic inequalities. Previous studies have highlighted numerous barriers that hinder the widespread adoption of EVs [73]. The high initial costs associated with EVs pose a significant challenge. Without adequate financial subsidies, steep upfront expenses can deter potential buyers. As discussed above, this challenge has been the main focus of national efforts worldwide. The time required for a full battery recharge - which far surpasses the refueling time for fossil fuel vehicles - further complicates the attractiveness of EVs [19]. Furthermore, driving range anxiety, i.e., the fear that an EV will not have sufficient battery capacity to complete a desired journey, remains a significant concern. Recharge time and range anxiety have mostly been addressed from the perspective of investments in the development of battery technology, lightweight body and material design, and improved powertrain, leading to better utilization, higher capacity, and improved driving range. How-

ever, effective installation of charging infrastructure directly or indirectly addresses the above concerns of potential EV adopters. Early adopters of EV technology tend to be homeowners residing in high-income areas, utilizing home-based charging [23]. On the other hand, to achieve widespread adoption of EVs, a substantial number of newer EV adopters should belong to moderate-income groups residing in multifamily residential communities that are less likely to have access to home charging [30]. Thus, public EVCS infrastructure holds a substantial importance for the adoption of EVs, especially in deprived communities [54]. Previous research suggests that areas with greater deprivation and lower economic position are associated with higher levels of pollutants [41]. In the context of EV adoption, this means that those who could benefit the most from low-carbon technology are the least able to afford it. Therefore, as EV adoption continues to rise, ensuring that EV infrastructure is placed in a fair and just way that benefits all members of society is of utmost importance for achieving an equitable transition towards net-zero, in line with UN Sustainable Development Goals (SDGs) [13], specifically, SGD 9.1 regarding equitable access to infrastructure and SDG 11.2 regarding sustainable and accessible transport systems for all.

To address these disparities, researchers have proposed various approaches to incorporate equity considerations into EVCS planning. Quantification of social equity w.r.t. EVCS access is proposed in [77] with a social equity access function that incorporates the Quality of Life (QoL) index and spatial factors, including distance to EVCSs, saturation index of current EVCSs, distance to EV demand points, and average traffic flow, aiming to identify regions with low QoL index scores and poor accessibility to EVCSs. A combination of socioeconomic indicators, spatial accessibility measures, and policy interventions is suggested in [54] to address equity concerns in EVCS placement. A multi-objective optimization model that considers site development costs, social equity access, and EV demand fulfillment is proposed in [79], whereby

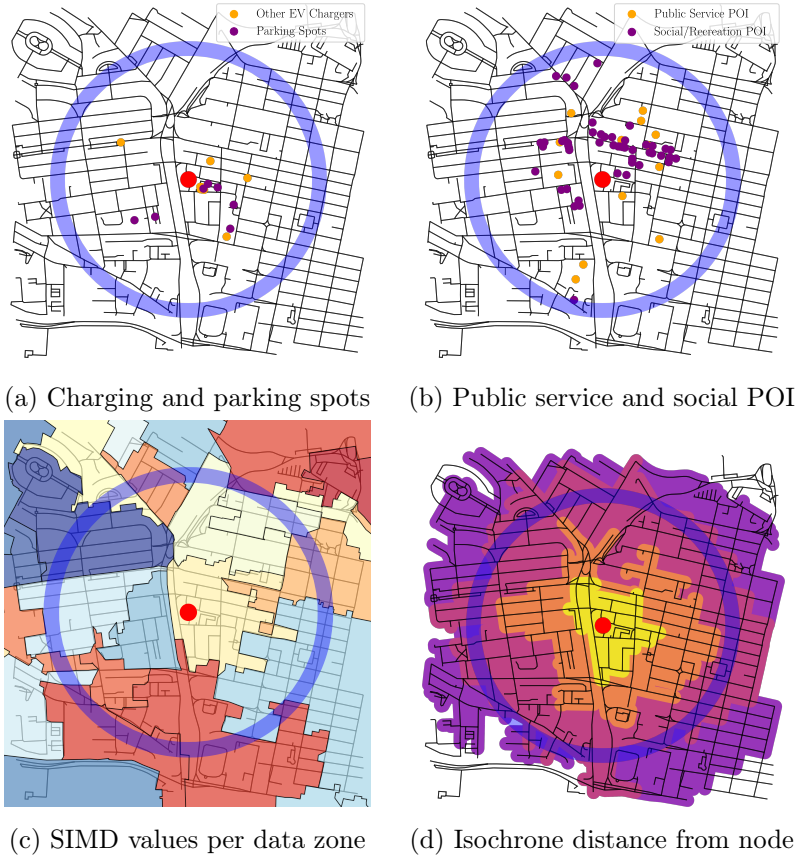
social equity access is quantified using a combination of socioeconomic factors and spatial accessibility measures. A systematic and analytical approach is presented in [100] to investigate the equitable distribution of public EV charging infrastructure, considering both horizontal and vertical equity, using census data to measure accessibility and evaluate inequity using spatial autocorrelation analysis and the Gini index. A more design-oriented approach investigates the influence of charging station accessibility and agglomeration effect on utilization rates in [39], considering the operational aspects of charging stations and the market competition among multiple operators. Authors in [45] present a methodology for optimal placement of charging station energy hubs (CS-EHs) using data aggregation from various open-source datasets, including information on renewable energy sources, traffic density patterns, and EV charging behaviors in Norway. [30] models the choice of charging location for PEV drivers as a function of various demand drivers, including the cost of charging at home and work, characteristics of charging infrastructure, and demographic factors. The viewpoints of EVCS owners, grid operators, and EV users are incorporated in the multi-objective optimization model of [77], balancing the trade-offs between minimizing site development costs, maximizing social equity access, and maximizing EV demand fulfillment. The spatial clustering of public EVCS infrastructure and their associated characteristics in the Chicago metropolitan area is proposed in [27], considering the perspectives of both EV users and charging infrastructure providers, highlighting the wealth disparity and the need to balance efficiency with equity in EVCS placement.

In contrast to prior work, this chapter presents a methodology that proposes a novel approach leveraging GNNs to enhance EVCS placement. By embedding diverse multi-modal data sources, including existing EVCS utilization, infrastructure information, traffic flow patterns, points of interest, deprivation indices, and parking infrastructure, our model provides a holistic and practical solution to EVCS distribution. The focus on underserved areas

through the proposed placement utilization metric and the consideration of specific land uses directly addresses equity concerns often overlooked in previous studies. Our approach offers targeted recommendations for different land use types (residential, working/industrial, commercial), acknowledging the varying charging needs across urban contexts. We conduct extensive infrastructure placement evaluations using multi-source data from the selected urban areas of major cities of Scotland. Specifically, the methodology brings the following contributions by modelling of GNNs aimed to achieve equitable, geodemographic aware EVCS planning:

- Integration of multiple socio-economic and environmental factors to support equitable EVCS infrastructure location and capacity placement.
- Modeling of urban EV charging demand via graphs and development of a novel GNN architecture to learn complex urban dynamics and correlation between charging demand influencing factors to facilitate identification of optimal areas for EVCS placement.
- Detailed analysis of EVCS utilization and distinct usage patterns in residential, working/industrial, and commercial zones, as well as between deprived and non-deprived areas in major cities of Scotland.
- Targeted placement decisions informed by urban land use requirements, and historical EVCS utilization and capacity, ensuring that placements are aligned with local policy goals.
- The evaluation of impact of targeted EV infrastructure deployment in promoting higher EV uptake in deprived areas with significant potential.

6.1 Data Processing and Labeling Methodology



(a) Charging and parking spots (b) Public service and social POI

(c) SIMD values per data zone (d) Isochrone distance from node

Fig. 6.1. Charging demand node area description and data used in this study

6.1.1 Case Study

The primary objective of the electrification of transport is to mitigate Greenhouse gas (GHG) emissions on a global scale. However, EVs are often powered by electricity derived from non-renewable energy sources, specifically power stations run on fossil fuels. This means that while EVs themselves do

not directly emit GHG, the overall reduction in GHGs depends largely on how the electricity they use is generated. Therefore, for EVs to effectively benefit the climate and to ensure the sustainability of the entire energy system, the use of renewable energy sources for EV charging is essential. To this end, Scotland represents an interesting case study, as it is the UK leader in renewable energy production, continuously generating more electricity than it needs, with net electricity exports amounting to 15.9 TWh in 2023. In Scotland, the total electricity generation from renewable sources in 2023 was 33.3 TWh, of which 77.5% came from wind energy, 13.8% from hydro, and the rest from biomass and other sources [124]. As a result, in the context of EV ownership, Scotland was identified as a UK region with the highest lifecycle assessment evaluations aimed at quantifying the reduction of carbon footprint per vehicle [131]. Given its leading position in renewable energy generation, Scotland can more effectively leverage public EV charging to achieve higher EV uptake, charging utilization, and a substantial decrease in GHG emissions, setting a standard for other UK regions to follow. Glasgow and Edinburgh were selected as primary case cities to extend the relevance and application of this study. These cities represent distinct urban morphologies with significant variations in spatial layout, population density, and transportation infrastructure, which are primary factors affecting EV charging demand patterns. The socio-demographic diversity within these cities spans multiple critical dimensions, including substantial variations in income levels (from affluent neighborhoods to areas of high deprivation), car ownership rates, racial makeup of population, housing types (detached homes to high-density apartments), and transportation access.

Glasgow and Edinburgh are not only the two largest cities in Scotland but also present largest EV adoption and existing infrastructure, vital for a comprehensive analysis: Glasgow contains 391 or 7.8% of Scotland's total number of active charging points, whereas Edinburgh has a total of 298 charging points. Furthermore, the selected cities exhibit high variability in

socio-economic conditions, reflected in their deprivation indices. Consideration of such factors is essential to understand potential barriers to EV adoption and to ensure that the benefits of the electrification of transport are equitably distributed.

6.1.2 Data Collection and Processing

To fully capture metrics that drive charging demand, we base our analysis on the various charging demand influencing factors [3]. To effectively capture these factors within a specific area, we introduce the concept of charging demand nodes. These nodes are defined as 500-meter radius circles centered around charging stations or parking spots, a distance that aligns with the observed preference for shorter walking distances to charge vehicles [128]. This approach allows us to encapsulate and analyze the relevant influencing factors within a practical and accessible range. To illustrate this concept, we provide visual representations of these charging demand nodes in Fig. 6.1, where each subfigure showcases a demand node centered on a charging station. The 500-meter radius of these nodes encompasses a variety of pertinent influencing factors that contribute to charging demand. Our study takes into account a diverse range of these factors, which we will explore in detail, to provide a holistic understanding of the dynamics driving electric vehicle charging demand in urban environments.

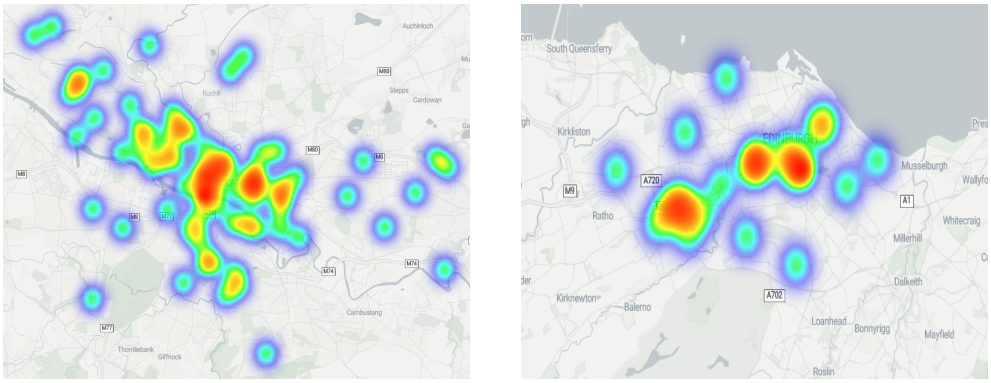
Existing EVCS Infrastructure and Charging Utilization Data

To collect EVCS infrastructure data, National Chargepoint Registry [38] was used. The registry contains detailed records of over 4,000 public EVCSs in Scotland, including station name, location, operational status, tariffs, availability, charging power output, charging plug type, etc. To collect EVCS session data, ChargePlace Scotland registry [32] was utilized. Developed as a national network of EVCS on behalf of the Scottish Government, Charge-

Place Scotland registry includes detailed historical records of public EV charging across Scotland. For the purpose of this study, key factors such as the geographic coordinates of each station, charging power output, types of connectors available, and the frequency and duration of charging sessions are used. Fig. 6.2 illustrates the public EVCS heatmaps for Glasgow and Edinburgh, which indicates locations where charging sessions were performed between October 2022 and January 2024. Only active, public EV chargers were considered.

Parking Infrastructure Data

The process of identifying potential locations for EVCS relies on OpenStreetMap (OSM) data, a comprehensive and crowd-sourced mapping resource [95]. The data collection encompasses a wide range of parking locations, including on-street parking, as well as parking facilities at event venues, hospitals, universities, and other key locations, all identified through the 'parking' amenity type in OSM. While our study does not explicitly model investment costs, the methodology inherently considers cost efficiency through its strategic focus on existing parking infrastructure. By identifying potential EVCS locations exclusively from existing parking spots, we significantly reduce installation costs by eliminating the need for land acquisition, demolition, or major construction work. In addition, this approach significantly reduces initial capital expenditure by eliminating the need for land acquisition and new construction. Second, existing parking lots are already integrated into the urban fabric, ensuring immediate accessibility and connectivity – factors crucial for user convenience and adoption rates. By repurposing existing parking spaces, this strategy aligns with sustainable urban planning principles, minimizing the environmental impact typically associated with new construction projects. Lastly, this approach also facilitates quicker deployment of charging infrastructure, accelerating the transition to EVs and supporting broader environmental and public health goals.



(a) Glasgow (b) Edinburgh

Fig. 6.2. Heatmaps of historical EVCS utilization

Demographic and Deprivation Data

The study incorporates elements related to the Human Development Index, as detailed in the Scottish Index of Multiple Deprivation (SIMD) dataset [108]. Notably, SIMD data is organized into “data zones", which are specific areas designated for small-scale statistics related to deprivation in Scotland (as seen in Fig. 6.1.c). These statistics include relevant domains such as income, employment, education, housing, health, crime, and geographical access. Integrating SIMD data into the analysis offers valuable insights into the socioeconomic context of potential EVCS sites, ensuring that infrastructure is strategically placed to be effective and beneficial, particularly in areas that might otherwise lack sufficient EV infrastructure. For each existing or potential charging spot, we collect data based on the data zone it is located within, enabling us to precisely align socio-economic and spatial factors in our placement strategy. In this study, we classify areas as “deprived" if they fall within the lower 50th percentile of the SIMD index, while those above this threshold are categorized as “non-deprived". For the purpose of this study, the latest SIMD report was utilized [108]. To address potential biases, instead of relying on the aggregate SIMD rank, a charging demand node incorporates individual socioeconomic indicators that comprise SIMD

as node attributes, specifically: Population density, Working population, Income rate, Employment count, Reachability metrics, and Crime rate.

Point of Interest Data Categorization

Point of Interest (POI) represents a specific location or landmark within a city that is relevant to travelers, residents, or urban planners. These are typically places that people might want to visit, navigate to, or use as reference points when moving around a city. Previous research consistently identifies POIs as crucial indicators of EV charging demand [2,29,45]. Our study leverages the comprehensive OpenStreetMap (OSM) [95] dataset to collect and analyze POI data, providing a rich source of information on potential charging demand hotspots. In our approach, we categorize POIs into two primary types: social and recreational. Social POIs encompass locations such as educational institutions (schools and universities), healthcare facilities (hospitals and GP practices), financial centers (banks), and other essential services (pharmacies). These represent areas where people spend significant time during their daily routines. Recreational POIs, on the other hand, include leisure and entertainment venues like restaurants, cafes, theatres, cinemas, and shopping centers, which attract visitors for shorter duration but often in higher volumes. The inclusion of both social and recreational POIs ensures that our charging infrastructure planning accounts for a wide spectrum of public activities, from daily necessities to leisure pursuits. The distribution of POIs aids in strategically placing charging stations in areas where drivers are likely to spend significant time, enhancing charging convenience.

Traffic Flow Data Creation

Traffic flow data is sourced from the UK Government's Road Traffic Statistics [37], and includes vehicle movement patterns, traffic volumes, and peak usage times across the national road network. By incorporating this data into our model, we can accurately identify high-traffic areas where the demand for

EV charging is likely to be significant. Traffic count data is typically collected at specific points along roads, using methods such as fixed sensors or periodic manual counts. However, EV charging demand is not limited to these specific data collection points but extends to broader areas where vehicles might park and charge. This mismatch between point-based data collection and area-based charging demand necessitates a method to approximate traffic data to cover potential charging locations. To overcome this, we developed a robust, adaptive approach that captures spatial variability and accounts for data sparsity. Starting with a 1km radius around each potential charging location, we collect all available traffic flow data points within this area. We then calculate an initial average traffic flow for each potential charging location. Subsequently, the algorithm enters a recursive phase, expanding the dataset in each iteration by incorporating both official traffic data and previously calculated approximations for charging points. This expansion allows for the estimation of traffic flow at locations initially lacking data by considering newly calculated values from nearby areas. The process iterates, progressively refining and propagating traffic flow information across the network of potential charging locations. Convergence is reached when successive iterations yield no significant changes in traffic flow estimates, indicating a stable and comprehensive set of values.

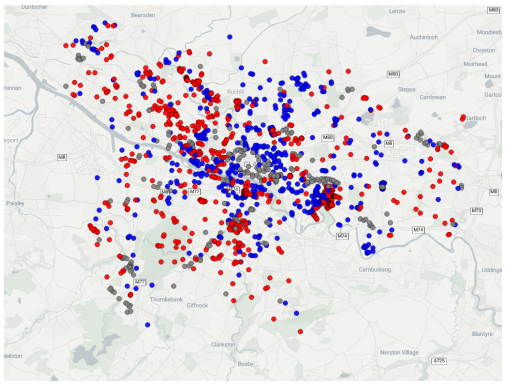
Land Use Label Creation via Clustering

Land use is a critical factor in determining optimal locations for EV charging stations. It influences accessibility, demand patterns, and dwell times, which are essential for meeting charging needs efficiently. Different land uses offer varying levels of existing electrical infrastructure, affecting installation costs and feasibility. Zoning regulations and future development plans tied to land use also impact where stations can be placed. However, detailed land use information for Scotland is not easily accessible. To improve the placement of new EVCS infrastructure, we conducted a comprehensive land use labeling

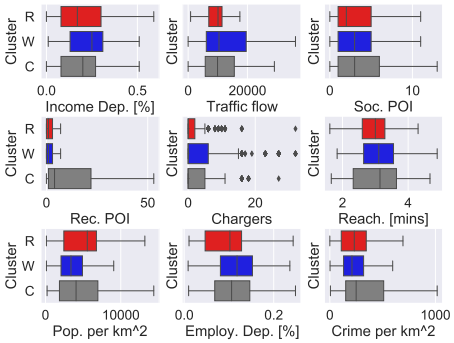
process for both existing and potential locations. This approach is crucial for maximizing accessibility and efficiency, as strategic alignment with existing and planned land use ensures that charging infrastructure effectively supports high-demand areas such as residential neighborhoods, commercial zones, industrial hubs, and transportation hubs, thereby enhancing convenience for EV users. Our methodology employs k-means clustering on a range of geodemographic influencing factors throughout the whole of Scotland. These factors included income deprivation rates, traffic flow, counts of social and recreational POIs, existing charger numbers, reachability metrics, normalized population and employment deprivation figures, and crime rates. Setting $k = 4$, this initial clustering yielded four distinct land use groups that were labeled as: Residential, Rural, Working/Industrial, and Commercial areas, closely aligning with previous findings reported in [2].

This initial clustering, performed on a Scotland-wide scale, provided a general categorization of land use. However, refinement was necessary due to the distinct urban fabric of Glasgow and Edinburgh compared to the broader Scottish context used in the initial clustering. To refine the initial cluster labels, we utilized OSM Landuse data [107]. OSM Landuse dataset describes the primary use of land by humans, where land use features are identified with a landuse tag. The database contains over a thousand tag values for landuse used in the OSM Landuse dataset. In this chapter, we refine our results by using 'residential' tag to denote Residential land use, 'industrial' tag to denote Industrial/Working land use, and 'retail' tag to denote Commercial land use. In cases where initial clustering labels differed, the land use was refined based on the corresponding OSM Landuse tag. This allowed us to more accurately classify areas within these cities, particularly in distinguishing between residential and working/industrial areas, a challenge in urban settings where these uses often overlap or exist in close proximity. The impact of this refinement is substantial and clearly demonstrated by the shifts in classification for both cities.

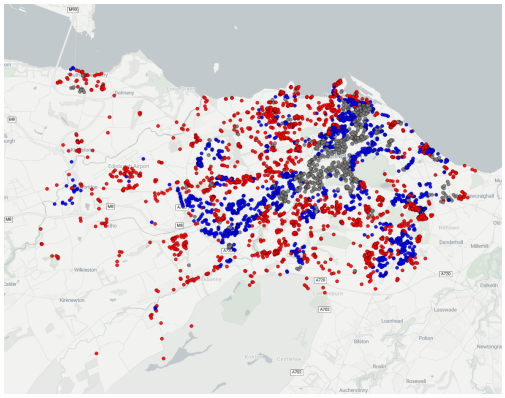
OSM land use data was collected based on the location of each potential EVCS location within OSM tiles, supplemented with gap-filled data if stations fell outside tile boundaries. No rural areas were detected within Glasgow City or City of Edinburgh councils. This refined approach enabled a more nuanced classification, rectifying many areas initially identified as working/industrial that were, in fact, predominantly residential. The result is a more accurate representation of land use patterns, crucial for informed EVCS placement decisions. To illustrate the outcomes of this labeling process, we provide a visual overview of the labeled areas for all charging and available parking infrastructure in Glasgow and Edinburgh, along with per-area statistics, in Fig. 6.3. Differences in charging frequency, duration, energy consumption, utilization, and charging power output within various land use types are shown in Table 6.1.



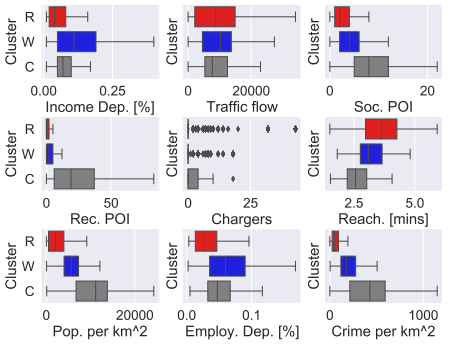
(a) Refined GL Sites per Area



(b) Refined GL Sites per Area Statistics



(c) Refined EDI Sites per Area



(d) Refined EDI Sites per Area Statistics

Fig. 6.3. Site statistics per location (R - residential, W - working/industrial, C - commercial) after refinement with land use data for Glasgow City (GL) and City of Edinburgh (EDI).

Table 6.1
Daily average EV charging session statistics for different locations

City	Location	Daily Sessions	Charge Duration per Session (mins)	Energy Delivered per Session (kWh)	Utilization Rate (%)	Charger Capacity (kW)
Glasgow	Residential	1.35	198.11	19.70	7.36	17.31
	Working/Industrial	1.00	132.87	23.38	4.05	35.95
	Commercial	1.43	167.82	12.58	4.56	16.60
Edinburgh	Residential	1.96	172.86	16.71	5.33	25.42
	Working/Industrial	0.83	134.16	17.98	4.70	18.53
	Commercial	0.96	209.44	14.78	2.68	22.00

6.1.3 Utilization-based Charging Demand Node Labeling

In our effort to improve the placement of EVCSs, we recognize the critical importance of leveraging historical data to inform future infrastructure decisions. To this end, we have developed a methodology that utilizes historical utilization rates of existing charging stations to identify areas with potential for successful EVCS deployment. The cornerstone of our approach is the analysis of EVCS utilization that includes detailed historical charging demand information. By examining this data, we aim to uncover patterns and factors that contribute to the success of charging stations in different locations. This data-driven method allows us to move beyond theoretical assumptions and base our decisions on actual usage patterns, thereby increasing the likelihood of placing new EVCS infrastructure in areas where they are most needed and likely to be well-utilized. To facilitate this analysis, we classify existing charging nodes into three distinct categories based on their historical utilization records: low, medium, and high utilization. This classification serves as a ground-truth labeling system, enabling us to identify the characteristics and contextual factors associated with well-performing charging stations. In calculating EVCS utilization, we adopt an energy-based metric rather than a time-based one, as proposed by [19]. This choice is motivated by the need to account for potential overstay periods and to more accurately reflect the actual usage and efficiency of each charging station. The utilization rate for an EVCS over a T-hour period is calculated as:

$$U_j^T = \frac{1}{c_j * T} \sum_i \epsilon_i^j, \tag{6.1}$$

where U_j^T represents the utilization rate of EVCS j over period T, c_j denotes the power output of EVCS j in kilowatts (kW), and ϵ_i^j signifies the energy consumed by EV i from station j.

Crucially, we extend this classification system to nearby parking infras-

tructure. Parking spots within the charging demand node radius of an existing EVCS are labeled with the same utilization potential as the charging station itself. This process allows us to identify the potential of parking areas for EVCS installation, even if they do not currently have charging facilities. To ensure comprehensive coverage and account for the influence of neighboring areas, we employ a recursive labeling algorithm. Initially, we label parking spots within the immediate radius of existing charging stations based on the historical utilization potential of the station. For subsequent iterations, we consider both the labeled parking spots and the original charging stations. This expanded dataset allows us to label previously unlabeled parking areas that fall within the radius of newly labeled spots. We repeat this process, propagating potential labels across the network of parking infrastructure until reaching convergence, a point where no new parking spots are labeled or changed in an iteration. By analyzing the utilization patterns of existing and potential stations, we can identify areas with similar characteristics that currently lack adequate charging infrastructure. These areas are then classified as having potential for new EVCS placement. We split the data into three balanced classes: low, medium, and high utilization.

6.2 Methodology for Geodemographic-aware EVCS Location Planning for Equitable Placement

Our proposed framework for optimizing EVCS infrastructure placement integrates geodemographic data with a spatially-aware GNN approach, as illustrated in Fig. 6.4. This approach comprises four key components designed to capture the complex dynamics of urban charging demand and inform strategic infrastructure decisions: 1) First, we employ graph representation learning to model and analyze the intricate relationships within the urban charging ecosystem, capturing spatial dependencies and connectivity patterns [Subsecs. 6.2.1, 6.2.2]; 2) Building upon these representations, we uti-

lize a clustering module to categorize nodes based on their characteristics and identify low, medium and high potential installation areas based on historical EVCS utilization [Subsec. 6.2.3]; 3) Finally, we employ a utility-informed site selection process based on area potential, taking into account installation requirements, such as the number of chargers to be installed, their installation utility, and how they affect the overall charging potential of surrounding areas within specific land use [Subsec. 6.2.3]. This integrated approach allows for a holistic evaluation of potential EVCS sites, considering both micro-level factors and macro-level impacts. In the following subsections, we describe each component in detail.

6.2.1 Geodemographic-aware GNN

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represent a graph where $\mathcal{V}$ and $\mathcal{E}$ denote sets of nodes and edges, respectively. In the context of the EVCS placement problem, we define a charging demand node $v \in \mathcal{V}$ as the area within a radius r of an existing EV charger site. The edges $\mathcal{E}$ represent undirected connections between proximate nodes, with two nodes being connected if their physical distance does not exceed r. These edges define the physical and functional connectivity within the network, affecting the flow and demand of EV charging. We define $A \in 0, 1^{|\mathcal{V}|\times|\mathcal{V}|}$ as the adjacency matrix of the graph, where $a_{u,v} = 1$ if an edge exists between nodes u and v, and $a_{u,v} = 0$ otherwise, for all pairs of nodes $u, v \in \mathcal{V}$. Additionally, we set $a_{u,u} = 0$ for all $u \in \mathcal{V}$, as self-loops are not considered in this model.

Each node in the graph is characterized by an F-dimensional feature vector, $\mathbf{h} \in \mathcal{H}$, where $\mathcal{H} \subset \mathbb{R}^{|\mathcal{V}|\times F}$. This feature vector encapsulates a diverse array of information within the radius r of each node, including socio-demographic composition, land use, density and types of POIs, traffic patterns, and existing parking infrastructure. By incorporating this multifaceted dataset into the graph structure, we can develop a sophisticated understanding of the EV charging demand dynamics at each node.

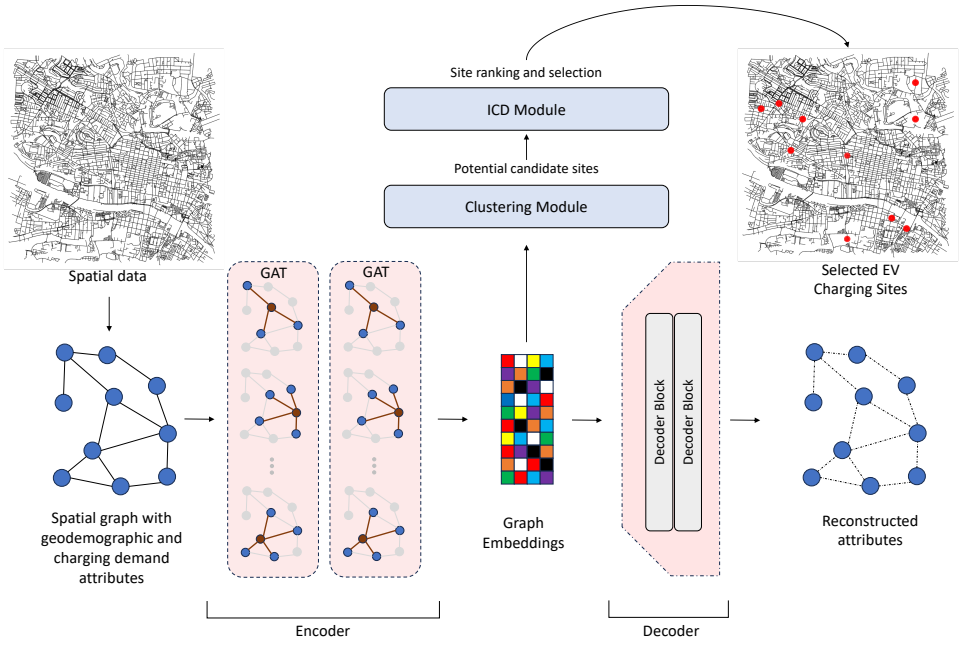


Fig. 6.4. ChargeDEM EV charging site selection approach.

GNN are particularly well-suited for the EVCS placement problem due to their ability to model complex spatial and relational data structures inherent in transportation networks. This approach enables the creation of EVCS placement systems that accurately capture the intricate topology of urban road networks, with nodes representing potential station locations and edges denoting connecting routes. Furthermore, this framework facilitates the seamless integration of heterogeneous data sources, such as geospatial, demographic, land use, and traffic flow information, as node and edge features, providing a comprehensive basis for informed decision-making in the placement of EVCS.

6.2.2 Urban Charging Graph Representation Learning

We approach the challenge of capturing structural information and encoding the relational context of urban charging demand through the lens of graph

representation learning, specifically utilizing graph autoencoders. This approach allows us to effectively capture and leverage the complex relationships inherent in the graph structure of urban charging demand. These relationships encompass a wide array of influencing factors, including spatial characteristics, existing demand patterns, urban vitality indicators, demographic composition, traffic flow dynamics, area reachability, and safety considerations. Using graph auto-encoders, we can distill these multifaceted and interconnected elements into a comprehensive representation that faithfully reflects the intricacies of urban EV charging ecosystems.

The application of autoencoders for unsupervised graph representation learning based on GNNs has been proposed in [68]. An autoencoder architecture typically comprises three key components: an encoder, latent representations, and a decoder. The primary role of an encoder function is to map the input data into a compact latent space, while the decoder attempts to reconstruct the original input from these latent representations. This reconstruction process is guided by a specified reconstruction criterion, ensuring that the learned representations capture essential features of the input data. In the context of graph autoencoders, let f_e denote the graph encoder function and f_d represent the graph decoder function. The fundamental objective of graph autoencoders is to learn the following mappings:

$$H' = f_e(A, X), G' = f_d(A, H'), \tag{6.2}$$

where for an input X, H' denotes the latent space, while G' represents the reconstructed features of the graph computed by the decoder.

[99] introduced completely symmetric graph convolutional autoencoders that leverage both the graph structure and node attributes throughout the entire encoding-decoding process. This approach addresses the instability issues commonly associated with graph convolutional layers by incorporating Laplacian sharpening layers, which counteract the smoothing effects typically observed in these models. While this method effectively resolves several com-

mon graph representation learning challenges, it does not provide a mechanism for dynamically weighting node importance. In the context of GNNs, node representations are typically learned using a set of node features. The conventional GNN approach involves node-level feature aggregation within a defined neighborhood, iteratively learning node representations by aggregating information from neighboring nodes to create a latent representation H'. However, this standard methodology often assumes uniform importance across all neighbors, assigning aggregation weights based solely on degree distance. This uniform weighting approach, while computationally efficient, may not accurately capture the nuanced relationships and varying degrees of influence between nodes in complex networks, such as those representing urban EV charging demand. The inability to dynamically adjust node importance can potentially limit the model's ability to discern and leverage critical patterns in the data, particularly in scenarios where certain nodes or relationships carry disproportionate significance in determining optimal EVCS placement.

To address these limitations, we employ the self-attention mechanism during the encoding phase using Graph Attention Networks (GATs) [132]. Unlike conventional GNNs, GATs dynamically assign weights to neighboring nodes based on their relative importance within the neighborhood, utilizing a masked attention mechanism. The input to a GAT layer is a set of node features $X = X_1, X_2, ..., X_{|\mathcal{V}|}$, where $X_i \in \mathbb{R}^F$ represents the feature vector of node i. The layer then computes an output $H' = H'_1, H'_2, ..., H'_{|\mathcal{V}|}$, where $H'_i \in \mathbb{R}^{F'}$ and cardinality F' may differ from the input feature dimension F. The key innovation of GATs lies in their computation of attention coefficients α_{ij}. These coefficients quantify the importance of the feature vector of node j to node i. The coefficients are computed only for nodes $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ represents a defined neighborhood of node i in the graph.

$$\alpha_{ij} = \frac{exp(a(\mathbf{W}H_i, \mathbf{W}H_j))}{\sum_{k \in \mathcal{N}_i} exp(a(\mathbf{W}H_i, \mathbf{W}H_k))}, \tag{6.3}$$

where $\mathbf{W} \in \mathbb{R}^{F' \times F}$ represent a network weight matrix, and a denotes the attention function. This masked self-attention mechanism allows the model to focus on relevant local structures while ignoring irrelevant or distant nodes.

As an urban graph tends to have a large amount of neighbors, to address the “neighbor explosion" problem often encountered in large urban graphs, we implement a data sampling procedure inspired by [137]. This approach involves obtaining a set of subgraphs by sampling the original training graph and then constructing the graph autoencoder based on these subgraphs. This sampling strategy allows for efficient processing of large-scale urban networks while maintaining the integrity of local structures.

6.2.3 Node Clustering and Site Selection Algorithm

The generated graph embeddings are utilized as input for a k-means clustering procedure, employed to create a classification of the utilization potential of EVCS locations. The classification scheme is based on utilization data, which was categorized into three balanced classes: low, medium, and high utilization. By classifying areas into these utilization categories, we can prioritize medium and high utilization zones for further analysis in the charging station placement process, thereby optimizing the potential impact and efficiency of new EVCS installations.

To identify the best charging location among a set of potential candidate sites, we focus our analysis on medium and high utilization areas identified through clustering techniques. Our primary objective is to maximize area demand coverage by simulating the impact of adding new charging stations that can fully meet existing local demand. To quantify this impact, we introduce Incremental Coverage Difference metric (ICD). The ICD metric evaluates the “usefulness" of potential infrastructure placement by measuring the incremental change in an area’s total charging output when a new station is added. This metric is designed to favor locations where new chargers can completely fulfill the existing demand in the area, thus maximizing the

“impact" of each new installation. The ICD metric is based on the core charging demand factors and requires knowledge of three key elements: 1) total annual EV flow within the demand node, 2) average annual power requirement of an EV, and 3) maximal annual power output of a demand node. To estimate the approximate number of EVs in each city, we multiply the number of registered private vehicles N^c_{CAR} with the assumed 10% EV penetration rate: $N^c_{EV} = 0.1 \times N^c_{CAR}$. Applying this formula to our case study cities yields the following estimates: $N^{Glasgow}_{EV} = 20,480$ and $N^{Edinburgh}_{EV} = 17,850$.

Vehicle Model	**Battery Capacity (kWh)**	**Range (km)**
Tesla Model Y	60-75	455-542
MG4	51-77	349-520
Audi Q4 e-tron	82	455-543
Tesla Model 3	60-78	513-528
Polestar 2	82	555-653
Volkswagen ID.3	62-82	430-558
Kia e-Niro	68	463
BMW i4	83.9	413-589
Volkswagen ID.4	82	515-550
Skoda Enyaq iV	82	538-547

Table 6.2
Battery capacity and range of the most sold EVs in the UK in the year 2023 [114].

To quantify the charging frequency of EVs within city limits, we first establish key parameters based on existing data. Drawing from statistics on the most purchased vehicles in the UK shown in Table 6.2, we set the average EV driving range to $range_{avg} = 466.45$ km and the average battery capacity to $capacity_{avg} = 68.20$ kWh. These figures provide a baseline for our calculations. We then incorporate traffic data from the UK Department of Transport Road Traffic Statistics for 2022, which reports total annual traffic for private cars and taxis as $t^{Glasgow} = 2.684$ billion kilometers in Glasgow City and $t^{Edinburgh} = 2.293$ billion kilometers in the City of Edinburgh. Using

these figures, we calculate the average annual travel distance per car in each city using the formula:

$$d^c_{CAR} = \frac{1}{N^c_{CAR}} t^c [\frac{km}{year}]. \tag{6.4}$$

This yields $d^{Glasgow}_{CAR} = 13{,}105.47\ \frac{km}{year}$ for Glasgow and $d^{Edinburgh}_{CAR} = 12{,}845.94\ \frac{km}{year}$ for Edinburgh.

Assuming that EV users typically recharge their vehicles only when the battery capacity falls below 20%, we can calculate the yearly charging frequency of an EV:

$$f^c_{EV} = d^c_{CAR} \div (range_{avg} \times 0.8)[\frac{charges}{year}]. \tag{6.5}$$

This yields $f^{Glasgow}_{EV} = 35.12 \frac{charges}{year}$ for Glasgow and $f^{Edinburgh}_{EV} = 34.42 \frac{charges}{year}$ for Edinburgh.

Similarly, the average charging need of an EV per year is:

$$e^c_{EV} = f^c_{EV} * (capacity_{avg} * 0.8)[\frac{kW}{year}]. \tag{6.6}$$

which results in $e^{Glasgow}_{EV} = 1916.14 \frac{kW}{year}$ for Glasgow and $e^{Edinburgh}_{EV} = 1877.96 \frac{kW}{year}$ for Edinburgh.

To determine the charging power output of each station, we focus on a circular area with a radius of 500 meters centered on each charging station. This radius is chosen based on previous studies, such as [128], which indicate that 500 meters is generally considered a comfortable walking distance to a charging station. This approach allows us to define charging demand nodes that realistically represent areas where EV users are likely to utilize a given charging station. Within each of these 500-meter radius nodes, we calculate the annual traffic flow, taking into account the assumed 10% EV penetration rate. This calculation is crucial for estimating the potential charging demand in each node i and is expressed as:

$$flow^i_{EV} = 0.1 \times flow^i_{CAR}. \tag{6.7}$$

where $flow^i_{EV}$ represents the annual EV traffic flow within node i, and $flow^i_{CAR}$ is the total annual traffic flow in that node.

Building upon our previous calculations, we next determine the total annual energy requirement for each demand node:

$$C^i_{total} = flow^i_{EV} \times e_{EV}[kWh]. \tag{6.8}$$

To assess the current charging need within each demand node, we calculate its annual output based on the existing infrastructure. This calculation takes into account the average charger power output (O_{avg}) measured in kW, and assumes maximal (24-hour) energy utilization. The current annual output of a demand node is expressed as:

$$C^i_{current} = 24 \times 365 \times no\_chargers^i \times O^i_{avg}[kWh]. \tag{6.9}$$

As a result, the current demand node coverage can be estimated as:

$$C^i = min[\frac{C^i_{current}}{C^i_{total}}, 1][\%]. \tag{6.10}$$

To evaluate the impact of adding a new charging station to a demand node, we calculate the updated coverage of the node. This calculation considers the additional charging power output provided by the new charger, which has an output of c_{out} measured in kilowatts (kW). The new coverage of the demand node after adding this charger is expressed as:

$$C^i_{new} = min[\frac{C^i_{current} + 24 \times 365 \times c_{out}}{C^i_{total}}, 1][\%]. \tag{6.11}$$

In this formula, C^i_{new} represents the new total coverage of demand node i after the addition of the new charger.

To quantify the impact of adding a new charging station to a demand

node, we introduce the ICD. This metric provides a measure of the improvement in output relative to the node's energy requirements. The ICD is calculated using the following formula:

$$ICD^i = C^i_{new} - C^i[\%]. \tag{6.12}$$

As a result, this study adopted a problem setup tailored to identify areas or communities that would benefit most from the installation of a single charging point. Our approach focuses on a methodology that assesses the impact and accessibility for underserved regions rather than following traditional optimization frameworks. We model the urban context by framing a graph-based problem where each potential or existing charging point represents a charging demand node that holds information about existing charging and parking infrastructure, traffic flow, socio-demographic factors, POIs, and charging utilization within the demand node area, as defined in GNN graph construction process explained in Section 6.2.1. Following the node clustering step, as indicated in Fig. 6.4, the potential sites are ranked based on their ICD values. Each potential site is a binary variable: either it gets a charging point or not, while constraints include the number of sites selected, ensuring geographic spread, and the selected nominal power of a charging point. Algorithm 1 presents a systematic method for identifying locations for new charger installations based on their ICD scores. This approach considers the installation of k number of new chargers with a predefined power output and assesses the incremental benefit of adding a new charging station to an existing demand node. The algorithm focuses on a charging demand node area, which contains $n_{charging}$ existing chargers and $n_{parking}$ parking stations. As a result, the computed ICD values are sorted in descending order, as seen in Fig. 6.7 and Fig. 6.8. A higher ICD score indicates a demand node with unmet charging needs due to insufficient infrastructure, where installation of a new charging point would make a significant impact on meeting the area needs, making it a prime candidate for new charger installation. Conversely,

sites that already meet or exceed the local charging demand are excluded from consideration, ensuring efficient resource allocation. This data-driven approach allows for targeted expansion of the charging network, prioritizing areas where new installations will have the most substantial impact on improving charging accessibility and meeting the growing demands of EV users. As a result, by systematically evaluating potential sites based on their ICD scores, urban planners and policymakers can make informed decisions that optimize the distribution of charging infrastructure across the city, ultimately enhancing the overall EV ecosystem.

6.2.4 EVCS Access Equity Evaluation

To address broader transportation justice concerns and ensure that the proposed methodology does not exacerbate existing socio-spatial disparities, Lorenz curves and the associated Gini coefficients were calculated based on data-zone-level EVCS accessibility. Given increasing policy emphasis on just transition in transportation electrification, this metric provides valuable insight in distributional inequality. This is further motivated by previous work in evaluating equity of spatial planning [25, 129]. In the context of EVCS infrastructure, the Lorenz curve plots the cumulative percentage of the population on the horizontal axis against the cumulative percentage of accessibility to EVCS infrastructure on the vertical axis. A perfectly equal distribution corresponds to a 45-degree line, often called the line of equality, where each fraction of the population has equal access to EVCS infrastructure. The area between the Lorenz Curve and the line of equality quantitatively captures inequality within the distribution, where a larger area indicates greater inequality. Formally, the Lorenz Curve $L(p)$ can be defined mathematically as:

$$L(p) = \frac{\int_0^p F^{-1}(q)\,dq}{\int_0^1 F^{-1}(q)\,dq}.$$

$L(p)$ is the value of the Lorenz curve at percentile p, $F^{-1}(x)$ is the inverse cumulative distribution function of EVCS accessibility values, while p represents the proportion of the population. The Lorenz Curve facilitates the computation of numerical measures of inequality, such as the Gini coefficient, which quantifies the deviation of the observed distribution from perfect equality. The Gini coefficient is calculated as twice the area between the line of equality and the Lorenz Curve:

$$\text{Gini} = 1 - 2 \int_0^1 L(p)\, dp.$$

A Gini coefficient of 0 represents perfect equality (the Lorenz curve follows the line of perfect equality), while a value of 1 indicates maximum inequality.

Finally, it is important to note that, although not a conventional optimization formulation, the presented approach could be viewed as a greedy search method based on GNN embeddings. The objective function includes the equity aspect through calculation of Gini coefficient, in addition to the coverage delta calculated through Gini index. Decision variables depend on the parking lots that are assigned as new charging stations, while constraints include the number of new stations installed, as well as the charger nominal power.

Algorithm 1 Site Selection Algorithm for New Charging Stations

Require: Charging nodes $\mathcal{V}$, Number of chargers to be installed k, New chargers power output c_{out} [kW]

Initialize list of new stations: $newStations \leftarrow []$
for $i = 1$ to k **do**
  $maxICD \leftarrow 0$
  $maxICDStation \leftarrow null$
  for each node $v \in \mathcal{V}$ **do**
    Calculate $ICD^v (Eq.6.12)$
    if $v_{type} == parking$ and $n^v_{parking} \geq 1$ **then**
      if $ICD^v > maxICD$ **then**
        $maxICD \leftarrow ICD^v$
        $maxICDStation \leftarrow v$
      end if
    end if
  end for
  Add site $maxICDStation$ to $newStations$
  for $v \in \mathcal{V}$ **do**
    Calculate the distance $dist^v$ from node v to $maxICDStation$
    if $dist^v \leq r$ **then**
      $n^v_{parking} \leftarrow n^v_{parking} - 1$
      $n^v_{charging} \leftarrow n^v_{charging} + 1$
      Recalculate node coverage given newly added c_{out} (Eq. 6.11)
    end if
  end for
end for
return $newStations$

6.3 Experimental Results and Discussion

6.3.1 Clustering Evaluation Metrics

To assess the quality of utilization-based clustering, we use three common performance evaluation metrics: Accuracy, Adjusted Rand Index (ARI) [88], and Normalized Mutual Information (NMI) [117]. Accuracy evaluates classification accuracy within three possible utilization classes, while the Rand Index (RI), defined as:

$$RI = \frac{p+q}{\binom{n}{2}}, \tag{6.13}$$

calculates a similarity between two cluster results by comparing all points within the same cluster. p is the number of pairs correctly placed in the same cluster, q is the number of pairs correctly placed in different clusters, and n is the total number of elements. Adjusted Rand Index (ARI), defined as:

$$ARI = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]}, \tag{6.14}$$

extends RI by accounting for different models of random clusterings, with values ranging from approximately 0 for random labeling to 1 for perfect agreement. Normalized Mutual Information (NMI), calculated as:

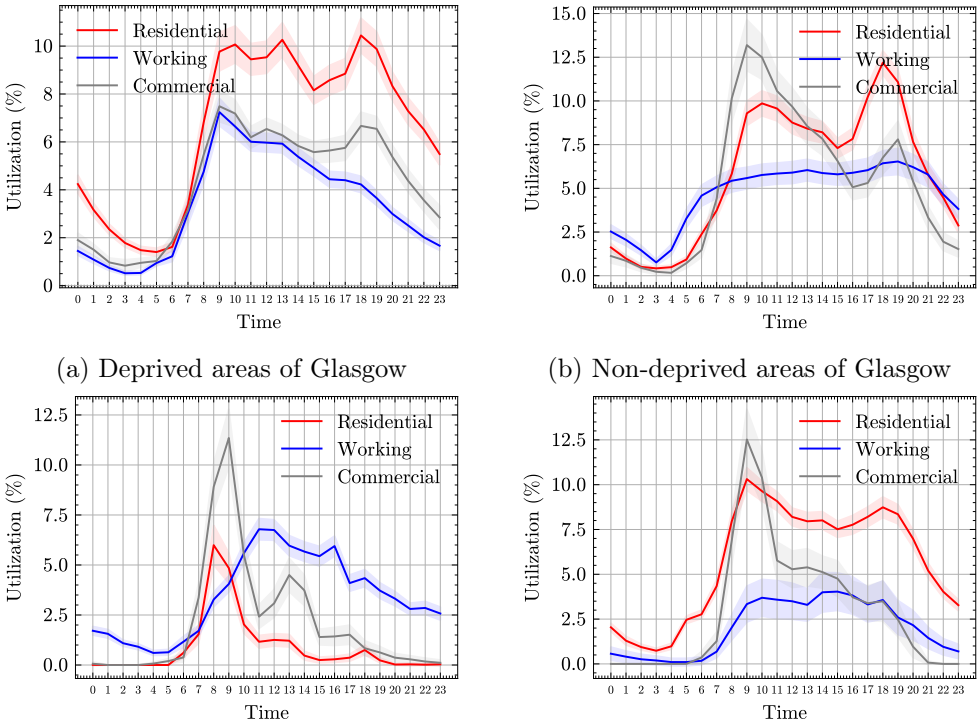
$$NMI = 2 \times \frac{I(U,V)}{H(U) + H(V)}, \tag{6.15}$$

quantifies the shared information between predicted and true clusterings, where $I(U,V)$ is the mutual information and $H(U)$, $H(V)$ are the entropies of the clusterings. NMI ranges from 0 (no sharing) to 1 (perfect correlation).

The above complementary metrics offer a comprehensive quantitative assessment of our algorithm's clustering performance, capturing different aspects of the results' quality and reliability.

6.3.2 Results: EVCS Land Use Identification and Statistical Analysis

A key component of our proposed approach involves the localization of land uses within the council limits of Glasgow City and the City of Edinburgh. The land use clustering procedure, as illustrated in Fig. 6.3, reveals distinct patterns in local land use characteristics across both cities. In Glasgow and Edinburgh, the analysis identifies commercial areas predominantly within the city centers. These zones correlate strongly with a high concentration of social and especially recreational POIs, reflecting the diverse entertainment options typically found in urban cores. Notably, these commercial areas also coincide with lower car traffic volumes, likely influenced by the implementation of low-emission zones in both city centers. The distribution of EVCS infrastructure shows a significant concentration within commercial zones. This pattern may indicate a potential saturation of charging facilities in these areas, suggesting a need for strategic reassessment of future EVCS placements. Commercial zones also exhibit lower rates of income (measured by the percentage of the population receiving income support) and employment deprivation (measured by the percentage of the population who are employment deprived), but also the highest population density and the highest incidence of reported crime, aligning with typical urban center characteristics. Working areas, as identified by our analysis, are characterized by higher traffic counts and notably, the highest rates of income and employment deprivation. The distribution of EVCS infrastructure in working areas appears less consistent, with some zones showing a lack of facilities compared to residential areas, while others contain large concentrations of deployed EVCS infrastructure. Residential areas, in contrast, generally report lower levels of traffic in Glasgow, while showing high variability in Edinburgh, possibly due to a large number of residential housing along major roads. Results indicate that EVCS infrastructure is severely lacking in residential areas in both cities.



(a) Deprived areas of Glasgow

(b) Non-deprived areas of Glasgow

(c) Deprived areas of Edinburgh

(d) Non-deprived areas of Edinburgh

Fig. 6.5. Average hourly EVCS utilization statistics per area

Observing per-area hourly EVCS utilization shown in Fig. 6.5, we quantify distinct patterns in EVCS usage across Glasgow and Edinburgh, highlighting urban disparities and lifestyle differences. Glasgow generally exhibits higher utilization rates than Edinburgh, particularly in residential charging within deprived areas. In Glasgow, deprived residential areas show similar utilization compared to non-deprived areas. This suggests possible under-utilization of non-deprived charging, explained by the fact that the majority of Glasgow falls within deprived areas, as seen in Fig. 6.7. The city center falls within non-deprived areas, where the majority of EVCSs are located. Thus, the results could indicate a general under-utilization of city center EVCS charging. In terms of time patterns, the utilization generally peaks around 6 PM, coinciding with the approximate time of commute home from

work. Commercial areas indicate utilization peak around 9 AM, possibly due to the large amount of commuters parking their cars within the city center charging locations, where there is the highest overall recorded utilization peak. Working areas of Glasgow showcase similar utilization patterns as commercial areas in deprived areas, peaking in the morning hours and gradually lowering the utilization, which is explained by workers commuting back home. In non-deprived areas, this trend is less apparent and the utilization is relatively consistent throughout the working hours.

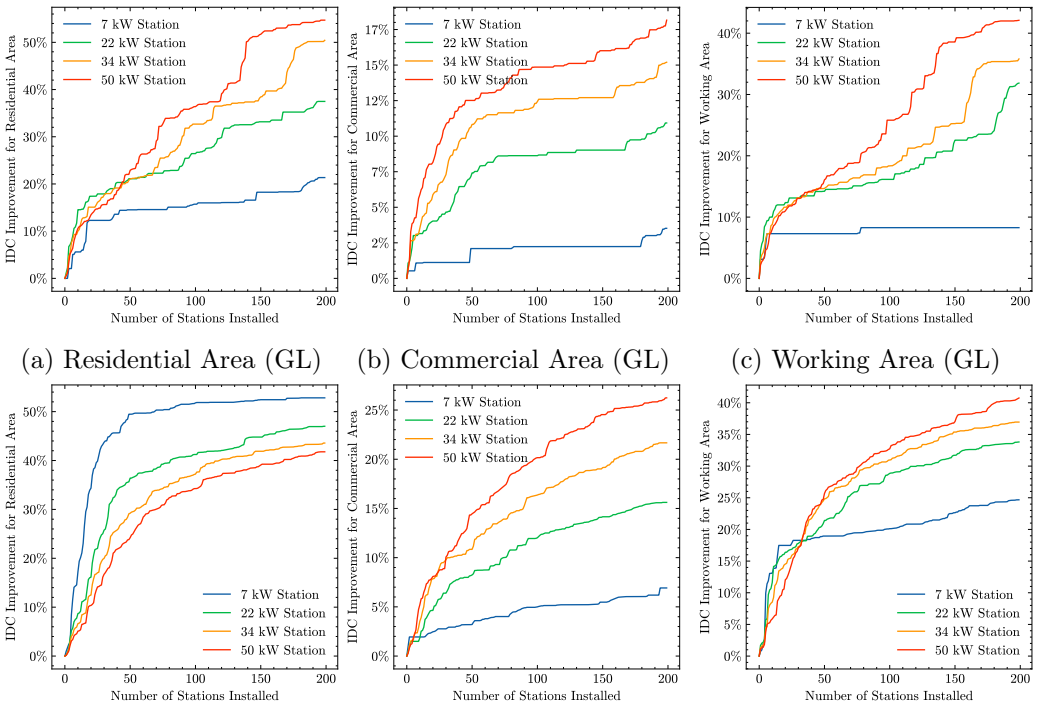
In Edinburgh, the most striking difference in utilization is between deprived and non-deprived residential areas. While non-deprived utilization peaks at around 10% in the morning, deprived areas experience only 5% utilization, which rapidly lowers throughout the day. This indicates limited access to public charging within deprived areas of Edinburgh, which has a direct impact on utilization due to high overstay periods and limited charging opportunities in the second half of the day. Utilization in commercial areas is relatively consistent between deprived and non-deprived communities, peaking in the morning, and sharply declining until the end of the day. Interestingly, utilization in working areas, while lower, experiences a similar pattern between non-deprived working areas of Glasgow and Edinburgh, indicating similar charging behavior within these areas.

6.3.3 Results: Utilization-based Clustering

In this subsection, we present detailed quantitative analysis of our proposed placement methodology, focusing on the performance of our GNN-based clustering approach for selecting potential candidate sites based on their utilization potential, as described in Subsection 6.1.3. Table 6.3 compares the performance of our proposed method against several widely-used clustering techniques, including K-means, Spectral, and Hierarchical Agglomerative Clustering, as well as other graph-based approaches such as [50]. This comparison provides a comprehensive quantitative evaluation of our approach

within the broader context of clustering methodologies. K-means and spectral clustering analyse the data based solely on its inherent features, offering a baseline for traditional clustering techniques. In contrast, GraphSAGE employs an unsupervised graph representation technique, aggregating feature information from a node's location and environment.

The performance discrepancies observed in Table 6.3 between traditional clustering methods and our proposed GNN-based approach highlight the inherent complexity of identifying high charging demand areas for EV infrastructure. Traditional clustering techniques, while effective in many scenarios, struggle to capture the non-linear relationships between the diverse factors influencing charging demand. This limitation is particularly evident in Glasgow, where the accuracy of conventional methods is notably lower, primarily due to the city's intricate urban topology and mixed-use nature of many neighborhoods. Such a diverse and interconnected urban landscape makes it challenging to delineate clear boundaries between high, medium, and low utilization areas using conventional clustering techniques. The blending of different land uses and activities creates a more nuanced charging demand pattern that requires a more sophisticated analytical approach. In contrast, Edinburgh presents a somewhat easier scenario for traditional clustering methods. The city's urban structure exhibits a clearer separation between high, medium, and low utilization areas, likely due to a more distinct spatial organization of land use areas. Our GNN-based approach demonstrates superior accuracy by effectively capturing the spatial relationships between charging demand nodes. By propagating information through the graph structure, the GNN incorporates broader contextual information about the surrounding area, which is crucial for understanding the intricate dynamics of urban charging demand. This ability to account for complex spatial interactions and the multifaceted nature of factors influencing EV charging demand allows our method to outperform traditional techniques, particularly in complex urban environments like Glasgow.



(a) Residential Area (GL) (b) Commercial Area (GL) (c) Working Area (GL)

(d) Residential Area (EDI) (e) Commercial Area (EDI) (f) Working Area (EDI)

Fig. 6.6. ICD performance results for Glasgow (GL) and Edinburgh (EDI)

Table 6.3
Clustering performance for Glasgow (top) and Edinburgh (bottom)

Algorithm	Accuracy	ARI	NMI
Kmeans	0.2566	0.0015	0.0009
Spectral	0.3837	0.0394	0.0213
Agglomerative	0.2020	0.0017	0.0002
GraphSage	0.3589	0.0124	0.0079
Proposed (Ours)	0.5770	0.0458	0.0138

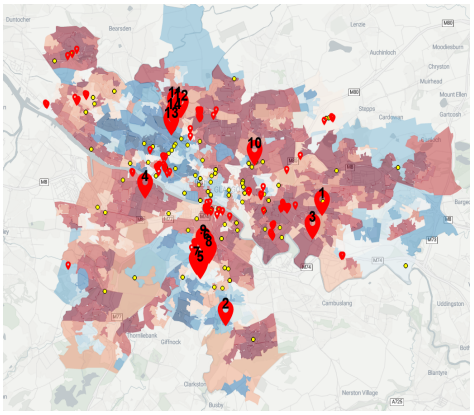
Algorithm	Accuracy	ARI	NMI
Kmeans	0.2383	0.0682	0.0520
Spectral	0.3262	0.0279	0.0303
Agglomerative	0.3262	0.0656	0.0510
GraphSage	0.3817	0.0293	0.0264
Proposed (Ours)	0.6258	0.1467	0.0838

6.3.4 Discussion: EVCS Localization

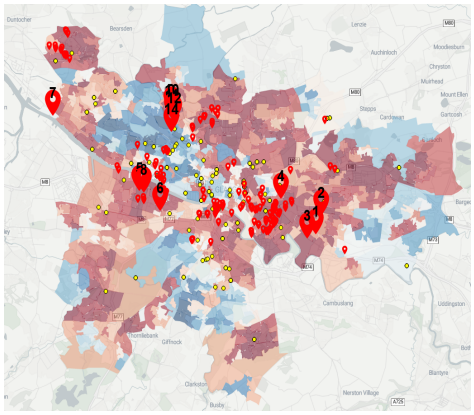
The three proposed placement strategies (working/industrial, residential and commercial) are defined based on the land use processing described in the study (see Subsection 6.1.2). The “residential area" policy aims to populate public parking available in residential areas with appropriate power output of charging stations based on their ICD metric values. Similarly, the other two policies focus on commercial and working/industrial areas. As shown in Fig. 6.6, the proposed model behaves differently depending on the proposed land use and battery capacity. In residential areas, noticeable differences in behavior were observed between Glasgow and Edinburgh. In Glasgow, the installation of higher power output stations (34 kW and 50 kW) significantly improved ICD, with the 50 kW stations achieving around 55%. Conversely, lower power output stations (7 kW) showed more modest increases, highlighting the limited efficacy of low-power stations in residential urban settings, likely due to already sufficient residential charging within the city limits. On the other hand, 22 kW stations showed the best improvement over the initial 50 installations, indicating high potential for 22 kW infrastructure placement in Glasgow's residential areas.

Observing the EVCS placements shown in Fig. 6.7, the east end of Glasgow has been identified as a high-potential area, particularly the Shettleston constituency (Fig. 6.7.a - rank 1,3, and Fig. 6.7.b - rank 1,2,3), which is a good site for fair infrastructure placement, providing much-needed infrastructure to region scoring higher than average on the deprivation index. Additionally, this area also offers high utilization potential due to its proximity to large supermarkets and parks. The residential EVCS placement with rank 2 is located in an underserved area near major Linn Park, filling the infrastructure gap between the non-deprived area in the northwest and the deprived area in the southeast of the potential EVCS placement location. Other high-potential residential areas include the borders of Maryhill and Canal wards (Fig. 6.7.a - rank 11-14), as well as Southside Central ward

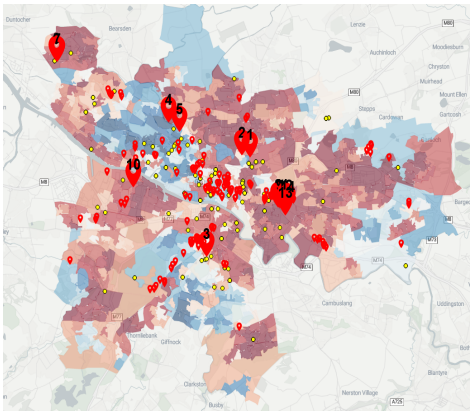
areas around Queens Park, known for its many restaurants and other POIs (Fig. 6.7.a - rank 5-9). Interestingly, Edinburgh exhibited a more pronounced response to slow residential charging, with 7 kW stations facilitating up to a 50% improvement in ICD, indicating that residential areas of Edinburgh significantly lack needed infrastructure that would fulfill demand by installation of slow overnight charging. To this end, as suggested in [33], utilization of existing lamp posts for on-street EV charging in residential areas might be an effective method for faster expansion of slow overnight charging.



(a) Residential 22kw Glasgow



(b) Working Area 22kW Glasgow

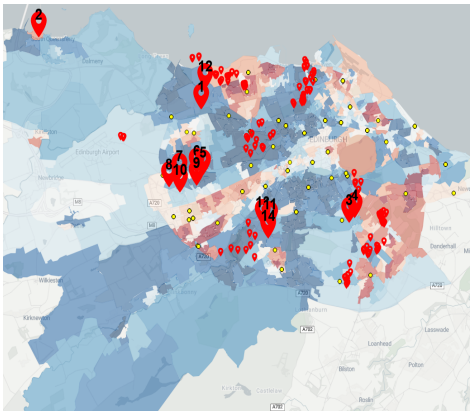


(c) Comercial Area 34kW Glasgow

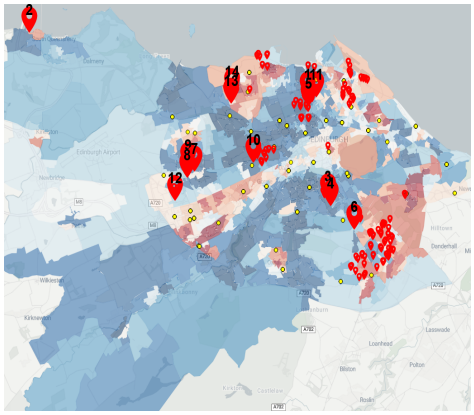
Fig. 6.7. Proposed EVCS infrastructure placement sites in Glasgow. The top 15 results are displayed with a number on top. Yellow markers are existing charging stations in the selected areas. Red-blue overlay corresponds to area deprivation, where blue is less deprived and red more deprived.

Policy focused around the installation of slow overnight charging is especially effective for deprived areas of Edinburgh, where statistical analysis performed in Subsection 6.3.2 showed a general lack of infrastructure which has a further pronounced impact on overall EVCS utilization (as seen in Fig. 6.5) due to high overstay periods resulting in unavailability of charging options. Observing the results in Fig. 6.8, the ward of Almond (Fig. 6.8.a - rank 2) shows high potential for EVCS placement due to its lower-than-

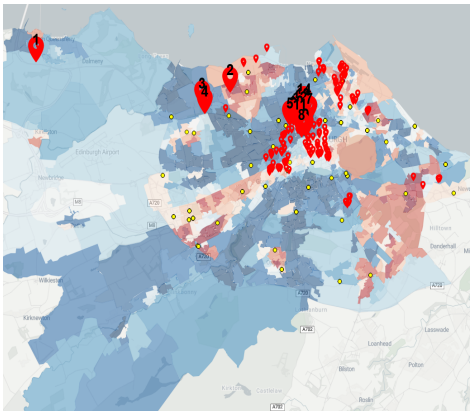
average EVCS infrastructure numbers, as well as the Liberton area (Fig. 6.8.a - rank 3-4), which is close to major roads and large retail areas. The non-deprived Corstorphine area (Fig. 6.8.a - rank 5-10), featuring numerous shops and major roads, as well as Edinburgh Zoo, also shows great potential for 7-22kW residential charging.



(a) Residential 22kW Edinburgh



(b) Working Area 22kW Edinburgh

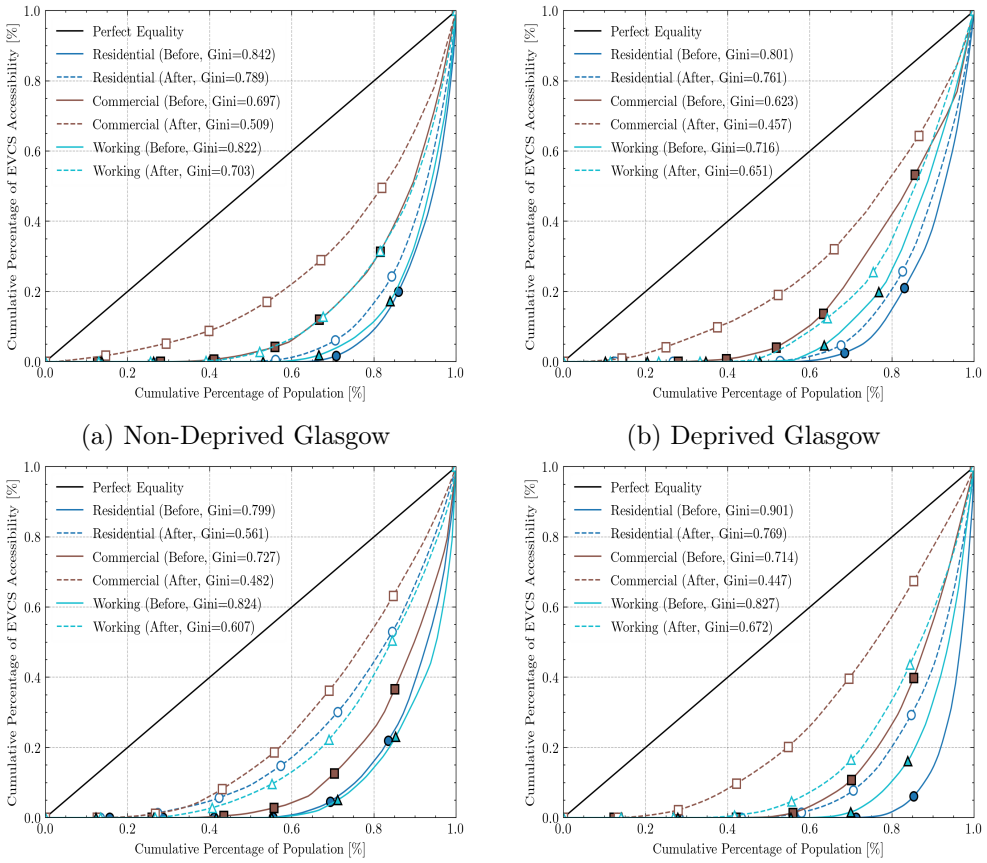


(c) Commercial Area 34kW Edinburgh

Fig. 6.8. Proposed EVCS infrastructure placement sites in Edinburgh. Top 15 results are displayed with a number on top. Top 15 results displayed with a number on top. Yellow markers are existing charging stations in the selected areas. Red-blue overlay corresponds to area deprivation, where blue is less deprived and red more deprived.

When it comes to EVCS placement in working and industrial areas of

Glasgow, the results suggest a similar strategy to residential areas, where 22 kW installations might provide the best utility-to-cost ratio, with around a 30% overall ICD improvement. Based on Fig. 6.7, high-potential areas include deprived areas at the eastern end of Shettleston (Fig. 6.7.b - rank 1-3), close to large factories, and the Haghill area near a university campus and a large retail park (Fig. 6.7.b - rank 4), largely due to a lack of infrastructure and high charging demand. Other high-potential areas include the Govan area (Fig. 6.7.b - rank 5,6,8), which hosts a large recycling center, business centers, the UK Visa and Immigration center, warehouses, and a subway depot, and lacks general charging infrastructure. Overall, Edinburgh displays higher potential for ICD improvement within working and industrial areas. While maximum improvement for fast charging is similar, 7kW EVCS placements display large improvements over the initial 50 installations. The highest potential lies in the South Queensferry area (Fig. 6.8.b - rank 2), an area of high importance due to the construction of a new bridge carrying the M90 motorway, connecting northern Scotland with Edinburgh. Contrary to residential land use, the difference in utility gained from installing slow 7kW and other charging stations, while low in Edinburgh, is significant in Glasgow, suggesting a need for rapid infrastructure in the working and industrial areas. This insight is invaluable for stakeholders, indicating that working and industrial areas of Glasgow, often close to major roads, are lacking in EVCS infrastructure.



(a) Non-Deprived Glasgow

(b) Deprived Glasgow

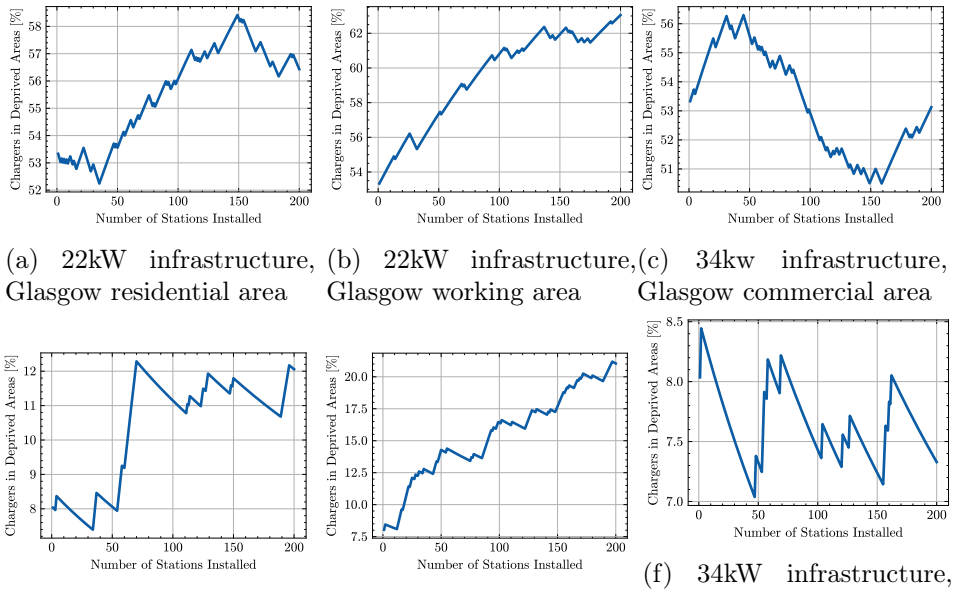
(c) Non-Deprived Edinburgh

(d) Deprived Edinburgh

Fig. 6.9. Lorenz curves for EVCS accessibility across different land use and deprivation areas before and after the proposed approach.

The proposed policy for infrastructure placement in commercial areas shows the lowest potential across Glasgow and Edinburgh. In Glasgow, the installation of higher power output stations (34 kW and 50 kW) demonstrated a modest improvement in ICD, with the 50 kW stations achieving around 17%. In contrast, Edinburgh achieved an improvement of more than 25% with 50 kW charging stations. Interestingly, the relative improvement between the desired charging capacities is similar between the two cities, suggesting predictable behavior within commercial areas due to similarities

in existing infrastructure, residential density, presence of points of interest, and traffic behavior. The preference for higher EVCS output suggests a potential saturation of slow EV charging and the need for faster infrastructure. In Glasgow, areas away from the city center are favored, which is a welcome policy, as the city center is designated as a low-emission zone. High-potential areas include deprived areas in Dennistoun, close to a large retail park, and retail centers in the deprived Drumchapel area. In Edinburgh, high-ranking areas include retail areas in South Queensferry, Silverknowes, and the area close to the historic city center.



(a) 22kW infrastructure, Glasgow residential area

(b) 22kW infrastructure, Glasgow working area

(c) 34kw infrastructure, Glasgow commercial area

(d) 7kW infrastructure, Edinburgh residential area

(e) 22kW infrastructure, Edinburgh working area

(f) 34kW infrastructure, Edinburgh commercial area

Fig. 6.10. Increase in the ratio of chargers in deprived zones per land use area after the proposed infrastructure placement strategy.

6.3.5 Discussion: EVCS Placement and Equity in Access to EVCS

The Lorenz curves shown in Fig. 6.9 reveal profound inequities in EVCS accessibility across different urban contexts. Prior to implementation of the new placement strategy, residential areas exhibited the most severe inequality, with high Gini coefficients ranging from 0.799 to 0.901, with deprived Edinburgh representing the most extreme case. The steep curvature of these residential lines indicates that approximately 80% of the population has access to less than 20% of available charging infrastructure, creating significant accessibility deserts. Commercial zones, while still inequitable, demonstrated relatively better distribution (Gini coefficients 0.623-0.727), potentially reflecting the concentration of existing infrastructure in business districts and retail centers. Working areas similarly suffered from poor distribution (Gini coefficients 0.716-0.827). Notably, a clear socioeconomic gradient emerged, with deprived areas consistently experiencing higher inequality than their non-deprived counterparts, except for commercial zones in Glasgow where the pattern was reversed. Geographic disparities were also evident, with Edinburgh displaying more extreme inequality than Glasgow, particularly in residential contexts. This landscape reveals systemic biases in EVCS distribution that likely reflect broader patterns of infrastructure investment prioritizing commercial centers and affluent neighborhoods, while neglecting residential and working areas, especially in deprived communities. The extreme bowing of most curves indicates severe concentration of accessibility resources, potentially creating substantial barriers to EV adoption among disadvantaged populations and reinforcing existing transportation inequities.

Following implementation of the proposed placement strategy, the Lorenz curves reveal significant improvements in EVCS accessibility distribution across all urban contexts. Commercial areas exhibited the most dramatic transformation, with Gini coefficients falling to the 0.447-0.509 range, representing the closest approximation to equitable distribution among all land

use types. This substantial improvement indicates that strategic placement effectively countered pre-existing commercial concentration biases. Working areas showed moderate improvements (post-intervention Gini coefficients 0.607-0.703), with Non-Deprived Edinburgh experiencing the most substantial gains. Residential zones, while improved, retained the highest level of inequality (Gini coefficients 0.561-0.789), suggesting these areas remain the most challenging for achieving equitable EVCS distribution—likely due to complex residential density patterns and infrastructure limitations. Importantly, these positive outcomes stand in sharp contrast to what would have occurred with a poor placement strategy. Had the proposed EVCS installation strategy merely reinforced existing infrastructure patterns or prioritized areas with already sufficient coverage, Gini coefficients would have increased rather than decreased, further exacerbating socioeconomic and geographic disparities. Encouragingly, the intervention diminished the socioeconomic gradient, with deprived areas experiencing proportionally larger improvements than non-deprived areas in most contexts. This is particularly relevant for the case of Edinburgh, where the non-deprived residential areas achieved the lowest post-intervention Gini coefficient (0.561) among all residential contexts. While perfect equity remains unachieved, the consistent flattening of all Lorenz curves demonstrates that the proposed strategic placement can substantially redistribute accessibility resources. The post-intervention landscape shows that approximately 60% of the population now has access to 20-30% of charging infrastructure (compared to under 10% pre-intervention), representing a significant step toward more inclusive EV infrastructure development, though persistent gaps indicate ongoing challenges in achieving truly equitable distribution. Fig. 6.10 provides critical insight into the mechanism behind these equity improvements, illustrating how the proposed station ranking method progressively affects the proportion of chargers in deprived areas during installation. Glasgow's working areas (Fig. 6.10b) show the most consistent upward trajectory, increasing from approximately

54% to 62% charger presence in deprived areas, explaining the substantial Gini coefficient improvement in this context. Similarly, Edinburgh's working areas (Fig. 6.10e) demonstrate the most dramatic proportional increase, more than doubling from 7.5% to around 20%. The fluctuating patterns in commercial areas (Fig. 6.10c, Fig. 6.10f) align with their more moderate equity improvements, while residential installations show varied patterns between cities, with Glasgow (Fig. 6.10a) exhibiting earlier prioritization of deprived areas compared to Edinburgh, with sharp increase after initial 50 installations (Fig. 6.10d).

6.3.6 Discussion: Scalability and Transferability

The proposed methodology can be adapted across different cities, and more readily within the UK, including cities in England and Wales. Direct application is possible due to equivalent socio-economic indicators - the Index of Multiple Deprivation (IMD) for England and the Welsh Index of Multiple Deprivation (WIMD) for Wales are directly comparable to the SIMD data used in our study. These indices share similar underlying domains including income, employment, health, education, and geographical access, enabling consistent node feature construction across UK cities. Outwith the UK, while specific deprivation metrics may differ, a mapping of pertinent deprivation metrics to the application domain can be made. Regardless, the core methodology of constructing charging demand nodes and their relationships, as described in Section 6.2, is applicable to all application domains. The approach requires three fundamental data categories that are typically available in most urban areas: (1) spatial infrastructure data (obtainable through OpenStreetMap), (2) socio-economic indicators (available through national census or similar demographic surveys), and (3) mobility patterns (accessible through traffic counts or similar transportation data). Cities lacking historical EVCS utilization data could initially calibrate the model using proxy metrics such as vehicle ownership rates, parking utilization, or traf-

fic flow patterns. The graph construction methodology, based on 500-meter radius nodes and their spatial relationships, is geography-agnostic and can be applied to any urban environment. The GNN architecture itself is flexible enough to accommodate varying numbers and types of input features, allowing for adaptation to locally available data while maintaining the core principles of geodemographic-aware infrastructure planning.

6.4 Summary

This research presents a novel geodemographic aware approach to EVCS placement through GNN modelling. By fusing socio-demographic data, spatial dynamics, and post-installation impacts, our methodology addresses the critical gaps in existing infrastructure planning strategies. The case study of Glasgow and Edinburgh demonstrates the effectiveness of this approach, optimizing EVCS placement for efficiency and equity. Key advantages of using GNNs include consideration of underserved communities, nuanced understanding of urban dynamics, and maximization of new charging station utilization. Experimental results validate the utility of the proposed method, showing significant improvements in strategic placement and use of EV charging stations. The proposed GNN-based approach demonstrates strong scalability potential for larger urban environments. Our model leverages GATs which are inherently more efficient than traditional GNNs due to their selective attention mechanism that focuses on important node relationships, and can also mitigate over-squashing issues related to large-scale graphs [6]. From a computational perspective, the scalability of GNN architectures has been demonstrated in substantially larger applications, including citation networks with millions of nodes, social networks with billions of edges, and molecular graphs analyzing hundreds of thousands of compounds. The urban context, being relatively constrained in comparison, with node numbers in the range of thousands, presents a more computationally manageable en-

vironment. However, we acknowledge certain limitations of our work. The primary limitations center around the temporal analysis constraints, as our model primarily focuses on static spatial patterns and does not fully incorporate temporal aspects such as seasonal fluctuations or long-term EV adoption trends. Building on this, to improve estimations of the potential EV count used for ICD calculation, demographic profiling of potential EV consumers alongside comprehensive surveys could be utilized. Additionally, notable gaps include the absence of electrical grid capacity considerations and associated infrastructure upgrade requirements, which could significantly impact implementation feasibility, as well as aspects related to renewable energy availability and the carbon footprint associated with charging infrastructure deployment, both of which could greatly impact sustainability outcomes. Lastly, refining accessibility calculations at specific charging demand nodes could provide more precise insights, enabling more targeted analysis and improved placement strategies at localised levels.

Chapter 7

Conclusion and Future Work

7.0.1 Conclusion

The goal of this thesis is to improve trustworthiness of AI systems within Smart Grid management, firstly driven by a fundamental challenge: the inherent instability of XAI methods. Recognizing that true trust in AI cannot be built on unverified explanations, our initial step was to establish rigorous quantitative measures for explainability, specifically within the demanding context of NILM. This work provided a robust framework to not only visualize but also numerically assess the faithfulness, robustness, and complexity of explanations, transforming XAI from a purely observational tool into a verifiable component of AI design. With this foundation, uncovering an actionable link between transparency and robustness became a goal. We demonstrate that explainability is not merely a desirable post-hoc feature but can be an instrument for enhancing the robustness of NILM models. This was pursued through two novel approaches: First, by directly embedding explainability principles into the training process through use of regularization techniques. By actively guiding the NILM models towards explanations that are not only more transparent but also demonstrably more robust to perturbations and variations in input data, we show that it can lead to improved model ro-

bustness. Second, a novel explainability-guided knowledge distillation mechanism was introduced. Moving beyond traditional distillation where student models merely mimic teacher predictions, our approach conditioned the student to also replicate the teacher's explanations. This proved transformative, showing that by ensuring the faithful transfer of explanations, without any gradient regularization, even compact student models could achieve significant gains in robustness and interpretability, vital for resource-constrained edge deployments. The resulting framework of quantifying explainability and leveraging it to bolster robustness provides a powerful toolkit in the context of NILM. This was extended to address the equally critical, yet distinct, challenge of fairness and equity in the broader energy transition. Our work on GNN-informed EV charging station placement, while a different domain, was fundamentally informed by the principles of developing transparent, robust, and now equitable AI. By integrating geodemographic factors and systematically aiming for fair access, we demonstrated how the lessons learned in building trustworthy AI components can be applied to ensure that technological advancements in smart grids benefit all segments of society. In addressing the core principles of robustness, transparency, and fairness, this work also indirectly contributed to other facets of Trustworthy AI. The advancements in explainability bolster human oversight and are a step towards greater accountability. The research into edge deployment inherently supports data privacy. Furthermore, the focus on equitable infrastructure planning directly promotes societal well-being and non-discrimination, ensuring that the benefits of the energy transition are more broadly shared. In essence, this research has laid out the groundwork for Trustworthy AI in Smart Grids by showing how trustworthy principles can be integrated in applications such as NILM and EV charging placement problem to build more robust, transparent and ultimately fairer AI systems for the energy transition.

7.0.2 Future Work

Several promising directions for future research emerge from this work. First, it is important to extensively explore the relationship and trade-offs between the properties of faithfulness, robustness and complexity in XAI NILM approaches. For example, a highly faithful explanation that closely reflects the model's behavior may be more complex and harder to understand. Conversely, a simpler explanation may be more accessible but less faithful to the model's true decision-making process. Similarly, there may be cases where faithful explanations are sensitive to small changes in input data, resulting in a trade-off between faithfulness and robustness. Thus, striking the right balance between the metrics of explanation quality is crucial to ensure the usefulness of the XAI system. As one of the challenges in deploying NILM systems is the need for real-time processing and interpretation of energy consumption data, investigating the feasibility of real-time XAI methods for NILM applications would be a valuable contribution to the field, enabling more practical and actionable insights for users. Next, future work could investigate methods to incorporate direct human feedback or domain knowledge into the learning process through active learning and similar approaches, further strengthening the human agency and oversight principles. Additionally, future work could explore dynamic distillation strategies that adapt to specific appliance characteristics and operational conditions, as well as methods to reduce the computational overhead of explanation generation while maintaining explanation quality. Furthermore, the XAI methods employed primarily provide feature attribution, indicating what input features are important. They do not fully delve into the causal relationships or the internal reasoning at a mechanistic level. Future work could explore integrating causal inference techniques, counterfactual explanations or mechanistic interpretability to provide deeper insights into why models learn specific patterns and how their internal decision thresholds are formed. Lastly, in the domain of equitable infrastructure placement, future research could incorporate tem-

poral dynamics, analysing grid stability impacts, examine carbon footprint implications, investigate relationships with other transportation modes, and perform comparative analysis across different urban environments. Promising research directions include studying EVCS placement effects on power grid stability, enhancing GNN models to include alternative transportation options, interpretability techniques for GNNs and developing reinforcement learning frameworks for dynamic charging recommendations. Lastly, proximity to brownfield sites presents an opportunity for more sustainable placement decisions that could help mitigate potential grid constraints.

Bibliography

[1] Global EV outlook 2023 – analysis.

[2] Lennart Adenaw and Sebastian Krapf. Placing BEV charging infrastructure: Influencing factors, metrics, and their influence on observed charger utilization. 13(4):56.

[3] Lennart Adenaw and Sebastian Krapf. Placing bev charging infrastructure: Influencing factors, metrics, and their influence on observed charger utilization. *World Electric Vehicle Journal*, 13(4):56, 2022.

[4] Tanveer Ahmad, Rafal Madonski, Dongdong Zhang, Chao Huang, and Asad Mujeeb. Data-driven probabilistic machine learning in sustainable smart energy/smart energy systems: Key developments, challenges, and future research opportunities in the context of smart grid paradigm. *Renewable and Sustainable Energy Reviews*, 160:112128, 2022.

[5] Shamim Ahmed and Marc Bons. Edge computed nilm: a phone-based implementation using mobilenet compressed by tensorflow lite. In *Proceedings of the 5th international workshop on non-intrusive load monitoring*, pages 44–48, 2020.

[6] Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021.

[7] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv:1806.08049*, 2018.

[8] David Alvarez-Melis and Tommi S. Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks, December 2018.

[9] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv:1711.06104*, 2017.

[10] Georgios-Fotios Angelis, Christos Timplalexis, Stelios Krinidis, Dimosthenis Ioannidis, and Dimitrios Tzovaras. Nilm applications: Literature review of learning approaches, recent developments and challenges. *Energy and Buildings*, 261:111951, 2022.

[11] Paul Arévalo and Francisco Jurado. Impact of artificial intelligence on the planning and operation of distributed energy systems in smart grids. *Energies*, 17(17):4501, 2024.

[12] K Carrie Armel, Abhay Gupta, Gireesh Shrimali, and Adrian Albert. Is disaggregation the holy grail of energy efficiency? the case of electricity. *Energy policy*, 52:213–234, 2013.

[13] UN General Assembly. Transforming our world: the 2030 agenda for sustainable development, 21 october 2015. *Retrieved from*, 2015.

[14] Sotirios Athanasoulias, Fernanda Guasselli, Nikolaos Doulamis, Anastasios Doulamis, Nikolaos Ipiotis, Athina Katsari, Lina Stankovic, and Vladimir Stankovic. the plegma dataset: Domestic appliance-level and aggregate electricity demand with metadata from greece. *Scientific Data*, 11(1):376, 2024.

[15] Sebastian Bach, Alexander Binder, Gregoire Montavon, Frederick Klauschen, Klaus-Robert Muller, and Wojciech Samek. On pixel-wise

explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[16] Djordje Batic, Vladimir Stankovic, and Lina Stankovic. Towards transparent load disaggregation–a framework for quantitative evaluation of explainability using explainable ai. *IEEE Transactions on Consumer Electronics*, 2023.

[17] Djordje Batic, Vladimir Stankovic, and Lina Stankovic. Xnilmboost: Explainability-informed load disaggregation training enhancement using attribution priors. *Engineering Applications of Artificial Intelligence*, 2024.

[18] Djordje Batic, Giulia Tanoni, Lina Stankovic, Vladimir Stankovic, and Emanuele Principi. 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, 2023.

[19] I. S. Bayram, A. Saad, R. Sims, C. Herron, and S. Galloway. Usage analysis of public AC chargers in the UK. In *EVI: Charging Ahead (EVI 2023)*, pages 40–43. Institution of Engineering and Technology.

[20] Umang Bhatt, Adrian Weller, and José MF Moura. Evaluating and aggregating feature-based model explanations. *arXiv:2005.00631*, 2020.

[21] Verónica Bolón-Canedo, Laura Morán-Fernández, Brais Cancela, and Amparo Alonso-Betanzos. A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, 599:128096, 2024.

[22] Gail Helen Broadbent, Graciela Isabel Metternicht, and Thomas Oliver Wiedmann. Increasing electric vehicle uptake by updating public policies to shift attitudes and perceptions: Case study of new zealand. 14(10):2920.

[23] Gracia Brückmann, Fabian Willibald, and Victor Blanco. Battery electric vehicle adoption in regions without strong policies. *Transportation Research Part D: Transport and Environment*, 90:102615, 2021.

[24] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[25] Yuchao Cai, Jie Zhang, Quan Gu, and Chenlu Wang. An analytical framework for assessing equity of access to public electric vehicle charging stations: The case of shanghai. *Sustainability*, 16(14):6196, 2024.

[26] Hui Cao, Shubo Liu, Longfei Wu, Zhitao Guan, and Xiaojiang Du. Achieving differential privacy against non-intrusive load monitoring in smart grid: A fog computing approach. *Concurrency and Computation: Practice and Experience*, 31(22):e4528, 2019.

[27] Gregory J. Carlton and Selima Sultana. Electric vehicle charging station accessibility and land use clustering: A case study of the chicago region. 2:100019.

[28] Brian Caulfield, Dylan Furszyfer, Agnieszka Stefaniec, and Aoife Foley. Measuring the equity impacts of government subsidies for electric vehicles. *Energy*, 248:123588, 2022.

[29] Debapriya Chakraborty, David S. Bunch, Jae Hyun Lee, and Gil Tal. Demand drivers for charging infrastructure-charging behavior of plug-in electric vehicle commuters. 76:255–272.

[30] Debapriya Chakraborty, David S Bunch, Jae Hyun Lee, and Gil Tal. Demand drivers for charging infrastructure-charging behavior of plug-in electric vehicle commuters. *Transportation Research Part D: Transport and Environment*, 76:255–272, 2019.

[31] Prasad Chalasani et al. Concise explanations of neural networks using adversarial training. In *Int. Conf. Machine Learn.*

[32] CharePlaceScotland. . `https://chargeplacescotland.org`, 2024.

[33] Anna Charly, Nikita Jayan Thomas, Aoife Foley, and Brian Caulfield. Identifying optimal locations for community electric vehicle charging. *Sustainable Cities and Society*, 94:104573, 2023.

[34] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[35] Kunjin Chen, Yu Zhang, Qin Wang, Jun Hu, Hang Fan, and Jinliang He. Scale-and context-aware convolutional non-intrusive load monitoring. *IEEE Transactions on Power Systems*, 35(3):2362–2373, 2019.

[36] European Commission, Content Directorate-General for Communications Networks, and Technology. *Ethics guidelines for trustworthy AI*. Publications Office, 2019.

[37] Department for Transport. Road traffic statistics. `https://www.gov.uk/government/collections/road-traffic-statistics`, 2023.

[38] Department for Transport. Uk national charge point registry api documentation, 2024. Accessed: 2024-01-05.

[39] Zhili Du, Lirong Zheng, and Boqiang Lin. Influence of charging stations accessibility on charging stations utilization. 298:131374.

[40] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172*, 2019.

[41] Jonathan Fairburn, Steffen Andreas Schüle, Stefanie Dreger, Lisa Karla Hilz, and Gabriele Bolte. Social inequalities in exposure to ambient air pollution: A systematic review in the WHO european region. 16(17):3127.

[42] Zhong Fan, Parag Kulkarni, Sedat Gormus, Costas Efthymiou, Georgios Kalogridis, Mahesh Sooriyabandara, Ziming Zhu, Sangarapillai Lambotharan, and Woon Hau Chin. Smart grid communications: Overview of research challenges, solutions, and standardization activities. *IEEE Communications Surveys & Tutorials*, 15(1):21–38, 2012.

[43] Xi Fang, Satyajayant Misra, Guoliang Xue, and Dejun Yang. Smart grid—the new and improved power grid: A survey. *IEEE communications surveys & tutorials*, 14(4):944–980, 2011.

[44] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

[45] Michele Garau and Bendik Nybakk Torsæter. A methodology for optimal placement of energy hubs with electric vehicle charging stations and renewable generation. 304:132068.

[46] Clark W Gellings. *The smart grid: enabling energy efficiency and demand response*. River Publishers, 2020.

[47] Leilani H Gilpin et al. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th Int. Conf. Data Science and Advanced Anal. (DSAA)*.

[48] Hoe-Han Goh and Ricardo Vinuesa. Regulating artificial-intelligence applications to achieve the sustainable development goals. *Discover Sustainability*, 2:1–6, 2021.

[49] R. Gopinath and Mukesh Kumar. Deepedge-nilm: A case study of non-intrusive load monitoring edge device in commercial building. *Energy and Buildings*, 294:113226, 2023.

[50] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[51] Alon Harell, Stephen Makonin, and Ivan V Bajić. 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)*, pages 8335–8339. IEEE, 2019.

[52] Álvaro Hernández, Rubén Nieto, David Fuentes, and Jesús Ureña. Design of a soc architecture for the edge computing of nilm techniques. In *2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS)*, pages 1–6. IEEE, 2020.

[53] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[54] Emma Hopkins, Dimitris Potoglou, Scott Orford, and Liana Cipcigan. Can the equitable roll out of electric vehicle charging infrastructure be achieved? 182:113398.

[55] Chih-Wei Hsu and Kevin Fingerman. Public electric vehicle charger access disparities across race and income in california. *Transport Policy*, 100:59–67, 2021.

[56] Patrick Huber, Alberto Calatroni, Andreas Rumsch, and Andrew Paice. Review on Deep Neural Networks Applied to Low-Frequency NILM. *Energies*, 14(9):2390, January 2021.

[57] Jana Huchtkoetter and Andreas Reinhardt. 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, pages 270–273, 2020.

[58] Amnesty International and Access Now. *The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine-Learning Systems*. 2018. Accessed: 2024-01-05.

[59] International Energy Agency (IEA). Renewables 2021. Technical report, IEA, Paris, 2021. Licence: CC BY 4.0.

[60] Jinsheng Ji, Zhou Shu, Hongqun Li, Kai Xian Lai, Minshan Lu, Guanlin Jiang, Wensong Wang, Yuanjin Zheng, and Xudong Jiang. Edge-computing-based knowledge distillation and multitask learning for partial discharge recognition. *IEEE Transactions on Instrumentation and Measurement*, 73:1–11, 2024.

[61] Jie Jiang et al. Deep learning-based energy disaggregation and on/off detection of household appliances. *ACM Trans. on Knowledge Disc. from Data (TKDD)*, 15(3):1–21, 2021.

[62] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

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

[64] Jack Kelly and William Knottenbelt. The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes. *Scientific Data*, 2(1):1–14, 2015.

[65] Jack Kelly and William Knottenbelt. The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Scientific Data*, 2(150007), 2015.

[66] Hafiz Anwar Ullah Khan, Sara Price, Charalampos Avraam, and Yury Dvorkin. Inequitable access to ev charging infrastructure. *The Electricity Journal*, 35(3):107096, 2022.

[67] Wazir Zada Khan, Ejaz Ahmed, Saqib Hakak, Ibrar Yaqoob, and Arif Ahmed. Edge computing: A survey. *Future Generation Computer Systems*, 97:219–235, 2019.

[68] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[69] Christoph Klemenjak, Anthony Faustine, Stephen Makonin, and Wilfried Elmenreich. On metrics to assess the transferability of machine learning models in non-intrusive load monitoring. *arXiv preprint arXiv:1912.06200*, 2019.

[70] J Zico Kolter and Matthew J Johnson. Redd: A public data set for energy disaggregation research. In *Workshop on data mining applications in sustainability (SIGKDD), San Diego, CA*, volume 25, pages 59–62. Citeseer, 2011.

[71] Odysseas Krystalakos, Christoforos Nalmpantis, and Dimitris Vrakas. Sliding window approach for online energy disaggregation using artificial neural networks. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, pages 1–6, 2018.

[72] Rithwik Kukunuri, Anup Aglawe, Jainish Chauhan, Kratika Bhagtani, Rohan Patil, Sumit Walia, and Nipun Batra. Edgenilm: Towards nilm on edge devices. pages 90–99, 2020.

[73] Rajeev Ranjan Kumar and Kumar Alok. Adoption of electric vehicle: A literature review and prospects for sustainability. *Journal of Cleaner Production*, 253:119911, 2020.

[74] G Le Ray and Pierre Pinson. The ethical smart grid: Enabling a fruitful and long-lasting relationship between utilities and customers. *Energy Policy*, 140:111258, 2020.

[75] Rachel Lee and Solomon Brown. Social & locational impacts on electric vehicle ownership and charging profiles. 7:42–48.

[76] Guijun Li, Tanxiaosi Luo, and Yanqiu Song. Spatial equity analysis of urban public services for electric vehicle charging—implications of chinese cities. 76:103519.

[77] Abdolah Loni and Somayeh Asadi. Data-driven equitable placement for electric vehicle charging stations: Case study san francisco. 282:128796.

[78] Wenpeng Luan, Ruiqi Zhang, Bo Liu, Bochao Zhao, and Yixin Yu. Leveraging sequence-to-sequence learning for online non-intrusive load monitoring in edge device. *International Journal of Electrical Power & Energy Systems*, 148:108910, 2023.

[79] Ruichen Ma, Ailing Huang, Hongyang Cui, Rujie Yu, and Xiaojin Peng. Spatial heterogeneity analysis on distribution of intra-city public electric vehicle charging points based on multi-scale geographically weighted regression. 35:100725.

[80] R Machlev, A Malka, M Perl, Y Levron, and Juri Belikov. Explaining the decisions of deep learning models for load disaggregation (nilm) based on xai. In *2022 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5. IEEE, 2022.

[81] Ram Machlev et al. Measuring explainability and trustworthiness of power quality disturbances classifiers using xai—explainable artificial intelligence. *IEEE Trans. Ind. Informat.*, 18(8):5127–5137, 2021.

[82] Stephen Makonin and Fred Popowich. Nonintrusive load monitoring (nilm) performance evaluation. *Energy Eff.*, 8(4):809–814, 2015.

[83] Yuyi Mao, Xianghao Yu, Kaibin Huang, Ying-Jun Angela Zhang, and Jun Zhang. Green edge ai: A contemporary survey. *Proceedings of the IEEE*, 112(7):880–911, 2024.

[84] Simone Mari, Giovanni Bucci, Fabrizio Ciancetta, Edoardo Fiorucci, and Andrea Fioravanti. An embedded deep learning nilm system: A year-long field study in real houses. *IEEE Transactions on Instrumentation and Measurement*, 2023.

[85] Luca Massidda, Marino Marrocu, and Simone Manca. Non-intrusive load disaggregation by convolutional neural network and multilabel classification. *Applied Sciences*, 10(4):1454, 2020.

[86] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[87] Christine Milchram, Rafaela Hillerbrand, Geerten van de Kaa, Neelke Doorn, and Rolf Künneke. Energy justice and smart grid systems: evidence from the netherlands and the united kingdom. *Applied Energy*, 229:1244–1259, 2018.

[88] Glenn W Milligan and Martha C Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate behavioral research*, 21(4):441–458, 1986.

[89] Rachel Stephen Mollel, Lina Stankovic, and Vladimir Stankovic. Explainability-informed feature selection and performance prediction for nonintrusive load monitoring. *Sensors*, 23(10):4845, 2023.

[90] David Murray, Lina Stankovic, and Vladimir Stankovic. Transparent AI: explainability of deep learning based load disaggregation. In *Proc. the 8th ACM Int. Conf. Sys. Energy-Eff. Buildings, Cities, and Transp.*

[91] David Murray, Lina Stankovic, and Vladimir Stankovic. An electrical load measurements dataset of united kingdom households from a two-year longitudinal study. *Scientific Data*, 4(1):1–12, 2017.

[92] David Murray, Lina Stankovic, Vladimir Stankovic, Srdjan Lulic, and Srdjan Sladojevic. Transferability of neural network approaches for low-rate energy disaggregation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8330–8334. IEEE, 2019.

[93] OECD. *G20/OECD Principles of Corporate Governance 2023*. OECD Publishing, Paris, 2023.

[94] Leaders' Session of the AI Seoul Summit. *The Seoul Declaration for Safe, Innovative and Inclusive AI*. 2024. Accessed: 2025-01-01.

[95] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`, 2017.

[96] Peter Palensky and Dietmar Dietrich. Demand side management: Demand response, intelligent energy systems, and smart loads. *IEEE transactions on industrial informatics*, 7(3):381–388, 2011.

[97] Yungang Pan et al. Sequence-to-subsequence learning with conditional gan for power disaggregation. In *Proc. - ICASSP IEEE Int. Conf. Acoust. Speech Sig. Process.*

[98] Yael Parag and Benjamin K Sovacool. Electricity market design for the prosumer era. *Nature energy*, 1(4):1–6, 2016.

[99] Jiwoong Park, Minsik Lee, Hyung Jin Chang, Kyuewang Lee, and Jin Young Choi. Symmetric graph convolutional autoencoder for unsupervised graph representation learning. In *Proceedings of the*

IEEE/CVF international conference on computer vision, pages 6519–6528, 2019.

[100] Zhenhan Peng, Matthew Wan Hong Wang, Xiong Yang, Anthony Chen, and Chengxiang Zhuge. An analytical framework for assessing equitable access to public electric vehicle chargers. 126:103990.

[101] Sikandar Abdul Qadir, Furkan Ahmad, Abdulla Mohsin AB Al-Wahedi, Atif Iqbal, and Amjad Ali. Navigating the complex realities of electric vehicle adoption: A comprehensive study of government strategies, policies, and incentives. *Energy Strategy Reviews*, 53:101379, 2024.

[102] Hasan Rafiq, Prajowal Manandhar, Edwin Rodriguez-Ubinas, Omer Ahmed Qureshi, and Themis Palpanas. A review of current methods and challenges of advanced deep learning-based non-intrusive load monitoring (nilm) in residential context. *Energy and Buildings*, page 113890, 2024.

[103] Hasan Rafiq, Xiaohan Shi, Hengxu Zhang, Huimin Li, Manesh Kumar Ochani, and Aamer Abbas Shah. Generalizability improvement of deep learning-based non-intrusive load monitoring system using data augmentation. *IEEE Transactions on Smart Grid*, 12(4):3265–3277, 2021.

[104] David B Richardson. Electric vehicles and the electric grid: A review of modeling approaches, impacts, and renewable energy integration. *Renewable and Sustainable Energy Reviews*, 19:247–254, 2013.

[105] Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[106] Avipsa Roy and Mankin Law. Examining spatial disparities in electric vehicle charging station placements using machine learning. 83:103978.

[107] Michael Schultz, Hao Li, Zhaoyhan Wu, Daniel Wiell, Michael Auer, and Alexander Zipf. OpenStreetMap land use for Europe "Research Data", 2024.

[108] Scottish Government. Scottish index of multiple deprivation 2020. Online, 2020. Accessed: 2024-01-04.

[109] Ramprasaath R Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE Int. Conf. Comput. Vis.*

[110] Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146:102551, 2021.

[111] Donghee Shin. Why does explainability matter in news analytic systems? proposing explainable analytic journalism. *Journalism Studies*, 22(8):1047–1065, 2021.

[112] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[113] Daniel Smilkov et al. Smoothgrad: removing noise by adding noise. *arXiv:1706.03825*, 2017.

[114] Society of Motor Manufacturers and Traders. New uk ev and afv registrations, 2023. [Online; accessed 21-December-2023].

[115] Fariba Soltani Mandolakani and Patrick A. Singleton. Electric vehicle charging infrastructure deployment: A discussion of equity and justice theories and accessibility measurement. 24:101072.

[116] Benjamin K Sovacool, Paula Kivimaa, Sabine Hielscher, and Kirsten Jenkins. Vulnerability and resistance in the united kingdom's smart meter transition. *Energy Policy*, 109:767–781, 2017.

[117] Colin Studholme, Derek LG Hill, and David J Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern recognition*, 32(1):71–86, 1999.

[118] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Int. Conf. Machine Learn.*

[119] Stavros Sykiotis, Sotirios Athanasoulias, Maria Kaselimi, Anastasios Doulamis, Nikolaos Doulamis, Lina Stankovic, and Vladimir Stankovic. Performance-aware NILM model optimization for edge deployment. *IEEE Trans. Green Commun. and Netw.*, pages 1–1, 2023.

[120] Enrico Tabanelli, Davide Brunelli, and Luca Benini. A feature reduction strategy for enabling lightweight non-intrusive load monitoring on edge devices. In *2020 IEEE 29th International Symposium on Industrial Electronics (ISIE)*, pages 805–810. IEEE, 2020.

[121] Giulia Tanoni, Emanuele Principi, Luigi Mandolini, and Stefano Squartini. Weakly supervised transfer learning for multi-label appliance classification. In *Applied Intelligence and Informatics*, pages 360–375. Springer Nature Switzerland, 2022.

[122] Giulia Tanoni, Emanuele Principi, and Stefano Squartini. Multilabel appliance classification with weakly labeled data for non-intrusive load monitoring. *IEEE Transactions on Smart Grid*, 14(1):440–452, 2022.

[123] Giulia Tanoni, Lina Stankovic, Vladimir Stankovic, Stefano Squartini, and Emanuele Principi. Knowledge distillation for scalable nonintrusive load monitoring. *IEEE Transactions on Industrial Informatics*, 20(3):4710–4721, 2024.

[124] The Scottish Government. Energy statistics for scotland - q4 2023. `https://www.gov.scot/publications/energy-statistics-for-scotland-q4-2023/pages/key-points/`, 2024.

[125] Cory Thoma, Tao Cui, and Franz Franchetti. Secure multiparty computation based privacy preserving smart metering system. In *2012 North American power symposium (NAPS)*, pages 1–6. IEEE, 2012.

[126] Tamara Todic, Vladimir Stankovic, and Lina Stankovic. An active learning framework for the low-frequency non-intrusive load monitoring problem. *Applied Energy*, 341:121078, 2023.

[127] Innovation & Technology UK Department for Science. *The Bletchley Declaration by Countries Attending the AI Safety Summit*. 2023.

[128] Peter Van der Waerden, Harry Timmermans, and Marloes de Bruin-Verhoeven. Car drivers' characteristics and the maximum walking distance between parking facility and final destination. *Journal of transport and land use*, 10(1):1–11, 2017.

[129] Quintin Van Heerden, Carike Karsten, Jenny Holloway, Engela Petzer, Paul Burger, and Gerbrand Mans. Accessibility, affordability, and equity in long-term spatial planning: Perspectives from a developing country. *Transport policy*, 120:104–119, 2022.

[130] Apostolos Vavouris, Benjamin Garside, Lina Stankovic, and Vladimir Stankovic. Low-frequency non-intrusive load monitoring of electric vehicles in houses with solar generation: generalisability and transferability. *Energies*, 15(6):2200, 2022.

[131] Apostolos Vavouris, Lina Stankovic, and Vladimir Stankovic. Integration of drivers' routines into lifecycle assessment of electric vehicles. In *8th International Electric Vehicle Conference*, 2023.

[132] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[133] David Font Vivanco, René Kemp, and Ester Van Der Voet. How to deal with the rebound effect? a policy-oriented approach. *Energy policy*, 94:114–125, 2016.

[134] Qingshan Xu, Yan Liu, and Kaining Luan. Edge-based nilm system with mdmr filter-based feature selection. In *2022 IEEE 5th International Electrical and Energy Conference (CIEEC)*, pages 5015–5020. IEEE, 2022.

[135] Yao-Yuan Yang et al. A closer look at accuracy vs. robustness. *Adv. Neural Inf. Proc. Syst.*, 33:8588–8601, 2020.

[136] Zhenrui Yue, Camilo Requena Witzig, Daniel Jorde, and Hans-Arno Jacobsen. Bert4nilm: A bidirectional transformer model for non-intrusive load monitoring. In *Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring*, pages 89–93, 2020.

[137] Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. Graphsaint: Graph sampling based inductive learning method. *arXiv preprint arXiv:1907.04931*, 2019.

[138] Chaoyun Zhang et al. Sequence-to-point learning with neural networks for non-intrusive load monitoring. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 32, 2018.

[139] Xiao-Yu Zhang, Chris Watkins, and Stefanie Kuenzel. Multi-quantile recurrent neural network for feeder-level probabilistic energy disaggregation considering roof-top solar energy. *Engineering Applications of Artificial Intelligence*, 110:104707, 2022.

[140] Yu Zhang, Guoming Tang, Qianyi Huang, Yi Wang, Kui Wu, Keping Yu, and Xun Shao. Fednilm: Applying federated learning to nilm applications at the edge. *IEEE Transactions on Green Communications and Networking*, 2022.