Improving knowledge distillation for non-intrusive load monitoring through explainability guided learning

note finali coverpage

(Article begins on next page)

27 April 2024

IMPROVING KNOWLEDGE DISTILLATION FOR NON-INTRUSIVE LOAD MONITORING THROUGH EXPLAINABILITY GUIDED LEARNING

Djordje Batic[*]   *Giulia Tanoni* [†]   *Lina Stankovic*[*]   *Vladimir Stankovic*[*]   *Emanuele Principi*[†]

[*] Dept. Electronic & Electrical Engineering, University of Strathclyde, Glasgow, United Kingdom
[†] Dept. Information Engineering, Università Politecnica delle Marche, Ancona, Italy

ABSTRACT

Knowledge distillation (KD) is a machine learning technique widely used in recent years for the task of domain adaptation and complexity reduction. It relies on a Student-Teacher mechanism to transfer the knowledge of a large and complex Teacher network into a smaller Student model. Given the inherent complexity of large Deep Neural Network (DNN) models, and the need for deployment on edge devices with limited resources, complexity reduction techniques have become a hot topic in the Non-intrusive Load Monitoring (NILM) community. Recent literature in NILM has devoted increased effort to domain adaptation and architecture reduction via KD. However, the mechanism behind the transfer of knowledge from the Teacher to the Student is not clearly understood. In this work, we aim to address the aforementioned issue by placing the KD NILM approach in a framework of explainable AI (XAI). We identify the main inconsistency in the transfer of explainable knowledge, and exploit this information to propose a method for improvement of KD through explainability guided learning. We evaluate our approach on a variety of appliances and domain adaptation scenarios and demonstrate that solving inconsistencies in the transfer of explainable knowledge can lead to improvement in predictive performance.

Index Terms— Non-Intrusive Load Monitoring, Energy Disaggregation, Knowledge Distillation, Neural Networks, XAI

1. INTRODUCTION

Energy conservation plays a crucial role in providing energy efficiency in smart homes and smart buildings. Surveys, such as [1], report that energy consumption awareness can lead to a reduction of about 15% of consumer energy usage in the residential sector. Recent developments in the area of research centered around Non-intrusive Load Monitoring (NILM) have shown success in estimating the contribution of individual appliances to the total load, helping gain deeper insight in energy usage and consumption habits, and enriching energy feedback and energy saving advice as a result. In particular, Deep Neural Network (DNN) approaches have reached state-of-the-art performance in various NILM tasks across most publicly available datasets [2]. However, the inherent complexity of large DNN models requires a large amount of computational resources both during training and inference, hindering deployment of the DNN-driven NILM methods on edge devices with limited resources. To this end, in recent years, techniques for complexity reduction have been gaining considerable attention in the NILM community [3, 4, 5]. One of the most promising approaches for model compression and domain adaptation is Knowledge Distillation (KD) [6], a machine learning paradigm that relies on the transfer of knowledge from a large Teacher network to a less complex Student model that can be implemented on the edge. In other application domains,

KD has demonstrated effective results in maintaining the performance of the Teacher network, while facilitating scalability [7] and preservation of privacy [8].

Another important issue that has received considerable critical attention in the NILM community is algorithmic transparency [9, 10, 11]. Lack of interpretability brought by the inherent algorithmic complexity of DNN models has caused many to regard them as "black-box" algorithms, leading to concerns raised by the scientific community [12], as well as legislative bodies [13]. The aforementioned problem has spurred the field of explainable AI (XAI), aimed to derive methods for creation of more trustworthy deep learning systems by providing human-understandable explanations of DNN outputs. Previous studies in this area of research have sought to propose techniques for generating visual explanations that highlight the features of the input which are the most influential for the prediction of a model. A considerable volume of literature suggests that such approaches can facilitate more trustworthy machine learning systems by facilitating predictive transparency [14] and assessment of the levels of bias [15]. However, despite the apparent benefits of introducing XAI in DNN-based NILM systems, most studies in KD NILM have only focused on domain adaptation and architecture reduction [16, 17], and little is understood about the mechanism behind the transfer of knowledge from the Teacher to the Student model. Importantly, the relationship between the explanations of the Teacher model outputs and how they relate to explanations of the Student model decisions has not received any attention in the NILM community.

In this paper, we propose a methodology that establishes a link between KD and XAI approaches for NILM. A KD framework is used to train less complex networks (Students) for each appliance starting from a more complex network (Teacher) trained on a large quantity of samples from different domains. The Teacher network is a multi-label classifier used to distill the knowledge into a binary Student classifier model. By exploiting existing XAI tools, we first derive visual explanations of outputs generated by the components of the KD system, with the aim of understanding the distillation mechanism. We then use this information to identify the main type of inconsistencies w.r.t. transfer of explanation knowledge. Finally, we propose a method for improvement of predictive performance of KD NILM algorithms by guiding the distillation process towards correct transfer of explanation knowledge. We evaluate the proposed methodology using models trained for classification of five appliances on geographically distinct UK-DALE [18] and REFIT [19] datasets in two domain adaptation scenarios.

In summary, the contributions of this work are as follows: (i) We identify the main type of inconsistency in the process of transferring explanation knowledge in the KD framework for NILM (ii) We propose a technique for alleviation of explanation inconsistencies in KD NILM via a new loss function (iii) We analyse the effectiveness of

the proposed explainability guided learning in various domain adaptation scenarios

The rest of this paper is structured as follows: In Section 2 we present the methodology of our approach. Section 3 demonstrates the experimental setup, while the experimental results are presented in Section 4. Finally, we conclude our work in Section 5.

2. METHODOLOGY

2.1. Knowledge Distillation

Network compression techniques are used to reduce the architecture size and overall computational load during the training and inference process. In this work, we adopt a KD approach based on a Teacher-Student strategy, where the architecture of the Teacher network is a Convolutional Recurrent Neural Network (CRNN). The architectural reduction compared to the Teacher is achieved by reducing the number of convolutional blocks and gated recurrent units in the Student model, leading to a 6-fold reduction in the number of trainable parameters. The architectures of Teacher and Student networks considered are shown in Table 1. The Teacher network is pre-trained on a large set of aggregate smart meter load profiles and then fine-tuned on a smaller set of aggregate signals. The pre-training set is annotated with sample-by-sample labels (called *strong* labels) and window-level labels (called *weak* labels). For more information on strong and weak labels, please see [20]. The networks take as input a series of D disjointed aggregate windows with dimension L and produce as output two levels of predictions, a series of D sample-by-sample state predictions $\hat{x}_s \in R^{1 \times L}$ at the *strong* level and a series of D window predictions $\hat{w}_s \in R^{1 \times 1}$ at the *weak* level. Both levels are shown in Table 1. The pre-training loss at the Teacher network is formulated as $\mathcal{L}_{pt} = \mathcal{L}_s + \lambda \mathcal{L}_w$, with $\mathcal{L}_s$ and $\mathcal{L}_w$ being Binary Cross-Entropy (BCE) defined as in [20] for strong and weak predictions, respectively. Then, the Teacher network is fine-tuned on a set of mains, annotated only with weak labels and the same set is also used during the distillation process for the Student network training. Fine-tuning is performed by re-training the Teacher network with the loss function defined as $\mathcal{L}_{ft} = \mathcal{L}_w$. The distillation loss compares soft Teacher with soft Student predictions and weak level predictions with weak ground-truth, and it is formulated as:

$$L_{KD} = \beta \cdot \mathcal{L}_{soft}\left(\sigma\left(\frac{\hat{x}_s}{T}\right), \sigma\left(\frac{\hat{x}_t}{T}\right)\right) + (1-\beta) \cdot \theta(e) \cdot \mathcal{L}_w(\hat{w}_s, w),$$
(1)

with $\sigma(\hat{x}_s/T)$ being soft predictions of the Student and $\sigma(\hat{x}_t/T)$ soft labels from the Teacher, and σ being the sigmoid function. T is the temperature parameter used to soften Teacher predictions [6]. $\theta(e)$ is a dynamic weight that balances the magnitude of the two losses based on the formula $\theta(e) = 10^{-G(e)}$ where $G(e)$ is obtained by $G(e) = \log_{10}(\mathcal{L}_w(e)) - \log_{10}(\mathcal{L}_{soft}(e))$ and index e is the training epoch. Parameter β balances the contribution of the Teacher knowledge and the weak ground-truth.

At the end of the distillation process, Student predictions are quantized to obtain the state of the appliance, by applying a threshold selected based on the validation set.

2.2. Feature Importance Map Generation

As the need for explainability is becoming an increasingly important step for integration of AI systems, there has been a strong push towards development of practical tools that facilitate better understanding of complex, "black-box" algorithms. In order to incorporate XAI in the NILM KD framework, we devote our attention to

Model	Layer	Activation	Filters	Kernel	Units
Teacher	Convolutional Block 1	ReLu	32	5	-
	Convolutional Block 2	ReLu	64	5	-
	Convolutional Block 3	ReLu	128	5	-
	Bidirectional GRUs	-	-	-	64
	Fully Connected (*strong* level)	Sigmoid	-	-	5
	Linear Softmax Pooling	-	-	-	5
	Activation (*weak* level)	Sigmoid	-	-	-
Student	Convolutional Block 1	ReLu	32	5	-
	Bidirectional GRUs	-	-	-	32
	Fully Connected (*strong* level)	Sigmoid	-	-	5
	Linear Softmax Pooling	-	-	-	5
	Activation (*weak* level)	Sigmoid	-	-	-

Table 1: Architecture of Teacher and Student models.

GradCAM, one of the most cited explainability methods [21]. GradCAM aims solve the problem of assigning importance values to the input features of a DNN algorithm.

Given an input x to a DNN model, and a target concept c, the goal is to map the relevance of each input feature to the target concept, where the target concept can be represented as a class of interest in the case of classification tasks. GradCAM operates by computing the gradient w.r.t the final convolutional layer of a CNN network [21]. In order to generate an explanation map $h^c \in R^{W \times H}$ of width W and height H for a target concept c, the gradient of the output for the target concept y^c w.r.t the kth feature map activations A^k of the last convolutional layer is computed, i.e., $\frac{\partial y^c}{\partial A^k}$. Next, a global average pooling operation is applied over the height and width dimensions (indexed by i and j, respectively) on the computed gradients, to obtain neuron importance weights [21]:

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

The generated weights represent the importance of feature map k for the target concept c. In order to compute the explanation map h^c, weighted combination of feature map activations, followed by ReLU function, is performed [21]:

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

Note that ReLU operation ensures that only features with a positive influence on the target concept are considered.

2.3. Explainability Guided Learning

As previously stated, KD minimizes the divergence between the probability distributions of the Teacher and Student models, 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. 1 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
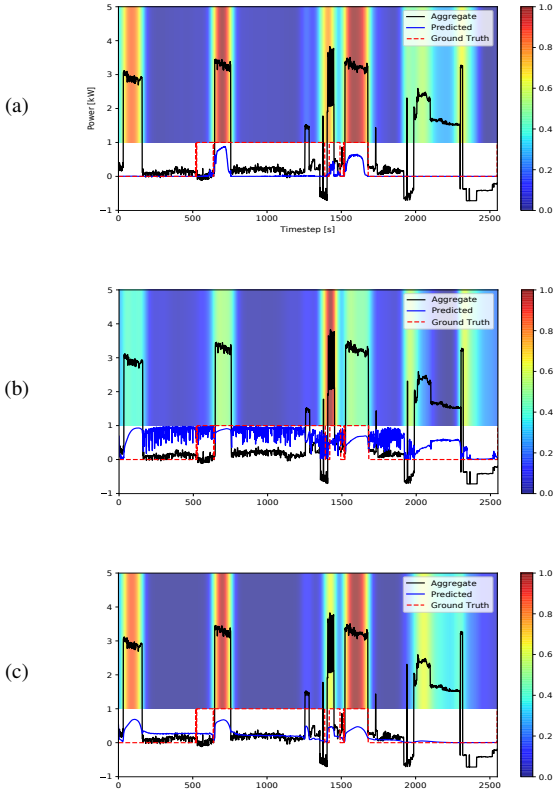
Fig. 1: Explanations for prediction of Washing Machine in the REFIT-to-REFIT domain adaptation scenario. a) Teacher explanation b) baseline Student explanation, displaying the inconsistent transfer of explanation knowledge c) Corrected Student explanation and prediction after explainability guided learning. Strong predictions are displayed before quantization.

optimization of the distillation process, leading to a more stable predictive performance.

To prevent inconsistencies in the transfer of explainable knowledge, we derive a learning technique for improvement of knowledge distillation, focusing on dissimilarities between the Teacher and Student explanations. We condition the distillation process to transfer the Teacher behaviour both in terms of output predictions and output explanations. This mode of learning, hereinafter *explainability guided learning*, is achieved through a new distillation loss function, modified to guide the learning process towards the resolution of explanation inconsistencies. As explanation heatmaps are represented in vector form, we quantify the inconsistency between two explanations through a loss function based on a measure of cosine similarity, defined as:

$$\mathcal{L}_{xai}^{\mu}(a, b) = -\frac{ab}{\|a\|\|b\|} = -\frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2}\sqrt{\sum_{i=1}^{n} (b_i)^2}}, \quad (4)$$

where a and b represent two generated explanations, while μ 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 pro-

Appliance	Scenario	γ	μ
Washing Machine	UK-DALE	0.50	*weak*
	REFIT	0.30	*strong*
Dishwasher	UK-DALE	0.85	*strong*
	REFIT	0.70	*weak*
Washer-Dryer	UK-DALE	0.60	*weak*
	REFIT	0.30	*weak*
Kettle	UK-DALE	0.30	*weak*
	REFIT	0.70	*weak*
Microwave	UK-DALE	0.70	*weak*
	REFIT	0.5	*strong*

Table 2: Training hyperparameters used for training of Student models for each of the two domain adaptation scenarios.

mote 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 can be defined as:

$$\mathcal{L}_{XGKD} = \mathcal{L}_{KD} + \gamma \cdot \mathcal{L}_{xai}^{\mu}(h_t, h_s), \quad (5)$$

where h_t and h_s represent explanations generated by Teacher and Student networks, respectively, while γ represents a parameter that adjusts the impact of the cosine similarity loss component $\mathcal{L}_{xai}^{\mu}$.

3. EXPERIMENTAL SETUP

3.1. Datasets

To validate our proposed approach, we use real-world UK-DALE [18] and REFIT [19] 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. To account for class imbalance, the datasets have been balanced as in [20]. 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 1) UK-DALE houses 1, 3, 4, and 5 (UK-DALE-to-REFIT scenario) and 2) REFIT houses 5, 6, 7, 10, 12, 13, 16, 17, 18 and 19 (REFIT-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, while the REFIT-to-REFIT scenario aims to evaluate the performance of the method when the pre-training domain is similar to the target environment signal domain. The validation set for each scenario is extracted from the pre-training set, as well as the mean and standard deviation values used to normalize the input signals.

3.2. Training Procedure

We evaluate our approach on five appliances (Washing Machine (WM), Dishwasher (DW), Washer-Dryer (WD), Kettle (KT), and Microwave (MW)), across two domain adaptation scenarios (UK-DALE-to-REFIT and REFIT-to-REFIT). Teacher is trained to perform multi-label classification of an input signal. As part of our

distillation framework, we design the Student model as a binary classifier with reduced architecture compared to the Teacher so that explainability guided learning can be focused on explanations for one appliance/class at a time. Moreover, the model can be used without re-training, even if some of the five appliances of interest are not present in the target house. We first perform knowledge distillation without explainability guided learning, using L_{KD} loss defined in Eq. (1), to create baseline Student models for each appliance in the two domain adaptation scenarios. Then, the same process is repeated with explainability guided learning with a loss function defined in Eq. (5). As each appliance model is sensitive to the choice of μ and γ, we report the chosen hyperparameters in Table 2. Hyperparameters and thresholds to quantize the predictions have been selected for each model such that they maximize the performance on the validation set. The input window dimension is $L = 2550$ which corresponds to 4h and 15min of measurements. The batch size is set to 64. Adam optimizer is used with a learning rate of 0.002, and a number of epochs is set to 1000. To prevent overfitting, we use early-stopping criterion. Standard classification metrics: Recall, Precision, and F1-score, are used for evaluation. We considered false positives as samples that have to be classified as inactive but have been predicted as active, while false negatives are samples that have been classified as active though they were inactive. Consequently, true negatives and true positives are the samples that have been correctly predicted as inactive and active, respectively.

4. EXPERIMENTAL RESULTS

We first present the results, in Table 3, for the case of domain adaptation scenario where the Teacher network is trained using UK-DALE, while the Student is trained using REFIT (UK-DALE-to-REFIT scenario). We observe that the proposed explainability guided learning led to an increase in performance compared to the baseline model for all appliances. 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. Results for the domain adaptation scenario where both Teacher and Student models were trained using REFIT data (REFIT-to-REFIT) and tested on unseen houses in REFIT are shown in Table 4. As in the first scenario, we observe improvements in the performance compared to the baseline and the Teacher, with the exception of MW, where all three methods provide similar performance. Results presented in Figure 1 suggest that explainability guided learning helps alleviate incorrect transfer of explanation knowledge, and through this process improves the predictive performance of the Student model. The results presented in Tables 3 and 4 show that the proposed explainability guided learning leads to improved knowledge distillation for most appliances in both domain adaptation scenarios. In the first scenario, we observe improvements of the F-Score measure ranging from 1.6% (for DW) up to 22.6% (for WM) compared to the baseline, while the improvements over the Teacher model ranged from 0% (for WD) up to 33.3% (for MW). Similar findings hold for the REFIT-to-REFIT domain adaptation scenario, where the maximum improvement over the baseline was 15.6% (for WD), while the maximum improvement over the Teacher was 25.5% (for DW).

Appliance	Model	Precision	Recall	F1-Score
Washing Machine	Teacher	**0.56**	0.69	0.62
	Baseline	0.70	0.43	0.53
	Ours	0.55	**0.81**	**0.65**
Dishwasher	Teacher	0.49	0.84	0.62
	Baseline	0.50	**0.88**	0.63
	Ours	**0.52**	0.83	**0.64**
Washer-Dryer	Teacher	0.79	0.77	**0.78**
	Baseline	**0.97**	0.52	0.68
	Ours	0.75	**0.81**	**0.78**
Kettle	Teacher	**0.77**	0.42	**0.55**
	Baseline	0.26	**0.98**	0.41
	Ours	0.31	0.97	0.47
Microwave	Teacher	0.43	**0.98**	0.60
	Baseline	**0.94**	0.52	0.67
	Ours	0.69	0.96	**0.80**

Table 3: Results for the UK-DALE-to-REFIT domain adaptation scenario.

Appliance	Model	Precision	Recall	F1-Score
Washing Machine	Teacher	0.57	0.91	0.70
	Baseline	0.60	**0.93**	0.73
	Ours	**0.76**	0.82	**0.79**
Dishwasher	Teacher	0.35	**0.97**	0.51
	Baseline	0.42	0.96	0.59
	Ours	**0.49**	0.93	**0.64**
Washer-Dryer	Teacher	0.93	0.52	0.67
	Baseline	**0.98**	0.47	0.64
	Ours	0.67	**0.82**	**0.74**
Kettle	Teacher	**0.92**	0.55	0.69
	Baseline	0.60	**0.95**	0.73
	Ours	0.72	0.79	**0.75**
Microwave	Teacher	0.79	**0.98**	**0.87**
	Baseline	0.77	0.95	0.85
	Ours	**0.93**	0.77	0.84

Table 4: Results for the REFIT-to-REFIT domain adaptation scenario.

5. CONCLUSIONS

We propose explainability guided learning for improvement of knowledge distillation for NILM. Driven by the increasing complexity of state-of-the-art architectures and the need for deployment on edge devices with limited resources, through this approach, we address three important challenges of NILM: scalability, algorithmic transparency and transferability. 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. Experimental results performed on actual smart meter household measurements suggest that our methodology helps mitigate explanation inconsistencies, and leads to improved predictive performance in the majority of domain adaptation scenarios.

6. ACKNOWLEDGEMENTS

7. REFERENCES

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

[2] Patrick Huber, Alberto Calatroni, Andreas Rumsch, and Andrew Paice, "Review on deep neural networks applied to low-frequency nilm," *Energies*, vol. 14, no. 9, pp. 2390, 2021.

[3] Rithwik Kukunuri, Anup Aglawe, Jainish Chauhan, Kratika Bhagtani, Rohan Patil, Sumit Walia, and Nipun Batra, "Edgenilm: Towards nilm on edge devices," in *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, New York, NY, USA, 2020, BuildSys '20, p. 90–99, Association for Computing Machinery.

[4] Jack Barber, Heriberto Cuayáhuitl, Mingjun Zhong, and Wenpeng Luan, "Lightweight non-intrusive load monitoring employing pruned sequence-to-point learning," in *Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring*, New York, NY, USA, 2020, NILM'20, p. 11–15, Association for Computing Machinery.

[5] 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.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[7] Wei Li, Shaogang Gong, and Xiatian Zhu, "Hierarchical distillation learning for scalable person search," *Pattern Recognition*, vol. 114, pp. 107862, 2021.

[8] Guyang Yu, "Data-free knowledge distillation for privacy-preserving efficient uav networks," in *2022 6th International Conference on Robotics and Automation Sciences (ICRAS)*, 2022, pp. 52–56.

[9] Maria Kaselimi, Eftychios Protopapadakis, Athanasios Voulodimos, Nikolaos Doulamis, and Anastasios Doulamis, "Towards trustworthy energy disaggregation: A review of challenges, methods, and perspectives for non-intrusive load monitoring," *Sensors*, vol. 22, no. 15, pp. 5872, 2022.

[10] R Machlev, L Heistrene, M Perl, KY Levy, J Belikov, S Mannor, and Y Levron, "Explainable artificial intelligence (xai) techniques for energy and power systems: Review, challenges and opportunities," *Energy and AI*, p. 100169, 2022.

[11] David Murray, Lina Stankovic, and Vladimir Stankovic, "Transparent ai: explainability of deep learning based load disaggregation," in *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 2021, pp. 268–271.

[12] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[13] Bryce Goodman and Seth Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[14] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.

[15] Christopher J Anders, Leander Weber, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin, "Finding and removing clever hans: Using explanation methods to debug and improve deep models," *Information Fusion*, vol. 77, pp. 261–295, 2022.

[16] Cheong-Hwan Hur, Han-Eum Lee, Young-Joo Kim, and Sang-Gil Kang, "Semi-supervised domain adaptation for multi-label classification on nonintrusive load monitoring," *Sensors*, vol. 22, no. 15, 2022.

[17] Binggang Peng, Leixin Qiu, Tao Yu, Lipeng Zhong, and Yikun Liu, "Incorporating knowledge distillation into non-intrusive load monitoring for hardware systems deployment," in *2021 IEEE 5th Conference on Energy Internet and Energy System Integration (EI2)*, 2021, pp. 3054–3058.

[18] Jack Kelly and William Knottenbelt, "The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes," *Scientific data*, vol. 2, no. 1, pp. 1–14, 2015.

[19] David Murray, Lina Stankovic, and Vladimir Stankovic, "An electrical load measurements dataset of united kingdom households from a two-year longitudinal study," *Scientific data*, vol. 4, no. 1, pp. 1–12, 2017.

[20] Giulia Tanoni, Emanuele Principi, and Stefano Squartini, "Multi-label appliance classification with weakly labeled data for non-intrusive load monitoring," *IEEE Transactions on Smart Grid*, pp. 1–1, 2022.

[21] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.