This is a peer-reviewed, author's accepted manuscript of the following conference output:
Mollel, R. S., Stankovic, L., & Stankovic, V. (Accepted/In press).
Using explainability tools to inform NILM algorithm performance: a decision tree approach. 1-5.
Paper presented at 6th International Workshop on Non-Intrusive Load Monitoring, Boston, Massachusetts, United States.

Using Explainability Tools to Inform NILM Algorithm Performance: A Decision Tree Approach

Rachel Stephen Mollel
University of Strathclyde
Glasgow, UK
rachel.mollel@strath.ac.uk

Lina Stankovic
University of Strathclyde
Glasgow, UK
lina.stankovic@strath.ac.uk

Vladimir Stankovic
University of Strathclyde
Glasgow, UK
vladimir.stankovic@strath.ac.uk

ABSTRACT

Over the years, Non-Intrusive Load Monitoring (NILM) research has focused on improving performance and more recently, generalizing over distinct datasets. However, the trustworthiness of the NILM model itself has hardly been addressed. To this end, it becomes important to provide a reasoning or explanation behind the predicted outcome for NILM models especially as machine learning models for NILM are often treated as black-box models. With this explanation, the models, not only can be improved, but also build trust for wider adoption within various applications. This paper demonstrates how some explainability tools can be used to explain the outcomes of a decision tree multi-classification approach for NILM and how model explainability informs feature selection and eventually improves performance.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

NILM, Decision Tree, Classification, Explainability

ACM Reference Format:
Rachel Stephen Mollel, Lina Stankovic, and Vladimir Stankovic. 2022. Using Explainability Tools to Inform NILM Algorithm Performance: A Decision Tree Approach. In *6th International Workshop on Non-Intrusive Load Monitoring (NILM '22), November 9–10, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3563357.3566148

1 INTRODUCTION

This paper is motivated by the potential of smart metering to provide real-time information on energy consumption in a household at any point. Load disaggregation can then be performed on smart meter data to help end users manage energy consumption and bills and help utilities implement effective demand response and tariffs. Since load disaggregation via Non-intrusive load monitoring (NILM) was initially proposed 30 years, NILM researchers have been focusing on improving disaggregation accuracy via ever more complex machine learning models [1, 5]. However, the following challenges, as summarised by [6] to enable trustworthiness of NILM, remain: generalisability across different load profiles, models enabling continuous learning and embedding user feedback, explaining NILM outcomes, fair performance evaluation and developing models that are privacy-preserving. The focus of this paper is on explaining the outcomes of a NILM model as well as improved learning through a better understanding of feature importance in the NILM multi-classifier model. The importance of explainability of a NILM model [9, 12], is: (a) to facilitate learning and satisfy curiosity as to why certain decisions have been made by the model to build trust, (b) for tuning purposes, as with explainability methods, one can learn important features that contribute significantly to the outcome and which do not, and (c) to debug the model in case of errors.

To date, there have been few attempts to explain NILM models, generally for neural networks. Heatmaps, a model agnostic way to visually interpret time series results, are demonstrated in [12] to explain NILM outputs, showing what the model considers the most impactful on its decision making. However, heatmaps may be difficult to explain to the end-user, who has little to no domain knowledge. Another attempt at explainability of NILM deep learning-based autoencoders [11], observed that the outcome of the network was improved by identifying which neurons and filters were most critical.

In this paper, the focus is on explainability of Decision Tree based NILM that acts as a multi-classifier taking as input features from smart meter measurements. Partial-Dependence (PD) and Individual conditional expectation (ICE) plots, and feature importance are explainability tools that are leveraged upon to explain the NILM multi-classifier outputs. In turn, this explanation is used to inform feature selection in order to improve the model.

Decision tree (DT) is a low-complexity supervised approach that requires only a small dataset to train the model. It has shown good performance for NILM [7, 8, 13] and can be used effectively as a multi-classifier. The tree is a hierarchical structure comprising nodes and branches that can be followed through (from the parent node to the leaf node) to understand how the outcome came to be, with Gini impurity measure used to determine the best splitting decision [14]. That is, DT method is interpretable by design [9], in the sense that it is possible to design a tree in way that decision outcomes can be mathematically explained and predicted. However, as the tree is becoming more complex with all the decision splits, the dependence of a predicted outcome on the feature is not easily seen. In other words, it is often very difficult for a human to infer how the outcome was generated. Therefore, additional explainability methods are needed to shed light on most important features that steer the model towards particular decision. Feature importance and PD plots are shown to provide user friendly global explanation

NILM '22, November 9–10, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9890-9/22/11…$15.00
https://doi.org/10.1145/3563357.3566148

[2] while ICE plots can generate intuitive local explanations for the DT model.

2 METHODOLOGY

For low-rate event-based NILM algorithms, the features are temporal and generally based on change in active power and duration [7, 8]. Let P_t be the aggregate power measurement, at sampling instant t, comprising the sum of power consumption of known individual appliances i, P_t^i, and measurement noise e_t. That is, $P_t = \sum_{i=1}^{M} P_t^i + e_t$. To generate features, automatic edge detection algorithm is used that isolates an edge if $\Delta P^i = |P_t^i - P_{t-1}^i| \geq T^i$, where T^i is an appliance-specific adaptive threshold. The event detection algorithm will output the following features: (a) *EDGE_P*: ΔP value when the appliance became ON (in Watts); (b) *EDGE_N*: ΔP value when the appliance went OFF (in Watts), (c) *DURATION*: time difference (in seconds) between time at *EDGE_P* and time at *EDGE_N*. All other ΔP edges that do not belong to the appliances of interest are labelled as "Other", representing unknown appliances contributing to e_t. Since DT requires labelled data during training, the generated features will be used as input features with output labels (appliances) during training. After training, the model is exported for prediction on unseen data without labels.

2.1 DT-based Multiclassifier

During training of the supervised DT-based multi-classifier, input features and their corresponding output (labels of appliances of interest) need to be set. The labels used in this paper correspond to the five popular appliances: Washing Machine, Dishwasher, Microwave, Toaster and Kettle. These appliances are present in most households. Of particular interest, since they are the cause of misclassifications, are appliances which have similar *EDGE_P* and *EDGE_N* such as microwave and toaster, as well as appliances with similar *DURATION* such as washing machine and dishwasher. Furthermore, these five appliances include a mixture of single state ON/OFF appliances, such as kettle, to multi-state appliances such as Washing Machine.

2.2 Explainability

In this paper, we leverage on model-agnostic interpretability tools, namely: partial dependence (PD) plots and Individual conditional expectation (ICE) plots, as well as feature importance. Feature importance gives a score for each input feature based on how useful it is for the prediction of an outcome for a given model as a whole. Feature importance is calculated as the number of samples of a feature that will reach the leaf node (predicted outcome) divided by the total samples of that particular feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain outcome. PD and ICE plots visually show the relationship between input features and predicted outcome [3]. PD plots and feature importance provide global explanation hence, do not explain individual instances of a feature and PD plots assume that the features(s) are not correlated with other features. ICE plots, on the other hand, focus on dependence of individual instances of a feature on the predicted outcome. They only display one feature at a time.

Table 1: F-SCORE Performance Comparison of the DT-based and Deep Learning Multi-classifiers.

Appliance	KE	WM	DW	MW	TOA
DT (proposed)	**1**	0.61	0.71	0.93	0.73
DT [4]	0.58	0.73	0.36	**0.95**	
Median filtering + 2-step DT [7]		0.52	0.77		
LSTM [15]	0.57	0.53	0.31	0.80	
S-CRNN [15]	0.64	0.65	0.65	0.80	
SSML-TCN [15]	0.77	0.60	0.52	0.73	
CRNN (strong labels) [15]	0.73	**0.77**	**0.78**	0.84	

For example, in Figure 3(a), the PD plot (PDP) is the highlighted thick red line, which shows, on average, the effect of feature *EDGE_N* on the predicted outcome "Kettle". The flat PDP, with low score (0.2), in Figure 3(a) indicates that feature *EDGE_N* is not important for kettle classification. The set of black lines on the x-axis indicate the distribution of all *EDGE_N* instances and the black dots are the actual individual instance values of a *EDGE_N* feature. From this figure, it is seen that, while the PD plot value is very small on average, there are a few individual instances that have a very high impact on the prediction "Kettle", as indicated by the dots that appear at the 1 "score" of y-axis on the plot, and all focused in the region between -3700W and -2000W.

3 RESULTS

In this section, a five-appliance classifier model is built using the three features described in Section 2.1. We use House 2 of the REFIT dataset [10], which contains all the appliances of interest and can be benchmarked against other NILM results that have been published in the literature. We train the classifier on 55 edge-pairs, per appliance, taken randomly from the dataset (except for the test month period) and test on the entire unseen months of October, November and December 2014. Adaptive thresholds T^i used are:(a) 2000W for Kettle, (b) 1900W for Dishwasher and Washing Machine, (c) 1000W for Microwave and (d) 700W for Toaster. For performance evaluation, the following standard classification metrics are used: Precision (PR), Recall (RE) and F-Score.

Table 1 compares F-SCORE performance of our DT multiclassifier with the following NILM multi-classifiers for Kettle (KE), Toaster (TOA), Washing Machine (WM), Microwave (MW) and Dishwasher (DW): (a) DT of [4] using *EDGE_P* and *EDGE_N* features and tested with REFIT House 2 (October 2015), (b) DT of [7] using *EDGE_P*, *EDGE_N* and active power as features, tested on REFIT House 2 (October 2014), (c) deep learning multi-classifiers, LSTM, Convolutional Recurrent Neural Networks (CRNN, S-CRNN) and Semi-Supervised Multi-Label TCN (SSML-TCN) whose results with 100% strong labels are reported in [15], trained on appliance activations and tested on unseen REFIT Houses 4, 9 and 15. Best accuracy score is indicated in bold for each appliance, showing that the proposed approach has comparable performance w.r.t other state-of-the-art multiclassifiers in the literature.

Table 2: Five Classifier Performance Metrics.

APPLIANCE	PR	RE	F-SCORE
Dishwasher	0.84	0.61	0.71
Washing Machine	0.51	0.77	0.61
Kettle	1	1	1
Microwave	0.98	0.89	0.93
Toaster	0.61	0.90	0.73

Table 3: Performance with explainability-informed feature selection.

	EDGE_N & DURATION			EDGE_N		
APPLIANCE	PR	RE	F-SCORE	PR	RE	F-SCORE
Dishwasher	**0.87**	**0.81**	**0.84**	0.71	0.60	0.65
Washing Machine	**0.69**	0.74	**0.72**	0.42	0.51	0.46
Kettle	0.98	0.99	0.98	0.94	0.95	0.94
Microwave	0.96	0.87	0.91	0.96	0.85	0.90
Toaster	0.61	0.90	0.73	0.57	0.90	0.70

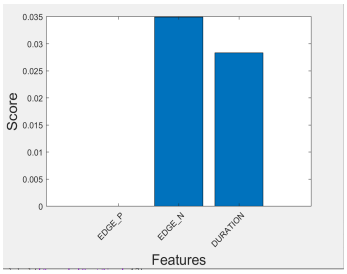


Figure 1: Feature Importance for 5-Appliance Classifier.

3.1 Explainability

Figure 1 shows the resulting feature importance scores for the proposed five-appliance classifier model, and Figures 2-4 show the ICE plots for each appliance. The y-axis of all figures shows the score of the predicted outcome (between 0 and 1) w.r.t the feature, and w.r.t instances of each feature in the ICE plots. Observing individual instances of a feature can explain the performance of each appliance. The instances that appear to have almost 1 "score" in ICE plots have very high impact on the predicted outcome.

Figure 1 shows that *EDGE_P* is the least important feature in our 5-appliance classifier model, on average across all 5 appliance classes. This is also observed in the PD plots in Figure 2, where the scores for most appliances, except kettle and microwave, are small. It is observed from Figure 2(a) that the trained model considers *EDGE_P* strongly for prediction "Kettle". Due to Kettle's high and distinct power signature, it is observed from the ICE plot of Figure 2(a) that the individual instance scores are very well clustered and rarely mixed. Indeed, all *EDGE_P* below 1800W are with score 0 and all values above 2700W are with score 1 (highlighted in green). That is why, the performance of kettle on unseen data is high as shown in Table 2. The only issue appear with a single *EDGE_P* instance of around 2400W that is mixed with the 0-score cluster of instance points between 2000W and 2600W (highlighted in yellow), where Kettle is likely to be confused as either Washing Machine or Dishwasher. This single instanced caused the Kettle's PD plot of Figure 2(a) to rise by only 60%.

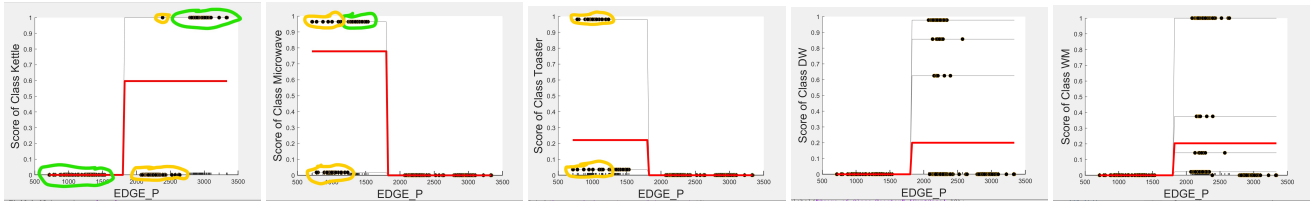
Microwave and Toaster have the lowest and similar power consumption. Therefore, they are confused with each other, as observed by multiple instances in *EDGE_P* around the same wattage in Figure 2(b-c) (multiple overlapping score 0 and 1 clusters highlighted in yellow). However, Microwave has high PD plot scores because between 1240W and 1530W all scores are 1 and the cluster is well separated (highlighted in green). With $EDGE_N$ feature in Figure 3(b-c), however, Microwave and Toaster can be separated. $EDGE_N$ instances lesser than -1180W are considered strongly for Microwave and higher that -1180W are considered strongly for Toaster (highlighted in green). The performance of the Microwave and Toaster is affected by the fact that the unseen data has $EDGE_N$ toaster values that are below -1180W and Microwave values that are above -1180W, hence false negatives for both Microwave and Toaster, as shown in Figure 5(a).

Dishwasher and Washing Machine have similar power signatures and therefore they are easily confused with each other, as observed in their ICE plots. The model leans towards Dishwasher with $EDGE_N$ instances between -2280W and -2567W, and -2215W and -2038W. The model leans towards Washing Machine with $EDGE_N$ instances between -2281W and -2216W, and -1937W. Though there is a rise of their PD plots of Figure 3(d-e) at those wattages, the PD plot scores are very low, indicating low confidence in prediction. With *DURATION* feature, however, for prediction "Dishwasher", most *DURATION* values between 570sec and 780sec have high impact. For prediction "Washing Machine", values between 270 and 540sec, and higher than 780sec, have higher impact. Even their PD plots of Figure 4(d-e) show a rise of about 50%. Hence, we conclude that including *DURATION* as a feature, can improve model performance for Washing Machine and Dishwasher as they can be distinguished from each other.

3.2 Feature Selection

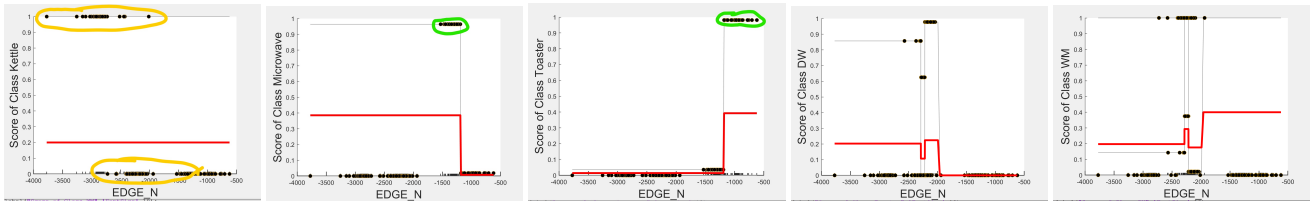
Figure 1 shows that the model considers *EDGE_N* and *DURATION* as the most important features overall. Thus, we attempt to improve the model by training with the features it considers the most important. Table 2 summarizes the classification results for the five classifier model with all 3 features, while Table 3 shows the results from the model that is trained with only two important features (discarding *EDGE_P*, as informed by explainability analysis). Improved performance, due to explainability-informed feature selection, is highlighted in bold.

As discussed previously, $EDGE_P$ is an important feature for Kettle. With $EDGE_N$ and *DURATION* as features, it is observed that there are a lot of mixed up instances seen in Kettle ICE plot of Figure 3(a) and 4(a) (highlighted in yellow). Hence, Kettle dropped its performance due to more false positives, as per Figure 5(b-c). Microwave has dropped its performance because from Figure 2(b), there are some instances between 1265W and 1532W that are not mixed up with other instances, raising the PD plot for Microwave up to about 70%. Hence, removing $EDGE_P$ affected Microwave as well. In turn, more false positives in Toaster as shown Figure 5(b-c). Since Microwave, Toaster and Kettle have all short and similar



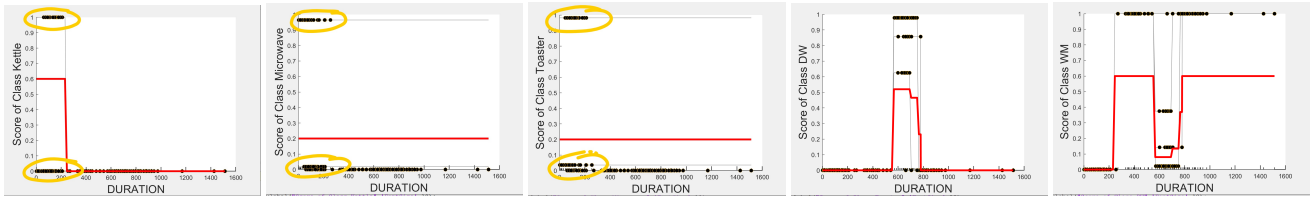
(a) Kettle (b) Microwave (c) Toaster (d) Dishwasher (e) Washing Machine

Figure 2: PD and ICE Plots for Predicted Outcome vs EDGE_P



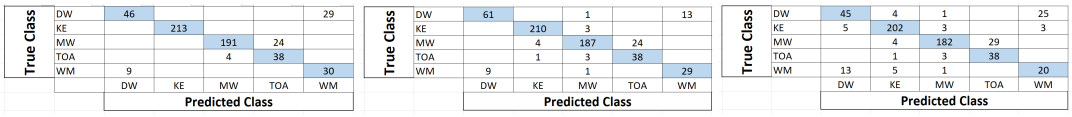
(a) Kettle (b) Microwave (c) Toaster (d) Dishwasher (e) Washing Machine

Figure 3: PD and ICE Plots for Predicted Outcome vs EDGE_N



(a) Kettle (b) Microwave (c) Toaster (d) Dishwasher (e) Washing Machine

Figure 4: PD and ICE Plots for Predicted Outcome vs DURATION



(a) Three Features (b) EDGE_N + DURATION Features (c) EDGE_N Feature

Figure 5: Prediction Confusion Matrix

duration of operation, *DURATION* as a feature brings confusion within these three appliances. In other words, Kettle, Microwave and Toaster prediction depends more on both $EDGE_P$ and $EDGE_N$ features. Therefore, with $EDGE_N$ only, the performance for these three appliances has not improved. With $EDGE_P$ feature, Dishwasher and Washing Machine are very much confused with each other as predicted by their low PD plots of Figure 2(d-e). Therefore, removing $EDGE_P$ as a feature, improves their performance as seen in Table 3. The model considers *DURATION* strongly for Washing Machine and Dishwasher, which can be seen by their drop in performance in Table 3 when *DURATION* is removed due to more false negatives as per Figure 5(c).

4 CONCLUSIONS

This paper proposes how explainability of a model yields a deeper understanding of the relative importance of features overall and on each instance of a prediction. This in turn can be used to improve the model performance, in addition to improving the trustworthiness of the model. A multiclassifier based on DT, with comparable performance to state-of-the-art classifiers, is used to demonstrate the value of explainability through evaluation using PD, ICE plots and feature importance. Although the overall model considers specific features more important than others, local explainability is critical to explain false positives. We also show that explainability-informed feature selection improves performance of the classifier in general.

5 ACKNOWLEDGMENTS

The authors acknowledge with gratitude the continuing financial support provided by the Commonwealth Scholarship Commission in conducting this research.

REFERENCES

[1] Georgios-Fotios Angelis, Christos Timplalexis, Stelios Krinidis, Dimosthenis Ioannidis, and Dimitrios Tzovaras. 2022. NILM applications: Literature review of learning approaches, recent developments and challenges. *Energy and Buildings* 261 (2022), 111951. https://doi.org/10.1016/j.enbuild.2022.111951

[2] Nadia Burkart and Marco F. Huber. 2021. A Survey on the Explainability of Supervised Machine Learning. *Journal of Artifical Intelligence Research* 70 (01 2021), 245–317. https://doi.org/10.1613/jair.1.12228

[3] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics* 24, 1 (2015), 44–65. https://doi.org/10.1080/10618600.2014.907095

[4] Kanghang He, Lina Stankovic, Jing Liao, and Vladimir Stankovic. 2018. Non-Intrusive Load Disaggregation Using Graph Signal Processing. *IEEE Transactions on Smart Grid* 9, 3 (2018), 1739–1747. https://doi.org/10.1109/TSG.2016.2598872

[5] Patrick Huber, Alberto Calatroni, Andreas Rumsch, and Andrew Paice. 2021. Review on Deep Neural Networks Applied to Low-Frequency NILM. *Energies* 14, 9 (2021). https://doi.org/10.3390/en14092390

[6] Maria Kaselimi, Eftychios Protopapadakis, Athanasios Voulodimos, Nikolaos Doulamis, and Anastasios Doulamis. 2022. Towards Trustworthy Energy Disaggregation: A Review of Challenges, Methods, and Perspectives for Non-Intrusive Load Monitoring. *Sensors* 22, 15 (2022). https://doi.org/10.3390/s22155872

[7] Mohammad Khazaei, Lina Stankovic, and Vladimir Stankovic. 2020. Evaluation of Low-Complexity Supervised and Unsupervised NILM Methods and Pre-Processing for Detection of Multistate White Goods. In *Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring* (Virtual Event, Japan) *(NILM'20)*. Association for Computing Machinery, New York, NY, USA, 34–38. https://doi.org/10.1145/3427771.3427850

[8] Jing Liao, Georgia Elafoudi, Lina Stankovic, and Vladimir Stankovic. 2014. Non-Intrusive Appliance Load Monitoring using Low-Resolution Smart Meter Data. In *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. 535–540. https://doi.org/10.1109/SmartGridComm.2014.7007702

[9] Christoph Molnar. 2022. *Interpretable Machine Learning* (2nd ed.). https://christophm.github.io/interpretable-ml-book

[10] David Murray, Lina Stankovic, and Vladimir Stankovic. 2017. An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study. Scientific data. *Scientific Data* 4, 1 (01 2017). https://doi.org/10.1038/sdata.2016.122

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

[12] David Murray, Lina Stankovic, and Vladimir Stankovic. 2021. Transparent AI: Explainability of Deep Learning Based Load Disaggregation. In *Proceedings of the 1st ACM SIGEnergy Workshop of Fair, Accountable, Transparent and Ethical AI for Smart Environments and Energy Systems* (Coimbra, Portugal) *(FATEsys '21)*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3486611.3492410

[13] M. Nguyen, Sami Alshareef, A. Gilani, and Walid G. Morci. 2015. A novel feature extraction and classification algorithm based on power components using single-point monitoring for NILM. In *2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE)*. 37–40. https://doi.org/10.1109/CCECE.2015.7129156

[14] Shan Suthaharan. 2016. *Decision Tree Learning*. Springer US, Boston, MA, 237–269. https://doi.org/10.1007/978-1-4899-7641-3_10

[15] Giulia Tanoni, Emanuele Principi, and Stefano Squartini. 2022. Multi-Label Appliance Classification with Weakly Labeled Data for Non-Intrusive Load Monitoring. *IEEE Transactions on Smart Grid* (2022), 1–1. https://doi.org/10.1109/TSG.2022.3191908