UNet-NILM: A Deep Neural Network for Multi-tasks Appliances State Detection and Power Estimation in NILM

Anthony Faustine
CeADAR, University College Dublin
sambaiga@gmail.com

Lucas Pereira
ITI, LARSyS, Técnico Lisboa, Portugal
lucas.pereira@tecnico.ulisboa.pt

Hafsa Bousbiat
University of Klagenfurt, Austria
hafsa.bousbiat@aau.at

Shridhar Kulkarni
Trinity College Dublin, Ireland
sskulki@gmail.com

ABSTRACT

Over the years, an enormous amount of research has been exploring Deep Neural Networks (DNN), particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for estimating the energy consumption of appliances from a single point source such as smart meters - Non-Intrusive Load Monitoring (NILM). However, most of the existing DNNs models for NILM use a single-task learning approach in which a neural network is trained exclusively for each appliance. This strategy is computationally expensive and ignores the fact that multiple appliances can be active simultaneously and dependencies between them. In this work, we propose UNet-NILM for multi-task appliances' state detection and power estimation, applying a multi-label learning strategy and multi-target quantile regression. The UNet-NILM is a one-dimensional CNN based on the U-Net architecture initially proposed for image segmentation. Empirical evaluation on the UK-DALE dataset suggests promising performance against traditional single-task learning.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

Non-Intrusive Load Monitoring, Energy Disaggregation, Multi-task Learning, Multi-label Classification, Deep Neural Networks, Convolutional Neural Networks, Quantile Regression

ACM Reference Format:
Anthony Faustine, Lucas Pereira, Hafsa Bousbiat, and Shridhar Kulkarni. 2020. UNet-NILM: A Deep Neural Network for Multi-tasks Appliances State Detection and Power Estimation in NILM. In *The 5th International Workshop on Non-Intrusive Load Monitoring (NILM'20), November 18, 2020, Virtual Event, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3427771.3427859

NILM'20, November 18, 2020, Virtual Event, Japan

ACM ISBN 978-1-4503-8191-8/20/11…$15.00
https://doi.org/10.1145/3427771.3427859

1 INTRODUCTION

The recent adoption of smart meters in emerging smart grids has increased the interest in Non-Intrusive Load Monitoring (NILM). NILM comprises computational methods that use aggregate power data to infer the power consumption of appliances in buildings [7].

Recently, deep neural networks (DNN) approaches to solve the NILM problem have moved to the spotlight [8, 9, 13, 15, 31]. Yet, scholars have focused mainly on single-appliance/single-task learning strategies, where one neural network is trained for one particular appliance a time. The main drawback of such strategies lies in ignoring the simultaneous activation of multiple appliances and the dependencies between their usage [18]. Additionally, most approaches focus on single tasks where one neural network is trained for either power estimation or appliance state detection. Furthermore, it should be noted that the main NILM tasks (appliance state detection and power estimation) depend on one another since the appliance stats are derived from their power profile.

Thus in this work, we propose UNet-NILM, a one-dimensional CNN based on the U-Net architecture [24] for multi-task-multi-appliances' state detection and power estimation. We leverage multi-appliances-multi-task learning strategy to improve generalization of DNN for NILM by sharing the feature space of the two dependent tasks [4, 32].

Multi-task learning is an approach to inductive transfer that improves generalization by using the domain information present in the training feature space of dependent tasks [4, 5, 11, 22, 32]. It learns several tasks in parallel while applying a shared representation. Consequently, the feature space learned for each task is used to improve the performance of other tasks. This learning paradigm has been applied successfully across several machine learning applications such as natural language processing [22], speech recognition [5], and computer vision [11].

2 PROPOSED METHOD

The proposed approach aims to identify the states $s_m(t)$ of an appliance m and estimate its power consumption $y_m(t)$ from the total power consumption x_t for $m = 1, 2, \dots M$, where M indicates the number of appliances. We refer to this as a *multi-task-multi-appliances NILM problem*, where given an observed aggregate power signal, unobserved states s_t of electrical appliances are detected and the corresponding power consumption y_t estimated. The aggregate power consumption x_t at time t can be expressed by Equation 1:

$$x(t) = \sum_{m=1}^{M} y_m(t) \cdot s_m(t) + \epsilon(t) \tag{1}$$

where ϵ_t represents the contributions from appliances not accounted for and measurement's noise [14]. Specifically, the problem is formulated as follows:

Let $\mathrm{X} \in \mathbb{R}^{T\times d} = \{x_1 \dots x_T\}$ denote a set of input features derived from the aggregate power consumption of M appliances and $\mathrm{Y} \in \mathbb{R}^{T\times M}$ indicates the power signals of the associated appliances, where each appliance has k states denoted as $\mathrm{s}_m(t) = \{s_m(t)^1, \dots s_m(t)^k\}$ such that $s_m(t)^k \in \{0, 1\}$. The matrix $\mathrm{S} \in \mathbb{R}^{T\times M}$ indicates the associated multi-label states for the M appliances and $\mathrm{Y} \in \mathbb{R}^{T\times M}$ is the corresponding power consumptions. Given $D = \{\mathrm{x}(t), \mathrm{s}(t) | t = 1, \dots, T\}$ datasets, the goal is to learn a multi-task model that predicts the state vector $\mathrm{s}(t) = \{s_m(t), \dots s_M(T)\}$ and power signal vector $\mathrm{y}(t) = \{y_m(t), \dots y_M(T)\}$ from the input aggregate power feature vector x_t.

2.1 Multi-target Quantile Regression for Power Estimation

Despite the success of DNN methods for power estimation in NILM, they are based on single-target regression, which creates a single model for each target, disregarding the possible inter-target correlation [31]. Multi-target regression, also called multi-output regression, is an alternative approach that aims to estimate the power of multiple appliances based on a shared set of feature space [1]. The joint modeling of multiple appliances allows exploiting the statistical correlation among appliances, which has been shown to be beneficial compared to modeling each appliance independently [26]. Therefore, in this work we propose multi-target quantile regression for power estimation.

Quantile regression aims to estimate either the conditional median or other quantiles of the target variable y conditioned on an observed feature x. The conditional distribution of y given $X = x$ can be expressed by a cumulative distribution $F(y|X = x) : \mathbb{R} \to [0, 1]$. For a continuous and strictly monotonically increasing cumulative distribution $F(y|X = x)$ the quantile function $q^{(\tau)}$ of the random variable y is simply the inverse of cumulative distribution: $q^{(\tau)} = F^{-1}(\tau)$ for all quantile levels τ [10, 23]. As a consequence the approximate of the entire conditional distribution of the target variable y is given as $\hat{y}^{(\tau)} = F^{-1}(\tau|\mathrm{x} = x)$ [23]. Thus a non-parametric probabilistic density estimate $\hat{f}(y)$ can be obtained by gathering a set of N_τ quantile estimates such that

$$\hat{f}(y) = \{\hat{y}^{(\tau_n)}, n = 1, \dots N_\tau \mid \tau \in (0, 1)\} \tag{2}$$

Learning is accomplished by minimizing the pinball loss defined by Equation 3, where $r = (y - \hat{y})$ denotes the residual power.

$$\rho_\tau(r) = \begin{cases} (\tau - 1) \cdot r & \text{if } r \geq 0 \\ \tau \cdot r & \text{otherwise} \end{cases} \tag{3}$$

The DNN for multi-target quantile regression can be trained by minimizing the objective function given by Equation 4:

$$\mathcal{L}_{(\rho_\tau)}(\hat{\mathrm{y}}^\tau, \mathrm{y}) = \frac{1}{TM} \sum_{t=1}^{T} \sum_{n=1}^{N_\tau} \sum_{m=1}^{M} \rho_{\tau_n}(\hat{y}_m(t)^{\tau_n} - y_m(t)) \tag{4}$$

The use of quantile regression provide uncertainty estimation almost for free. Uncertainty estimation can play an essential role in assessing how much to trust predictions produced by the DNN model [33]. Furthermore, quantifying predictive uncertainty is crucial in determining how much to trust the DNN model's prediction [33], resulting in an accurate, reliable, and interpretable machine learning system [30].

2.2 Multi-label Learning for State Estimation

Multi-label learning is a machine learning technique that aims at predicting one or more labels for each input instance [6, 31]. Several studies have demonstrated that multi-label learning represents a viable alternative to conventional NILM approaches [2, 3, 6, 17, 18, 20, 27]. For instance, the work by [2], investigated the possibility of applying a temporal multi-label classification approach in NILM where a novel set of meta-features was proposed. In [27] an extensive survey for the multi-label classification and the multi-label meta-classification framework is presented. Recent studies have explored deep neural networks for multi-label appliance recognition in NILM [19, 29], yet these approaches rely on single-task learning strategy.

A common approach that extends DNNs to multi-label classification uses one DNN to learn the joint probability of multiple label conditioned on the input features. The final predicted multi-label is obtained by applying a *sigmoid* activation function [29]. To this end, an additional threshold mechanism is required to transform the *sigmoid* probabilities to multi-label outputs.

In contrast, we propose DNN multi-label learning that uses *softmax* to implicitly capture the relations between multiple states and avoid th need for the thresholding mechanism as demonstrated in [6]. The cross-entropy between the predicted softmax distribution and the state of each appliance is thus used to learn the model parameters such that:

$$\mathcal{L}_{CE}(\hat{\mathrm{s}}, \mathrm{s}) = -\frac{1}{TM} \sum_{t=1}^{T} \sum_{m=1}^{M} s_m(t) \cdot \log \frac{\exp(\hat{s}_m(t))}{\sum_k^2 \exp(\hat{s}_m(t)^j)} \tag{5}$$

2.3 UNet-NILM for Multi-Task NILM

The UNet-NILM is a one-dimensional CNN architecture consisting of downsampling blocks, upsampling blocks, and an output layer. The downsampling and upsampling blocks use the U-Net's fundamental concepts to map the input sequence x_w to the feature map sequences z_m on a freely chosen temporal scale [24]. It achieves this by computing an increasing number of higher-level features on more granular time scales using downsampling (1D convolution) blocks. The higher-level features are combined with the earlier computed local, high-resolution features using upsampling (1D transposed convolution) blocks, producing multi-scale features to make predictions.

The output layer consists of an N-stages 1D CNN which capture non-temporal dependencies across the feature map sequences z_m and produces latent feature maps z_c that is shared among the multiple tasks. The $N-stages$ 1D CNN layers are followed by three Multilayer perceptron (MLP) layers with hidden sizes of N_h, $2 \times M$ and $N_\tau \times M$. The first MLP layer receives the latent feature maps z_c to produce a lower dimension output feature map z_o. The final

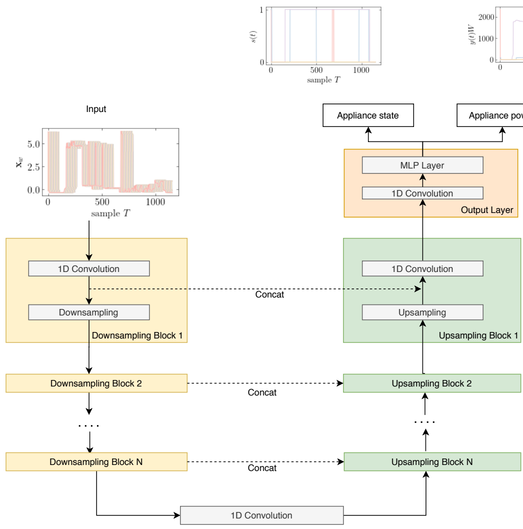

Figure 1: Architecture of UNet-NiLM for multi-task NILM.

predicted multi-label states $\hat{s}_t$ and power signal $\hat{y}_t$ is obtained as follows:

$$\hat{s}_t = \text{softmax}(\text{MLP}_{\theta_s}(z_o)) \tag{6}$$

$$\hat{y}_t = \text{MLP}_{\theta_p}(z_o) \tag{7}$$

It can be observed that the appliance state detection and power-estimation task share the feature-space z_o, while keeping task-specific output layers. The sharing of feature-space greatly reduces the risk of over-fitting while introducing an inductive bias.

The proposed approach is trained using a sliding window approach where the input to the model is a sequence of aggregate measurements $\bar{x}_L$ with window size L such that $\bar{x}_L = \{x(t-L+1), x(t-L+2), \ldots x(t)\}$. The DNN network learns to maps the sequence of observed inputs $\bar{x}_L$ to the corresponding targets' states $s(t)$ and power consumption $y(t)$. Through the experiment, we find that the window size of 100 gives optimum results and provides near real-time disaggregation. To learn the model parameters, a standard back-propagation is used to optimize the two objective functions:

$$\mathcal{L}_T = \mathcal{L}_{CE}(\hat{s}, s) + \mathcal{L}_{(\rho_\tau)}(\hat{y}^\tau, y) \tag{8}$$

The following values of τ were used throughout the experiment [2.5%, 10%, 50%, 90%, 97.5%]. Training is done for 50 iterations using the Adam optimizer with an initial learning rate of $\alpha = 0.001$, $\beta = (0.9, 0.98)$, and a batch size of 128. A factor of 0.1 reduces the learning rate once the learning stagnates for 5 consecutive iterations following *ReduceLROnPlateu* scheduler.

2.4 Appliance State Profile

We model each appliance as a binary-state machine ($k = 2$): *on*-state whenever it is consuming power and *off*-state otherwise. To generate appliance state $s_m(t)$ from the appliance power profile, a sequence to quantile sliding window (seq2quantile) approach is used in contrast to the slope algorithm approach used in previous works [17, 18]. Using seq2quantile, we first generate a sequence $y_m(t : t + w_m)$ from appliance power profile y and compute its quantile of $y_m^\tau = Q_\tau(y_m(t : t + w_m))$ where $w_m \propto \frac{T_{on}}{T_s}$ is the window size for each appliance, T_{on} is the mean ON-duration, T_s is the sampling rate and $\tau \in (0, 1)$. The $Q_{0.5}(y_m(t : t + w_m))$ quantile implies that 50% of $y_m(t : t + w_m)$ has values greater or equal to $y_m^{\tau=0.5}$. The quantile operation acts as a filter by removing noise which may add negative influence on state generation as illustrated in fig. 2. Given the appliance quantile y_m^τ and its on-power threshold p_m^{on}, the appliance is assumed to be in the *on*-state if $y_m^\tau \geq p_m^{on}$ and in the *off*-state otherwise.

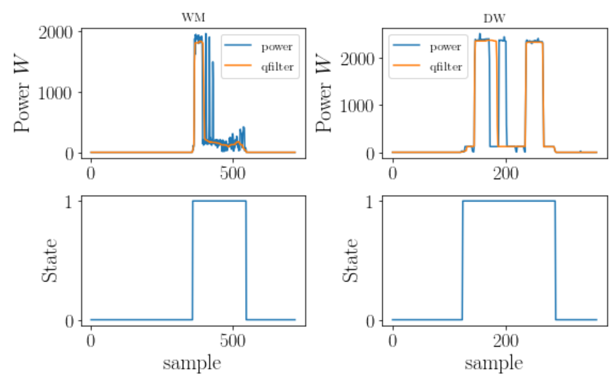

Figure 2: Example showing the process of generating appliance state profiles.

3 EVALUATION METHODOLOGY

To evaluate our method, we first implement a baseline model (1D-CNN) based on the CNN architecture presented in [6] to compare the performance of the proposed UNet-NILM. It consists of a four-stage CNN layer each with 16, 32, 64, and 128 feature maps, 2×2 strides. The first two CNN layers use a 5×5 filter size, while the last two layers use a 3×3 filter size. The four CNN layers are followed by a batch normalization layer and the PReLU activation function. The last CNN layer is followed by an adaptive average pooling layer with an output size of 16×16 and three MLP layers. The last two MLP layers are used to estimate the appliance's states and their consumption.

3.1 Dataset and Performance Metrics

The proposed method is evaluated on a UKDALE [12] dataset. The UKDALE dataset contains aggregate power consumption readings at 1 Hz resolution and sub-metered power consumption at 1/6 Hz resolution collected from five residential buildings in the UK. For this experiment, we used data from house one from January to March 2015. The following subset of appliances was selected for disaggregation; kettle (KT), fridge (FRZ), dishwasher (DW), washing machine (WM), and microwave (MW). Due to the high noise to signal ratio and some inconsistencies on the UKDALE's aggregated measurements the following data pre-processing were performed: 1) data were resampled to 6 seconds, 2) removed the missing values, and 3) constructed artificial aggregate consumption by taking the

Table 1: Experimental results in the UK-DALE dataset.

	EAC		MAE		NDE		$\text{exb} - \text{F}_1$	
Appliance	1D-CNN	UNet-NILM	1D-CNN	UNet-NILM	1D-CNN	UNet-NILM	1D-CNN	UNet-NILM
KT	0.589 ± 0.003	0.677 ± 0.017	20.390 ± 0.169	16.003 ± 0.824	0.674 ± 0.010	0.429 ± 0.039	0.944	0.956
FRZ	0.923 ± 0.000	0.937 ± 0.000	18.583 ± 0.006	15.124 ± 0.014	0.073 ± 0.000	0.072 ± 0.000	0.964	0.962
DW	0.875 ± 0.000	0.914 ± 0.000	9.884 ± 0.012	6.764 ± 0.012	0.126 ± 0.000	0.080 ± 0.000	0.913	0.909
WM	0.875 ± 0.000	0.909 ± 0.000	15.758 ± 0.009	11.506 ± 0.006	0.111 ± 0.000	0.062 ± 0.000	0.954	0.963
MW	0.630 ± 0.002	0.753 ± 0.003	9.690 ± 0.055	6.475 ± 0.072	0.656 ± 0.007	0.334 ± 0.005	0.907	0.916
Average	0.778 ± 0.003	0.838 ± 0.004	14.86 ± 0.050	11.174 ± 0.186	0.328 ± 0.003	0.195 ± 0.009	0.937	0.941

summation of selected appliances plus additional appliance (Television in this setting). We then applied quantile filter $Q_{\tau}^{w_\tau}$ with $\tau = 0.5$ and $w_\tau = 10$ presented in section 2.4 to smooth the signal. After that, z-score normalization, which normalizes the input data by using the mean and standard deviation, was applied.

We quantitatively evaluate the performance of the UNet-NILM model with both regression and classification metrics. To assess the performance of the regression task, we adopt three standard metrics used in NILM, namely the mean absolute error (MAE), estimated accuracy (EAC) and normalized disaggregation error (NDE) [21]. The MAE quantifies the error in predicted power at every time point such that $\text{MAE} = \frac{1}{TM}\sum_{i=1}^{T}\sum_{m=1}^{M}|\hat{y}_m(t) - y_m(t)|$ wheres the EAC metric gives the total estimated accuracy [9] defined as $\text{EAC} = 1 - \frac{\sum_{i=1}^{T}\sum_{m=1}^{M}|\hat{y}_m(t)-y_m(t)|}{2\sum_{m=1}^{T}\sum_{m=1}^{M} y_m(t)}$. The NDE metric measures the normalized error of the squared difference between the prediction and the ground truth defined as $\text{NDE} = \frac{\sum_{i=1}^{T}\sum_{m=1}^{M}(\hat{y}_m(t)-y_m(t))^2}{\sum_{i=1}^{T}\sum_{m=1}^{M} y_m(t)^2}$

To quantify the multi-label classification performance, we use label-based and instance-based metrics. Label-based metrics work by evaluating each label separately and returning the average (micro or macro) value across all appliances. In contrast, the instance-based metrics evaluate bi-partition over all instances. To this end, two metrics, namely example-based F_1 (F_1-eb) and macro-averaged F_1 (F_1-macro) measures are used. Example-based F_1 (F_1-eb) is an instance-based metric that measures the ratio of correctly predicted labels to the sum of the total true and predicted labels such that: $F_1\text{-eb} = \frac{\sum_{i=1}^{M} 2\cdot t_p}{\sum_{i=1}^{M} y_i + \sum_{i=1}^{M}\hat{y}_i}$ The F_1-macro is derived from F_1 score and measures the label-based F_1 score averaged over all labels and is defined as: $F_1\text{-macro} = \frac{1}{M}\sum_{i=1}^{M}\frac{2\cdot t_{pi}}{2\cdot t_{pi}+f_{pi}+f_{ni}}$ where t_p is true positive, f_p is false positive and f_n is false negative. High F_1-ma usually indicate high performance on less frequent labels [16].

4 RESULTS

Table 1 summarizes the outcomes of our experiments for both UNet-NILM and the CNN baseline. The table includes results for four metrics that were calculated for each appliance separately. The values reported consist of the predicted value of the target appliance consumption and the estimated uncertainty in the case of the regression metrics.

Concerning the multi-label classification, However, the F1-measure shows a slight improvement in detecting ON/OFF state for the UNet-NILM in all appliances included in our experiments. This result is also supported with the F_1-eb metric and F_1-macro.

The F_1-macro highlights the superiority of UNet-NILM with values of 0.941 and 0.93 for UNet-NILM and the baseline CNN, respectively.

The three regression metrics considered in our experiments also support the findings revealed by the classification metrics. According to the table, the three metrics show that UNet-NILM has better performance for all the appliances with reasonable confidence. The UNet-NILM reduces the MAE by 26.35% on average for each appliance compared to the baseline CNN. We argue that it is mainly due to the downsampling/upsampling blocks used to learn the feature maps' representation based on the sequences.

The values of the uncertainty highlight the strength of the UNet-NILM and stress the precision of the obtained results. The confidence of our model makes it a very competitive choice for several applications. Specifically, in emerging solutions for Ambient And Assisted Living approaches based on NILM [25] that aim to provide a level of security and independence for the elderly living alone.

5 CONCLUSIONS

In this paper, we propose UNet-NILM, a DNN based model, for multi-task NILM that provides both ON/OFF states of multiple running appliances and their corresponding power consumption. Experimental evaluation on the UK-DALE dataset shows good disaggregation performance for different appliances with hight confidence. Overall, the UNet-NILM can accurately capture the dependencies between the appliance's usage, efficiently detect states for multiple running appliances and estimate their corresponding power consumption. Compared to existing DNNs approaches for NILM, the advantage of the proposed approach is its ability to provide both appliance states and power consumption values and estimate the prediction's uncertainty.

At this point, we acknowledge that the employed benchmarking approaches have room for improvement. Therefore, future work should explore the model evaluation with a larger number of appliances and its comparison with more advanced NILM benchmarks such as state-of-the-art single-task DNN. Finally, it should be mentioned that the proposed method considered only quantile regression for uncertainty quantification. However, other approaches, such as Monte Carlo dropout (MCD) [28] and Batch normalization [28] have been extensively used to estimate the uncertainty of DNNx, which should be investigated in future work.

ACKNOWLEDGMENTS

Lucas Pereira received funding from FCT under grants CEECIND/01179/2017, and UIDB/50009/2020.

REFERENCES

[1] Gabriel Jonas Aguiar, Everton José Santana, Saulo Martiello Mastelini, Rafael Gomes Mantovani, and Sylvio Barbon Jr. 2019. Towards meta-learning for multi-target regression problems. *CoRR* abs/1907.11277 (2019). arXiv:1907.11277 http://arxiv.org/abs/1907.11277

[2] K. Basu, V. Debusschere, S. Bacha, U. Maulik, and S. Bondyopadhyay. 2015. Nonintrusive Load Monitoring: A Temporal Multilabel Classification Approach. *IEEE Transactions on Industrial Informatics* 11, 1 (Feb 2015), 262–270. https://doi.org/10.1109/TII.2014.2361288

[3] Bundit Buddhahai, Waranyu Wongseree, and Pattana Rakkwamsuk. 2018. A non-intrusive load monitoring system using multi-label classification approach. *Sustainable Cities and Society* 39 (2018), 621 – 630. https://doi.org/10.1016/j.scs.2018.02.002

[4] Rich Caruana. 1997. Multitask learning. *Machine Learning* 28, 1 (1997), 41–75. https://doi.org/10.1023/a:1007379606734

[5] L. Deng, G. Hinton, and B. Kingsbury. 2013. New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 8599–8603.

[6] Anthony Faustine and Lucas Pereira. 2020. Multi-Label Learning for Appliance Recognition in NILM Using Fryze-Current Decomposition and Convolutional Neural Network. *Energies* 13, 16 (Aug. 2020), 4154. https://doi.org/10.3390/en13164154

[7] A. Faustine, L. Pereira, and C. Klemenjak. 2020. Adaptive Weighted Recurrence Graphs for Appliance Recognition in Non-Intrusive Load Monitoring. *IEEE Transactions on Smart Grid* (2020), 1–1.

[8] Eduardo Gomes and Lucas Pereira. 2020. PB-NILM: Pinball Guided Deep Non-Intrusive Load Monitoring. *IEEE Access* 8 (2020), 48386–48398. https://doi.org/10.1109/ACCESS.2020.2978513

[9] A. Harell, S. Makonin, and I. V. Bajić. 2019. Wavenilm: A Causal Neural Network for Power Disaggregation from the Complex Power Signal. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8335–8339. https://doi.org/10.1109/ICASSP.2019.8682543

[10] Kostas Hatalis, A. Lamadrid, K. Scheinberg, and S. Kishore. 2019. A Novel Smoothed Loss and Penalty Function for Noncrossing Composite Quantile Estimation via Deep Neural Networks. *ArXiv* abs/1909.12122 (2019).

[11] L. Kaiser, A. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One Model To Learn Them All. *ArXiv* abs/1706.05137 (2017).

[12] Jack Kelly and William Knottenbelt. 2014. UK-DALE': A dataset recording UK Domestic Appliance-Level Electricity demand and whole-house demand. *arXiv preprint arXiv:1404.0284* March (2014), 1–13. arXiv:1404.0284 http://arxiv.org/abs/1404.0284

[13] Jack Kelly and William Knottenbelt. 2015. Neural nilm: Deep neural networks applied to energy disaggregation. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. ACM, 55–64.

[14] Christoph Klemenjak, Stephen Makonin, and Wilfried Elmenreich. 2020. Towards Comparability in Non-Intrusive Load Monitoring: On Data and Performance Evaluation. In *2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. IEEE.

[15] Odysseas Krystalakos, Christoforos Nalmpantis, and Dimitris Vrakas. 2018. Sliding window approach for online energy disaggregation using artificial neural networks. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*. ACM, 7.

[16] Jack Lanchantin, Arshdeep Sekhon, and Yanjun Qi. 2019. Neural Message Passing for Multi-Label Classification. *CoRR* abs/1904.08049 (2019). arXiv:1904.08049 http://arxiv.org/abs/1904.08049

[17] D. Li and S. Dick. 2017. A graph-based semi-supervised learning approach towards household energy disaggregation. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 1–7. https://doi.org/10.1109/FUZZ-IEEE.2017.8015650

[18] Liangda Li and Hongyuan Zha. 2016. Household Structure Analysis via Hawkes Processes for Enhancing Energy Disaggregation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*. 2553–2559.

[19] Luca Massidda, Marino Marrocu, and Simone Manca. 2020. Non-Intrusive Load Disaggregation by Convolutional Neural Network and Multilabel Classification. *Applied Sciences* 10, 4 (Feb. 2020), 1454. https://doi.org/10.3390/app10041454

[20] Christoforos Nalmpantis and Dimitris Vrakas. 2020. On time series representations for multi-label NILM. *Neural Computing and Applications* (May 2020). https://doi.org/10.1007/s00521-020-04916-5

[21] Lucas Pereira and Nuno Nunes. 2018. Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 6 (Nov. 2018), e1265. https://doi.org/10/gfb8gr

[22] Hemant Pugaliya, Karan Saxena, Shefali Garg, S. Shalini, P. Gupta, Eric Nyberg, and Teruko Mitamura. 2019. Pentagon at MEDIQA 2019: Multi-task Learning for Filtering and Re-ranking Answers using Language Inference and Question Entailment. In *BioNLP@ACL*.

[23] Yaniv Romano, Evan Patterson, and Emmanuel Candes. 2019. Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 3543–3553. http://papers.nips.cc/paper/8613-conformalized-quantile-regression.pdf

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Lecture Notes in Computer Science*. Springer International Publishing, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

[25] Antonio Ruano, Alvaro Hernandez, Jesus Ureña, Maria Ruano, and Juan Garcia. 2019. NILM techniques for intelligent home energy management and ambient assisted living: A review. *Energies* 12, 11 (2019), 2203.

[26] E. Spyromitros-Xioufis, K. Sechidis, and I. Vlahavas. 2020. Multi-target regression via output space quantization. In *2020 International Joint Conference on Neural Networks (IJCNN)*. 1–9.

[27] S. M. Tabatabaei, S. Dick, and W. Xu. 2017. Toward Non-Intrusive Load Monitoring via Multi-Label Classification. *IEEE Transactions on Smart Grid* 8, 1 (Jan 2017), 26–40. https://doi.org/10.1109/TSG.2016.2584581

[28] Mattias Teye, Hossein Azizpour, and Kevin Smith. 2018. Bayesian Uncertainty Estimation for Batch Normalized Deep Networks *(Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 4907–4916. http://proceedings.mlr.press/v80/teye18a.html

[29] Yandong Yang, Jing Zhong, Wei Li, T Aaron Gulliver, and Shufang Li. 2019. Semi-Supervised Multi-Label Deep Learning based Non-intrusive Load Monitoring in Smart Grids. *IEEE Transactions on Industrial Informatics* (2019).

[30] Shuochao Yao, Yiran Zhao, Huajie Shao, Chao Zhang, Aston Zhang, Dongxin Liu, Shengzhong Liu, Lu Su, and Tarek Abdelzaher. 2018. ApDeepSense: Deep learning uncertainty estimation without the pain for IoT applications. *Proceedings - International Conference on Distributed Computing Systems* 2018-July (2018), 334–343. https://doi.org/10.1109/ICDCS.2018.00041

[31] Chaoyun Zhang, Mingjun Zhong, Zongzuo Wang, Nigel Goddard, and Charles Sutton. 2018. Sequence-to-point learning with neural networks for non-intrusive load monitoring. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[32] Yu Zhang and Qiang Yang. 2017. A Survey on Multi-Task Learning. *CoRR* abs/1707.08114 (2017). arXiv:1707.08114 http://arxiv.org/abs/1707.08114

[33] Lingxue Zhu and Nikolay Laptev. 2017. Deep and Confident Prediction for Time Series at Uber. *IEEE International Conference on Data Mining Workshops, ICDMW* 2017-Novem (2017), 103–110. https://doi.org/10.1109/ICDMW.2017.19