Investigation of Deep Learning-based Techniques for Load Disaggregation, Low-Frequency Approach

Abdolmaged Alkhulaifi[1]
Department of Information and Computer Science
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

Abdulah J. Aljohani[2]
Center of Excellence in Intelligent Engineering Systems
Department of Electrical and Computer Engineering
King Abdulaziz University, Jeddah 21589, Saudi Arabia

***Abstract*—Unlike sub-metering, which requires individual appliances to be equipped with their own meters, non-intrusive load monitoring (NILM) use algorithms to discover appliance individual consumption from the aggregated overall energy reading. Approaches that uses low frequency sampled data are more applicable in a real world smart meters that has typical sampling capability of $\leq 1Hz$. In this paper, a systematic literature review on deep-learning-based approaches for NILM problem is conducted, aiming to analyse the four key aspects pertaining to deep learning adoption. This includes deep learning model adoption, features selection that are used to train the model, used data set and model accuracy. In our study, analyses the performance of four different deep learning approaches, namely, denoising autoencoder (DAE), recurrent long short-term memory (LSTM) , Recurrent gate recurrent unit (GRU), and sequence to point. Our experiments will be conducted using the two data sets, namely, REDD and UK-DALE. According to our analysis, the sequence to point model has achieved the best results with an average mean absolute error (MAE) of** 14.98 **watt when compared to other counterpart algorithms.**

***Keywords*—NILM; deep learning; load disaggregation; recurrent long short-term memory; gate recurrent unit**

I. Introduction

Energy disaggregation, also called non-intrusive load monitoring NILM, is the method of decomposing the aggregated energy consumption of the whole household down to individual appliance usage. The problem was firstly introduced in Hart's seminal paper in 1992 [1], and has been investigated intensively since then. NILM aims to allow the household occupants to understand the consumption of each appliance and hence take effective action towards reducing the power consumption. Reporting individual appliance consumption could lead to energy consumption reduction by more than 15% [2], [3]. Numerous NILM algorithms have been proposed, where they can be divided into two main categories according to type of data they employ, namely: low frequency and high frequency. The former approach uses data that are collected in low sampling rate typically ($< 1Hz$), while the latter relies on data that are collected at high sampling rate ($> 50Hz$). Researchers have been focusing on low frequency approaches as they can be readily applied to current smart meters [4]–[7].

One of the most widely used metric in measuring an energy disaggregation algorithm is the mean absolute error (MAE), which can be formulates as:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|\widehat{y}_t^{(i)} - y_t^{(i)}\right| \tag{1}$$

where $y_t^{(i)}$ and $\widehat{y}_t^{(i)}$ are the actual and estimated power consumption of the i^{th} appliance at t instance, respectively. The appliances that are usually picked for testing are: kettle, microwave, fridge, dish washer and washing machine. Note that, the analysis uses MAE as one of the metric measurement.

Recently, deep learning techniques have been widely used in solving the low-frequency-based NILM problem, due to their capabilities of extracting features and patterns [4]–[9]. For example, three models were proposed in [5]: first model was based on denoising autoencoder (DAE) that is aiming to reconstruct a clean target from the noisy data input. The second was based a convolutional neural network (CNN)-trained model with aim to estimate the start time, end time and mean power demand. While the third was based on the long short-term memory (LSTM) recurrent neural network (RNN) architecture. The study has concluded that the DAE, CNN,and LSTM-based RNN architectures performed adequately well achieving MAE score of 18, 14 and 70 in watts respectively, when compared to non-deep learning-based techniques counterparts of combinatorial optimization (CO) and factorial hidden Markov model (FHMM) both achieved higher error, i.e. MAE of 70 and 170 respectively [5]. Note that, all of the discussed approaches in [5] were compared using the Domestic Appliance-Level Electricity (UK-DALE) data set [10] and using active power as input features.

II. Background

In [11], a hybrid model based on both hidden markov model (HMM) and deep neural network (DNN) was proposed. It works by training HMM with two emission probabilities, one for the single load to be extracted and the other for the aggregate power signal. To elaborate a little, Gaussian distribution was used to model observations of the single load whereas observations of the aggregate signal are modeled with a DNN. Aiming to learn more features, MoWan He et al. [12] modified the RNN of [5] by adding multiple parallel convolutional layers with varying filter size to detect features from aggregated signal. This idea was borrowed from GoogleLeNet [13] model for image recognition and it's also used in natural language processing.

All approaches so far tackled NILM as a sequence to sequence, given a sequence of aggregated power try to find the sequence of the appliance disaggregated power. However, in [6] a sequence to point model was proposed, where given a sequence of aggregated power find the mid-point in the appliance disaggregated power sequence. By applying sliding window on the aggregated data, the model will cover all points in the disaggregated signal. This new approach was compared to the autoencoder approach of [5] and has achieved a significant low error of MAE= 15.47 across all appliances compared to 93.49 achieved by DAE counterpart.

Typically, the active power which is the actual power that is consumed measured in watts, was the only feature that was used in energy disaggregation in low frequency deep learning-based approaches. However, M. Valenti et al. [7] introduced the idea of using reactive power, The wasted power resulting from inductive and capacitive loads measured in volt-amperes reactive, with the active power. Two different data-sets were used, namely UK-DALE [10] and Almanac of Minutely Power data set (AMPds) [14], where the model of in [7] was able to outperform the model proposed in [5] by around 8.4% and 8.4% using UK-DALE data set and AMPds, respectively.

D. Murray et al. [4] presented a study on the transferability of neural network approaches across different data-set. The purpose of the study was to measure the scalability of neural network approaches in large scale smart meter deployment. Two architecture were proposed, a CNN architecture with 28,696,641 parameters and a gate recurrent unit (GRU) architecture with 4,861 parameters [4]. Evaluation was conducted across three data sets: REDD data set [15], UK-DALE [10], and REFIT [16], where models were trained on one data set and tested on another. Results from [4] showed that the two proposed architecture preformed well in transferability test with minimal performance drop compared to training and testing on the same data set. Although Both GRU-based network and CNN-based network showed similar performance, the GRU-based network was easier to train and less complex due to having less trainable parameters compared to CNN. In [17], C. Shin et al. explored a new direction for energy disaggregation by combining regression and classification network. By multiplying regression output with classification probability to form the final estimates, their proposed model which is employing subtask gated networks (SGN), outputs the power estimation gated with on/off classification. In their experiment in REDD and UK-DALE data-sets, they reported that SGN showed $15 and 30\%$ improved performance on average when compared to of the FHMM [18], and DAE of [5].

Against this background, we will analyze the performance of four different deep learning approaches, namely, DAE, Recurrent LSTM, Recurrent GRU, and Sequence to point, aiming to evaluate their accuracy within the NILM problem context. Our experiments will conducted using the two, well-known, data-sets [15] and UK-DALE [10]. The rest of the paper is organised as follows. The experiment design will be detailed in Section III, in which the data sets selection criteria will be explained. In Section IV the experiment performance is quantified. Finally, our conclusions will be offered in Section V.

III. Experimental Design

This section will discuss our experiment set-up in which we selected the two most widely used data set, namely REDD [15], and UK-DALE [10]. During our experiment, we will conduct a transfer-ability test on each of the following four models:

- Denoising Autoencoder: Denoising autoencoder (DAE) was introduced by J. Kelly et al. [5]. It's a sequence to sequence model that works by attempting to reconstruct a clean target from a noisy input. They showed that denoising autoencoders performed better than other architectures for sequence to sequence learning. The Keras implementation of the model (the building of layers) was taken from a reimplementation of Kelly DAE model in Keras by Taiwan Power Company[1].
- Recurrent Neural Network (LSTM): The recurrent LSTM model was also introduced by J. Kelly et al. [5]. It's a point to point model that keeps a memory of the previous entered point. The model implements LSTM layer to overcome the vanishing gradient problem where gradient information disappears over time.
- Sequence to Point:
  Sequence to point was introduced by C. Zhang et al [6]. It's a CNN model where it maps a sequence of the input power to a midpoint in the sequence of the appliance power consumption.
- Recurrent network with GRU: The idea of using gate recurrent unit (GRU) instead of LSTM in recurrent network was proposed by D. Murray et al. [4], Krystalakos et al. [9] and [19]. Since the there is different models with different implementations, we will use the LSTM model from (b) but we replace the use of LSTM with GRU to medicate the poor performance of LSTM achieved by J.Kelly experiment.

A. Data Set Selection

The data set that was selected for this experiment are REDD [15] and UK-DALE [10] data set. These two data set was the most used data set from our literature review. Due to the different sampling of the two data set, we re-sampled the data set to 1 sample per 6 seconds. For UK-DALE, we used data from house #1 and house #2 to train our models while we used data from house #5 for testing. For REDD, we used data from house #1 and house #2 to train and data from house #3 for testing. Since UK-DALE data set has data of a period of more than 4 years and REDD data set has a period of around 3 months, we only selected a small portion window frame of UK-DALE that is roughly of around 6 months and its the same time window that was used by J. Kelly et al. [5] in their experiment. The Data sets were converted to a NILMTK [18] compatible format NILMTK [2] (non-intrusive load monitoring toolkit) is a python library that simplify extracting, processing and handling data from NILM data set.

B. Appliance Selection

We want to test our model of the two type of appliances, on/off state and the multi-state appliances. Also, the selected appliance needed to exist on both the data set to be used for the transferability test. Two appliance were selected for the experiment, microwave which represents on/off state appliance and the dish washer which represents the multi-state appliance.

[1] github.com/hyl0327/neuralnilmtp

[2] github.com/nilmtk/nilmtk

We wanted to include the washing machine to our test, but due to some issues we faced during code implementation that prevented us from extracting and performing data augmentation on washing machine data from REDD data set we had to exclude it from our test.

C. Data Augmentation

Here we prepared the data according to the experiment design of J. Kelly et al. [5]. Instead of taking a portion of main data and the a portion of matching time-frame from the appliance, we follow a complex procedure where we select the data by the activation of the appliance. We select all the activation of the desired appliance in our data set that satisfy the criteria in Table I. This insures that only complete activation event of an appliance is used for the experiment. These activations are then matched with the main power data that aligns with it. Finally, a random portions of the main power data are selected with the condition that the target appliance is not active during the selected time-frame.

Synthetic data were also used in the experiment. We created synthetic data by combining the activations of multiple appliances with the target appliance to create a new input data. This procedure was suggested by J. Kelly et al. [5] paper which helped increasing the amount of the data to be used for training that is according to our activation selection criteria. Synthetic data were only used for training and it was created by combining the activation of the following appliances: kettle, washing machine, dish washer, microwave and fridge. The code for performing data augmentation and synthesising was taken from J. Kelly et al. [5] github repository[3], although we had to modify it since it does not work anymore due to the incompatibility of the python version using in the J.Kelly project with the minimum version of the dependencies it needs.

TABLE I. CRITERIA FOR SELECTING ACTIVATION, THE SAME CRITERIA USED IN J.KELLY ET AL. [5] EXPERIMENT

Appliance	Max power (watts)	On power threshold (watts)	Min. on duration (secs)	Min. off duration (Sec)
Microwave	3000	200	12	30
Dish washer	2500	10	1800	1800

D. Data Normalization

There is different approaches for normalizing energy data. In C.Zhang et al. [6] experiment, they subtracted the input and the target data by the mean and then divides them by the standard deviation. While in J. Kelly et al. [5] experiment, they only divided the input by the standard deviation of a random subset from the whole input, while for the target data they divided it by it's maximum power draw. However, they updated their project code to divides the target data by a standard deviation of a randomly selected subset from the whole target data. A common approach when normalizing data for deep learning is to just divide both the input and the target values by the maximum value in the input data. We performed a quick test using these approaches on small portion of data, and we found that by using the standard deviation of a randomly selected subset of the input and the target and then dividing them by their the computed standard deviation achieved better results. Hence we selected this approach for our experiment.

E. Training

The models were implemented using TensorFlow with Keras and were trained on our personal machine with NVIDIA GTX 970. The window length of data used are depending on the appliance and is taken from J. Kelly et al. [5] experiment. For microwave we used a window length of 288 sample (1728 seconds) while the dish washer we reduced the window length to 1024 sample (6144 seconds) from 1536 sample (9126 seconds) due to getting out of memory error in our computer. Furthermore, the window length used for training the sequence to point model were reduced to 288 samples for both appliance, this was done to overcome the out of memory error since when training the sequence to point model we need to apply a sliding window approach on the input data and map each subset to a single point. The training epochs varies from model to model. For DAE we used epoch of 30, while for RNN LSTM and GRU we reduced the number of epoch to 20 due to it taking longer to train in our hardware. For the sequence to point model, we used an epoch of only 10, this was done because it takes roughly around 40 minute per a single epoch due to the increase in the amount of data when sliding over the input.

F. Evaluation Criteria

Each approach is trained on an on/off state appliance and on a multi-state appliance to showcase it's capabilities in not only detecting the simple patterns of on/off appliances, but to also detect the complex patterns of the multi-state appliances like the dish washer. Furthermore, the transferability test will showcase models capabilities in generalizing the learned features to other unseen, new instances of the same appliance. Each model will try to estimate the active power (measured in watts) of each appliance when given the active power of the total load. In this experiment, will be using the mean absolute error (MAE) to measure the estimation accuracy of each models under the different mentioned circumstances. The formula for MAE can be defined as follows:

$$MAE = \frac{\sum_{i=1}^{n} abs(y_i - x_i)}{n}$$

The sum of the of the differences between the prediction value $(\mathrm{y_i})$ and the true value $(\mathrm{x_i})$ is then divided by the number of samples (n).

G. Validity Evaluation

In our experiment, we want to conduct a transferability test on each trained model. Due to the differences in power consumption of each house in the two data set used, this will result in different computed standard deviation that is used for normalizing the data. To overcome this issue, the input data on both data sets were normalized using the same standard deviation. The standard deviation that is used are the average of the standard deviation of two randomly selected subsets of each data set. For the target appliance data, it was normalized in regular manner. although this will results in the appliance being normalized differently in each data set, we believe that this won't hurt the results of the appliance as the normalization will not affect it's power distribution (i.e.

[3] github.com/JackKelly/neuralnilm

assuming that microwave on REDD data set consumes 1100 watt and microwave on UK-DALE consume 1500 watts, when both are normalized that will be in the range of small numbers that are almost identical to each other. After prediction when we multiple the predicted number by the standard deviation that was used for normalizing it, it should results on both 1100 and 1500 values being brought back). Our experiment results unfortunately might not be generalized, since we only used a small portion of the data sets. This also accompanied by using low and uneven epoch per model during training. This was necessary due to the limited hardware power on the conducted computer which takes a long time to train these networks and the time constraint on our experiment. Another validity concern regarding the transferability test is that all our models does not have dropout layers, which helps in reducing overfitting. The decision to not add dropout layer was that the original architecture of each model from their respective research paper did not include dropout layers. Also since we are performing low number of epochs during training, we felt that using dropout layers could have negative effects to our results.

IV. Experimental Results and Analysis

In Table II we showcase the mean absolute error of microwave and dishwasher when trained on each models in REDD data set. While Table III shows the results on UK-DALE data set. The sequence to point model achieved the highest accuracy on both data set and both appliances. This can be attributed to the deep architecture that consists of 5 Convolution layers, which adds more parameter that can capture more feature on power patterns. Although it took the sequence to point the longest to train, we only trained it for 10 epochs compared to 30 of DAE and 20 to the RNN's models and we used a shorter window length compared to DAE and to the suggested window length by C.Zhang et al. [6] to coup with our shortage in GPU memory. The DAE model outperformed both recurrent notworks on both data sets and appliances, which confirms J. Kelly et al. [5] finding that CNN models outperforms RNN model. On the other hand, the use GRU instead of LSTM did improve the performance on on/off state machine (microwave) while significantly improving the performance on the multi-state appliance (dishwasher). Looking at the results from both data sets we can notice that UK-DALE produce a greater challenge than REDD data set to our models, as UK-DALE contains more appliances per house that makes the input power signal more noisy. In Fig. 1 we visualize some example disaggregations that was performed by the four models on both data sets.

TABLE II. The appliance mean absolute error (MAE) in watts for REDD data set. Best results are shown in bold.

Appliance	DAE	RNN LSTM	Seq2Point	RNN GRU
Microwave	26.39	42.04	**13.15**	34.58
Dish washer	51.02	90.76	**9.93**	62.77

Regarding the transferability test, the results can be seen in Table IV for training on REDD and testing on UK-DALE, and Table V and the average MAE of both when trained and tested on same data set or different can be seen in Fig. 2. We can see that on UK-DALE to REDD data sets, the sequence to point model achieves better results than the others. However, on the REDD to UK-DALE data sets it fails behind the DAE and RNN LSTM. Although sequence to point model achieved better results in same data set tested and when trained on UK-DALE and tested on REDD, it sufferers the biggest increase in MAE when trained on REDD and tested on UK-DALE. We believe it was due to REDD data set containing less data (in terms of time-frame window and number of activations for trained appliance) compared to UK-DALE, this and with the low number of epoch used for training the sequence to point model might had an effect in it's performance when trained and tested on different data sets. The low number of data on the REDD data set also affected other models when trained on REDD and tested on UK-DALE (compared the other way around). The RNN LSTM model seems to be the least affected by testing on different data set than the one trained on, followed up by RNN GRU. This could be an indication on the capability's of RNN in transferring well to other data sets. Overall, the sequence to point model still has the best average MAE on both tests despite having a massive increase in error when transferred to other data sets.

TABLE III. The appliance mean absolute error (MAE) in watts for UK-DALE data set. Best results are shown in bold.

Appliance	DAE	RNN LSTM	Seq2Point	RNN GRU
Microwave	39.61	57.12	**20.21**	46.64
Dish washer	61.17	93.18	**16.61**	65.73

TABLE IV. The appliance mean absolute error (MAE) in watts when trained on REDD data set and tested on UK-DALE data set. Best results are shown in bold.

Appliance	DAE	RNN LSTM	Seq2Point	RNN GRU
Microwave	**46.19**	56.96	66.80	59.98
Dish washer	152.69	**92.94**	100.78	148.35

TABLE V. The appliance mean absolute error (MAE) in watts when trained on UK-DALE data set and tested on REDD data set. Best results are shown in bold.

Appliance	DAE	RNN LSTM	Seq2Point	RNN GRU
Microwave	49.22	42.14	**41.97**	54.39
Dish washer	87.35	138.90	**78.60**	88.26

V. Conclusion

A brief background on energy disaggregation techniques were discussed. Then a brief summary of deep learning-based approaches using low frequency data was presented. We implemented four models selected from our research literature that we mentioned in our literature review. We applied the four models on two data sets, REDD and UK-DALE which are the two most used data sets in energy disaggregation research papers. We conducted an experiment were we trained the mentioned four models twice for each of the two data sets. The models were then tested on both the same data set (testing on unseen data from another house) and on the other data set. The comparison was carried out between the four models using mean absolute error on two scenarios, training and tested on the same data set and training and testing on two different data

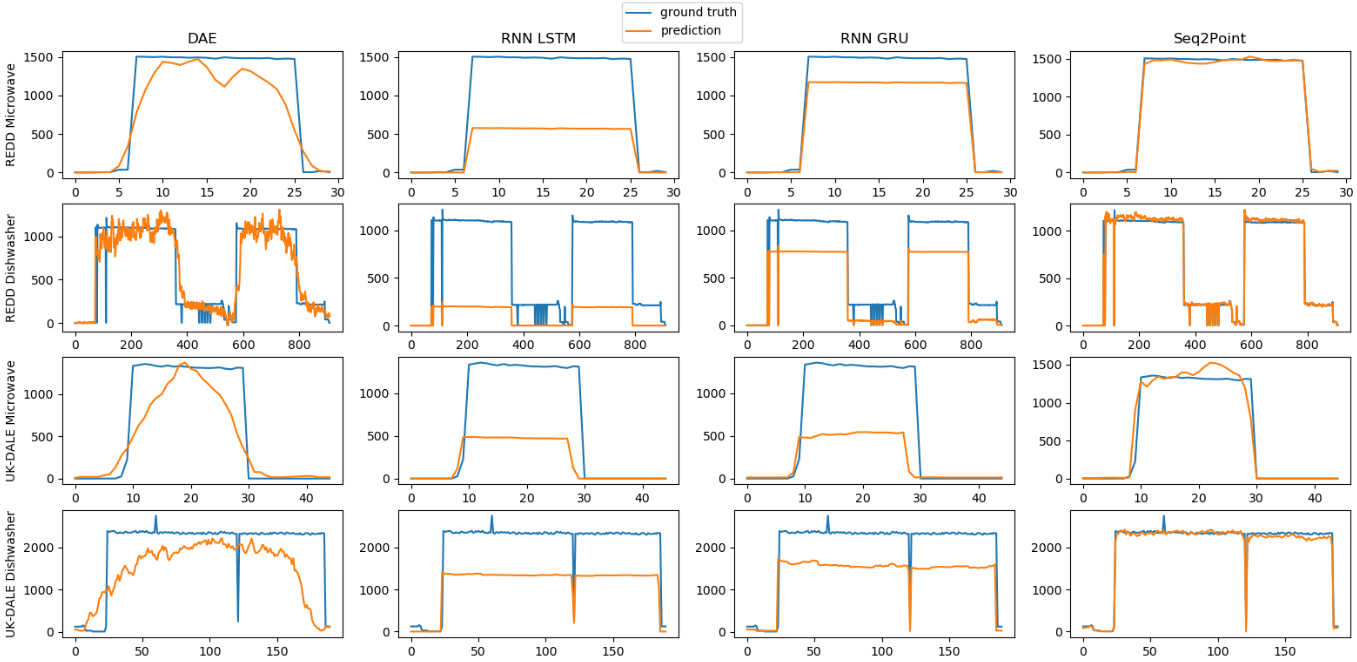

Fig. 1. Example of disaggregation results for microwave and dishwasher on both REDD and UK-DALE data set. The Y-axis corresponds to watts.

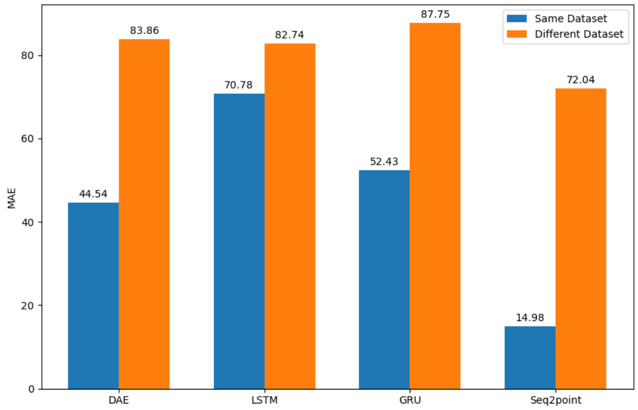

Fig. 2. The average MAE of each model when trained and tested on the same data set or on different data set.

sets. We noticed that the sequence to point model proposed by C.Zhang et al. [6] achieved the best results with an average MAE of 14.98 watts when tested on the same data set and 72.0 watts when transferred to a different data set. While on the other hand, both recurrent models performed the worst in the same data set testing and achieving close to the average score in the transferability test. Hence, the transferability is the most challenging issuing that is limiting the scalability of NILM-based solutions.

References

[1] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, Dec 1992.

[2] C. Fischer, "Feedback on household electricity consumption: a tool for saving energy?" *Energy Efficiency*, vol. 1, no. 1, p. 79–104, 2008.

[3] H. Rashid, P. Singh, V. Stankovic, and L. Stankovic, "Can non-intrusive load monitoring be used for identifying an appliance's anomalous behaviour?" *Applied Energy*, vol. 238, no. C, pp. 796–805, 2019. [Online]. Available: https://ideas.repec.org/a/eee/appene/v238y2019icp796-805.html

[4] D. Murray, L. Stankovic, V. Stankovic, S. Lulic, and S. 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)*, May 2019, pp. 8330–8334.

[5] J. Kelly and W. Knottenbelt, "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*, ser. BuildSys '15. New York, NY, USA: ACM, 2015, pp. 55–64. [Online]. Available: http://doi.acm.org/10.1145/2821650.2821672

[6] C. Zhang, M. Zhong, Z. Wang, N. H. Goddard, and C. A. Sutton, "Sequence-to-point learning with neural networks for nonintrusive load monitoring," in *AAAI*, 2016.

[7] M. Valenti, R. Bonfigli, E. Principi, and a. S. Squartini, "Exploiting the reactive power in deep neural models for non-intrusive load monitoring," in *2018 International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–8.

[8] G. Bejarano, D. Defazio, and A. Ramesh, "Deep latent generative models for energy disaggregation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 850–857, 2019.

[9] O. Krystalakos, C. Nalmpantis, and D. Vrakas, "Sliding window approach for online energy disaggregation using artificial neural networks," in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, ser. SETN '18. New York, NY, USA: ACM, 2018, pp. 7:1–7:6. [Online]. Available: http://doi.acm.org/10.1145/3200947.3201011

[10] J. Kelly and W. Knottenbelt, "The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes," *Scientific data*, vol. 2, p. 150007, 2015.

[11] L. Mauch and B. Yang, "A novel dnn-hmm-based approach for extracting single loads from aggregate power signals," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 2384–2388.

[12] W. He and Y. Chai, “An empirical study on energy disaggregation via deep learning,” in *2016 2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE 2016)*. Atlantis Press, 2016/11. [Online]. Available: https://doi.org/10.2991/aiie-16.2016.77

[13] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.

[14] S. Makonin, F. Popowich, L. Bartram, B. Gill, and I. V. Bajić, “Ampds: A public dataset for load disaggregation and eco-feedback research,” in *2013 IEEE Electrical Power Energy Conference*, Aug 2013, pp. 1–6.

[15] J. Z. Kolter and M. J. Johnson, “Redd: A public data set for energy disaggregation research,” in *Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA*, vol. 25, no. Citeseer, 2011, pp. 59–62.

[16] D. Murray, L. Stankovic, and V. Stankovic, “An electrical load measurements dataset of united kingdom households from a two-year longitudinal study,” *Scientific Data*, vol. 4, no. 1, May 2017.

[17] C. Shin, S. Joo, J. Yim, H. Lee, T. Moon, and W. Rhee, “Subtask gated networks for non-intrusive load monitoring,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, p. 1150–1157, 2019.

[18] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava, “Nilmtk: An open source toolkit for non-intrusive load monitoring,” 07 2014.

[19] P. Xiao and S. Cheng, “Neural network for nilm based on operational state change classification,” *ArXiv*, vol. abs/1902.02675, 2019.