A study of Sequence-to-point Learning for Non-Intrusive Load Monitoring

Filippos Andreadis (s4942906)

March 18, 2024

Abstract

Non-intrusive load monitoring (NILM) (also referred to as load disaggregation) is the procedure of detecting the activity of domestic appliances by analysing the total power consumption of a household. Such technology could potentially accelerate the renewable energy transition and also raise awareness to the public about energy waste. Load disaggregation is an inherently unidentifiable problem and thus, poses a demanding challenge to solve. While a variety of approaches have been proposed, deep neural networks have shown promising results in recent years. In this study the sequence-to-point learning technique, an approach that maps sliding windows of the total consumption sequence to single points of the target appliance reading, is explored in two distinct directions. Firstly, we examine how the disaggregation performance of this methodology is affected by changing the sampling frequency of the input data and secondly we evaluate the transferability of learned appliance features between two different domains. The results suggest that maintaining the original resolution of the datasets is typically the best choice, while cross-domain disaggregation is feasible for certain devices, such as the refrigerator or the microwave.

1 Introduction

The global demand for energy is constantly being increased in terms of both quantity and distribution complexity. This renders the traditional power grid system an unsustainable electricity distribution model, especially when it comes to achieving the pledge to bring the related carbon dioxide emissions to net zero by 2050 [1]. Studies show that the annual growth of energy needs stands at approximately 3.4% per year in the last decade [2], while residential and commercial buildings account for roughly 36% of the total electrical demand in the USA and 25% in the EU [3, 4]. At the same time, the recent advancements in Artificial Intelligence, Internet of Things and monitoring equipment along with the ability to store and process large amounts of data have kick-started a transition towards a redefined archetype of power grids, namely the smart grid. The development of smart grids aims to optimize the generation, distribution and consumption of energy by integrating multiple technologies such as advanced metering infrastructure, renewable and efficient energy resources and smart distribution systems.

One of the key factors of smart grid research is the accurate monitoring of appliance activity. Knowing the exact load of individual devices within a household, as well as the aggregate electrical energy consumption of the given dwelling, contributes to a better energy management in a major way [5, 6]. Non-intrusive load monitoring (NILM) refers to the process of inferring the operating times and energy consumption of individual house appliances by analysing the total consumption of a household [7]. It can enable residents to efficiently monitor and have better overall control over their energy consumption, without the extra cost of installing individual tracking devices (e.g. smart plugs) on each appliance. Additionally, having appliance specific information can help utility companies provide their customers with personalised recommendations about their energy usage or motivation to become more energy efficient through a reward system. For example, providers can inform customers about unusual patterns or which time during the day it is more economical to use their highest consuming devices.

NILM or load disaggregation has been mainly used for energy consumption reduction for both residential and industrial buildings [8, 9], but other applications include the optimization of smart grid management [10] and even human activity monitoring for assisted living [11]. It is inherently a difficult prediction problem as creating a reliable dataset of energy readings is a very hard task. Specifically, measuring devices need to be placed in every single household appliance and be made sure that they will accurately collect data, without any significant downtime and for a sufficient amount of time. Typically, this is not a fail-proof

process and thus real world data include noise and can even lack information about the true power consumption of each individual appliance. Moreover, the simultaneous switching of multiple devices, the consumption discrepancies among different households as well as the major imbalance among the energy loads of each appliance add more layers of complexity to the problem. This is why NILM has become an active area of research, with solutions ranging from traditional approaches such as Hidden Markov Models (HMM) to more recent state-of-the-art Deep Learning (DL) models.

One crucial aspect of NILM that has not been sufficiently explored yet is how the sampling rate of data can impact the disaggregation quality of a model. Public datasets can vary in temporal resolution, while proprietary load monitoring setups can also output signals in different frequencies. At the same time data storage space and computation time play an important role in the efficiency of a disaggregation system. The authors of [12] have investigated how altering the sampling rate affects the event detection accuracy of the the chi square method. Similarly, the same group presents some interesting findings after testing three state-of-the-art DL techniques on different frequencies in [13]. They conclude that favourable low-frequency sampling rates range between 1 Hz and 1/30 Hz and that lower sampling rates may not always have a deteriorating effect on disaggregation results, but can even improve performance in certain occasions.

Another interesting NILM direction is cross-domain transfer learning. By this we refer to the process of extracting features of a source domain and use them to detect appliances that belong to a new and previously unseen target domain. For example, one could train a disaggregation model over data that belong to a UK household and use it to predict the load of appliances in the US. This could benefit a NILM system by circumventing the expensive procedure of installing tracking devices in houses of a new region. Additionally, by utilizing a transfer learning scheme, a considerable amount of computational time could be avoided as pre-trained models could be readily deployed on appliances of the target domain.

The goal of this research is to implement an impactful Convolutional Neural Network (CNN) approach, defined as sequence-to-point learning in [14], and explore how its disaggregation quality is affected by changes in the sampling rate of input data. Moreover, we evaluate the cross-domain knowledge transferability of this methodology by training the model on appliances of a source domain and testing it on the respective appliances of a target domain. For this, two datasets that belong to two different regions (i.e., UK and US) have been used, where experiments have been conducted for both ways. The results suggest that keeping the original sampling frequency of the data yields the best performance, while downsampling is a viable option when a cut in resources is required. The cross-domain experiments have shown potential for a deployable model, however not for every appliance tested.

First, a survey of existing work on NILM approaches is provided in 1.1. Then, Section 2 includes preliminary information about the NILM problem and the disaggregation techniques explored in this work. Section 3 presents in detail the datasets used and the preprocessing applied to them, along with all the steps taken to perform the experiments. In Section 4 we present all our findings, which are analysed in 5. Lastly, we summarize all the insights gained during this research and discuss possible future extensions in Section 6.

1.1 Related Work

Earlier NILM work has focused on HMMs, which are typically used for probabilistic modelling of time series data. Typically, a number of states is defined for each appliance, with each state having its own probabilistic distribution. While various variants have been proposed, the most popular one has proven to be the Factorial HMM (FHMM) that generalizes the HMM state representation by letting the state be represented by a collection of state variables. The authors in [15] have proposed a promising FHMM approach for the NILM task, which is considered a fully unsupervised method for energy disaggregation. The main issue with HMM-based techniques is that they tend to become inefficient when the number of disaggregated appliances increases as they suffer from high computational complexity. A different approach lies in optimization methods where the main idea is finding the optimal combination of individual appliances that compose the aggregate signal. Genetic algorithms (GAs) have shown promising performance [16], while a Graphical Signal Processing (GSP) method, proposed in [17], offers a solution with reduced computational complexity. Traditional machine learning approaches have also been considered as solutions to NILM. Implementations of Support Vector Machines (SVM) [18], Naive Bayes classifiers [19], K nearest neighbors (K-nn) algorithms [20] and tree-based methods [21, 22] have been explored. The advantage of such models is that they are easier to implement, while also they have lower computational complexity compared to deep neural networks and in certain instances they can yield adequate results. However, in most cases a handcrafted feature extraction process is required, something that implies domain expert knowledge.

This is where the value of neural networks becomes recognizable, as they have the advantage of automatically extracting features from data. Even in their simplest form, namely feed-forward neural networks (FFNN), deep learning models can be applied to the NILM task as shown in [23], although more advanced and recent architectures have proven to be state-of-the-art. A. Harell et al. have adapted WaveNet [24] to the NILM problem in [25]. Specifically, WaveNILM consists of 1D dilated causal convolutions whose output is fed both into a gating mechanism and a rectified linear activation. The output of the two activations is multiplied and represents the output of the block. Recurrent neural networks (RNNs) along with their variations, Gated Recurrent Units (GRU) and Long Short Term Memory (LSTM) networks, have been extensively used in NILM literature [26]. Additionally, a Variational Autoencoder (VAE) was trained in [27] to extract individual appliance loads from the total consumption signal achieving impressive results. Finally, S. Sykiotis et al. proposed ELECTRIcity in [28], a trans-

former based approach which introduces an efficient training routine that is split into unsupervised pre-training and downstream task fine-tuning. This proposed routine has yielded performance increases in both predictive accuracy and training time decrease, compared to other state-of-the-art methodologies.

2 Background

This section formulates the NILM task and provides details about the baseline model used for benchmarking and the sequence-to-point architecture.

2.1 NILM Problem Formulation

The goal of load disaggregation is to accurately estimate the power consumption of individual appliances from the aggregate consumption signal of the given household. Let M be the number of household appliances and i be the index referring to the $i-th$ appliance, where $i = 1, ..., M$. The aggregate power consumption y at a given time t is the sum of the power consumption x of the individual appliances M, denoted by $x_i \ \forall \ i = 1, ..., M$. The total power consumption y at a given time t is:

$$y(t) = \sum_{i=1}^{M} x_i(t) + \epsilon(t), \tag{1}$$

where ϵ is a noise term. Hence, NILM is categorised as a blind-source separation problem, typically with a very large number of degrees of freedom. The non-linearity of the consumption patterns and the deep temporal dependencies among the appliance activations comprise a few of the severe challenges of building a disaggregation model with promising generalization capability.

2.2 Baseline model

A baseline model is used as reference point for the performance of the main DL approach that is being examined. In particular, the Combinatorial Optimization (CO) [7] model is chosen as it is frequently used as a benchmark in NILM literature. It is a simple model, easy to implement and provides fast results in terms of computation time.

This model addresses the load disaggregation task as an optimization problem where the goal is to find the combination of appliance states (i.e., each state corresponds to a certain consumption value) that best explain the observed total power consumption. The states can be calculated either automatically by clustering or manually by visual inspection of the load pattern of each appliance. Specifically, each appliance i, has a finite number of possible states C_i, each with the corresponding consumption ${B_i}^{(1)}, \ldots, {B_i}^{(C_i)}$. The allocated consumption of device i at timestep t, is given by (2), where ${\theta_i}^{(j)}(t)$ are binary variables representing if state j of appliance i is active on timestep t.

$$\hat{x_i}(t, \vec{\theta_i}) = [{B_i}^{(1)} \ldots {B_i}^{(C_i)}] \begin{bmatrix} {\theta_i}^{(1)}(t) \\ \vdots \\ {\theta_i}^{(C_i)}(t) \end{bmatrix} + e_i(t), \tag{2}$$

where $e_i(t)$ represents the intrinsic modeling error accounting for the fact that the states are just approximations.

The CO model solves the following optimization problem.

$$\min_{{\theta_i}^{(1)}(t), \ldots, {\theta_i}^{(C_i)}(t)} |y(t) - \sum_{i=1}^{M} \hat{x_i}(t, \vec{\theta_i})|$$

$$s.t. \qquad \sum_{j=1}^{C_i} {\theta_i}^{(j)}(t) = 1, i = 1, \ldots, M,$$

$$t = 1, \ldots, T$$

Here, $y(t)$ refers to the aggregated consumption at timestep t, and $\hat{x_i}(t, \theta_i)$ to the allocated load of device i, at timestep t. The constraints make sure that each appliance is operating at exactly one state at each timestep.

2.3 Sequence-to-point learning

In the context of deep learning applied to NILM, a neural network must learn a nonlinear regression between a sequence of the total power consumption and the sequence of the consumption of an individual appliance. An important condition is that the two sequences must have the same sampling rate and be aligned in time. While multiple ways of achieving this exist, a popular approach proposed by Zhang et al. in [14] is termed sequence-to-point learning. The idea is that the network learns to represent the midpoint of the appliance sequence given a window of the total consumption, also referred to as mains, as input. Specifically, the input of the network consists of sliding windows of the mains power $Y_{t:t+W-1}$ and the output is the midpoint element x_τ of the corresponding window of the target appliance. W represents a predefined window size and $\tau = t + \lfloor W/2 \rfloor$. Thus, it is assumed that the midpoint element x_τ is a nonlinear function of the total consumption window and it is also expected that the value of x_τ is connected to the information of the mains before and after it.

The sequence-to-point approach does not perform sequence to sequence mapping, but rather defines a neural network f that maps the sliding windows $Y_{t:t+W-1}$ of the input to the midpoint x_τ of the corresponding windows $X_{t:t+W-1}$ of the output. Hence, the network is modelled as $x_\tau = f(Y_{t:t+W-1}) + \epsilon$, where ϵ is a noise term. The loss function used to train such a model can be formulated as

$$L = \sum_{t=1}^{T-W+1} \log p(x_\tau | Y_{t:t+W-1}, \vec{w}), \tag{3}$$

where $\vec{w}$ are the model parameters. Finally, in order to receive predictions for the entire input sequence $Y = (y_1, \ldots, y_T)$, Y is padded with $\lceil W/2 \rceil$ zeros on both sides.

3 Methodology

In this section we provide information about the two datasets that are used, along with the architecture of the CNN model at hand. Additionally, we describe all the necessary preprocessing steps taken, the performance metrics used for evaluating the model and finally how the experiments have been conducted.

3.1 Data

A variety of both domestic and industrial open datasets exist that have been specifically created for the NILM task. In this research we choose to focus on the REDD and UK-DALE datasets. This is done because they are two of the most prevalent datasets in the literature, meaning there is a plethora of different methodologies and results to compare. Additionally, since the data have been collected from different regions (i.e., US and UK respectively) it would allow us to observe how well knowledge can be transferred between the two domains.

3.1.1 REDD

The Reference Energy Disaggregation Dataset (REDD), published by Massachusetts Institute of Technology (MIT) in 2011 [29], consists of recordings from six households in the United States with an approximate duration of 20 days per house. The version of the dataset used for this work includes a 1 Hz reading of the total power consumption of each residence along with 1/3 Hz readings of each individual appliance load. The number of monitored appliances for each house ranges between 8 and 14, while existing appliances do not entirely overlap in every house. This means that not all houses can be used to perform load disaggregation for every device present in the dataset and thus the experiments have been adjusted accordingly. Figure 1 presents the distribution of the load of the individual appliances of the first household in the dataset.

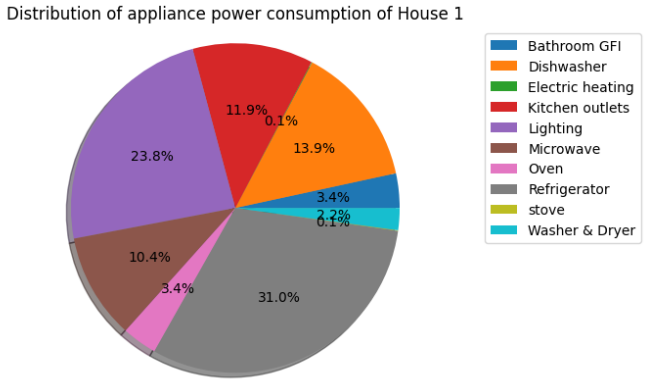


Figure 1: Distribution of the power consumption of the individual appliances in house 1 of the REDD dataset.

It is worth mentioning that the collected data account for roughly 47% of the total consumption of house 1, while this ratio ranges between 36% and 60% for the rest of the residences. Since the data corresponding to each appliance are highly inconsistent among the six houses, four specific appliances are chosen to be investigated. Specifically, four distinct models are trained for the refrigerator, the dishwasher, the microwave and the lighting using data from houses 2 to 6, and tested on house 1. Additionally, the models are validated on the last 10% of the training data. Table 1 shows details about the train-test split for each appliance.

Table 1: Train-test split for REDD.

Appliance	Training	Testing
Refrigerator	House 2,5,6	House 1
Dishwasher	House 4,5,6	House 1
Microwave	House 2,3,5	House 1
Lighting	House 2,3,5,6	House 1

3.1.2 UK-DALE

The second dataset used in this research is the UK Domestic Appliance-Level Electricity (UK-DALE) dataset, published by J. Kelly and W. Knottenbelt in [30]. It includes recordings of every six seconds of the total electricity consumption and individual appliances for five separate households from November 2012 to January 2015. A total of approximately 50 different devices have been monitored, however only the refrigerator, dishwasher, microwave and kettle are considered. This is done because the availability of data for most devices is highly inconsistent among the houses, while the refrigerator, the dishwasher and the microwave are the only appliances that exist in both REDD and UK-DALE. Houses 1 and 5 are selected for training and house 2 for testing. The reason for that is the fact that only houses 1, 2 and 5 include all the considered appliances, while house 2 misses very few recordings. Moreover, only roughly the first three months of house 1 are used since its total duration is more than four years, which is deemed to be unnecessarily long. Same as REDD, the neural network models are validated on 10% of the training houses data. Figure 2 shows the distribution of the load of the considered appliances of the second household.

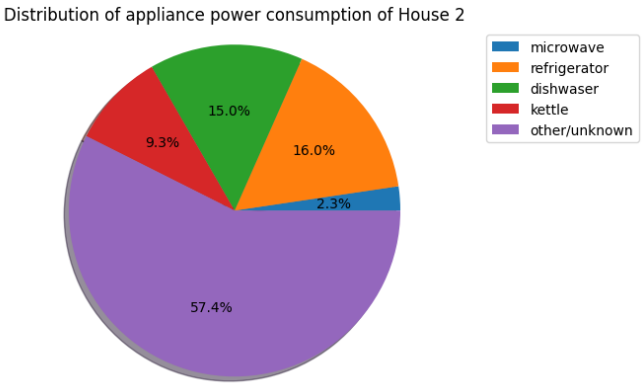


Figure 2: Distribution of the power consumption of the individual appliances in house 2 of the UK-DALE dataset.

The remaining 57.4% of the measured consumption comes from unknown sources (i.e., devices that have not been recorded) or from devices that are included in the

dataset but are irrelevant to this research. Finally, Table 2 shows details about the train-test split for UK-DALE and Figure 3 presents a sample day of the measured total and sub-metered loads taken from house 2.

Table 2: Train-test split for UK-DALE.

Appliance	**Training**	**Testing**
Refrigerator	House 1	House 2
Dishwasher	House 1,5	House 2
Microwave	House 1,5	House 2
Kettle	House 1	House 2

3.2 Preprocessing

Before the data are used as input to begin the training of the network, a set of preprocessing steps must take place. Initially, the sub-metered and total consumption data are aligned according to the respective date-time index of each dataset. Not all devices have been recorded for the same period of time, while also patches of measurements of varying length are missing possibly due to the downtime of the respective smart-plug. This results in the existence of missing values in various parts of both datasets and we choose to entirely remove them.

Since one of the research goals is to observe how the time resolution of the consumption data affects the quality of the disaggregation, resampling of both datasets must take place. Given that the original recording frequencies of the two datasets are not the same, different resampling techniques are applied to each one of them. We test REDD on three different resolution rates. Firstly, we upsample the appliance measurements from one recording every three seconds to one recording every second by linear interpolation. Additionally, the entire dataset (i.e., both the individual and aggregate loads) is downsampled to recordings of every six and every ten seconds. It is not possible to follow the same process for the UK-DALE dataset because the original recording frequency of the aggregate consumption is 1/6 Hz and not 1 Hz as in REDD. This means that even if we upsample the data to 1 Hz and train the model using them, there will be no objective way to perform the evaluation as we will have to also use artificially upsampled test data. Hence, UK-DALE is used only in its original 1/6 Hz resolution and downsampled to 1/10 Hz, as well. The performance results for each sample frequency are presented in Section 4.

Then, the total and individual power consumptions are scaled using their respective mean and standard deviation according to the following formula:

$$z = \frac{x - \text{mean}}{\text{std}}, \tag{4}$$

where x is the original value, *mean* is the mean of the consumption and *std* is the standard deviation of the consumption.

Finally, the data are brought to the sequence-to-point format, as described in 2.3, before they are fed as input to the network.

3.3 CNN architecture

The CNN architecture used for the experiments is similar to the one proposed in [14] and can be seen in Figure 4. In particular, it consist of five 1-dimensional convolutional layers with varying kernel size and number of filters, a dense layer with 1024 units and the output layer that has a single linear unit. Both the dense and the convolutional layers have LeakyReLu activation functions. The input of the network is windows of the total consumption sequence that have a size of W timesteps. The authors of the paper suggest that a window size of 599 produces adequate results, and thus W is set at 599 here as well. The model's output is a single scalar value, which corresponds to the predicted consumption of the target appliance at the midpoint of the input window, as described in 2.3.

3.4 Performance metrics

Four widely used metrics are computed to evaluate the disaggregation performance of the CNN model. Namely, Mean Squared Error (MSE), Mean Absolute Error (MAE), Accuracy and F1-score. MSE and MAE are measured by using the predicted consumption of the target appliance and the ground truth according to the following formulas:

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{x}_t)^2 \tag{5}$$

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|x_t - \hat{x}_t|, \tag{6}$$

where T represents the total number of timesteps of the target appliance consumption sequence, x_t denotes the ground truth and $\hat{x}_t$ the prediction of the appliance's load at time t. MSE being more sensitive to outliers provides a good estimate of the model's ability to predict sparse and high power activations, but also of isolated predictions that are particularly inaccurate. On the other hand, MAE is less sensitive to extreme values and provides an intuitive interpretation of the error rate at any given point in time.

Looking at the plots in Figure 3 it is clear that for the majority of the devices the consumption data are highly imbalanced. In particular, their consumption signal consists of mainly inactive periods with load values close to 0 watts and sparse activation peaks that last relatively short. This property of the data requires a thoughtful evaluation scheme, as basing results on the wrong performance metrics can lead to misleading interpretations of the results. F1-score is used to assess if the model can properly address this exact class imbalance in the data. In order to calculate the class-wise metrics (i.e., Accuracy and F1-score), the on-off status of the devices is required and can be inferred by comparing the appliance signature with the predefined thresholds of Table 3. The selection of these threshold values are based on literature conventions [28]. Accuracy is equal to the amount of correctly predicted time points over the entire sequence length, while F1-score is computed according to:

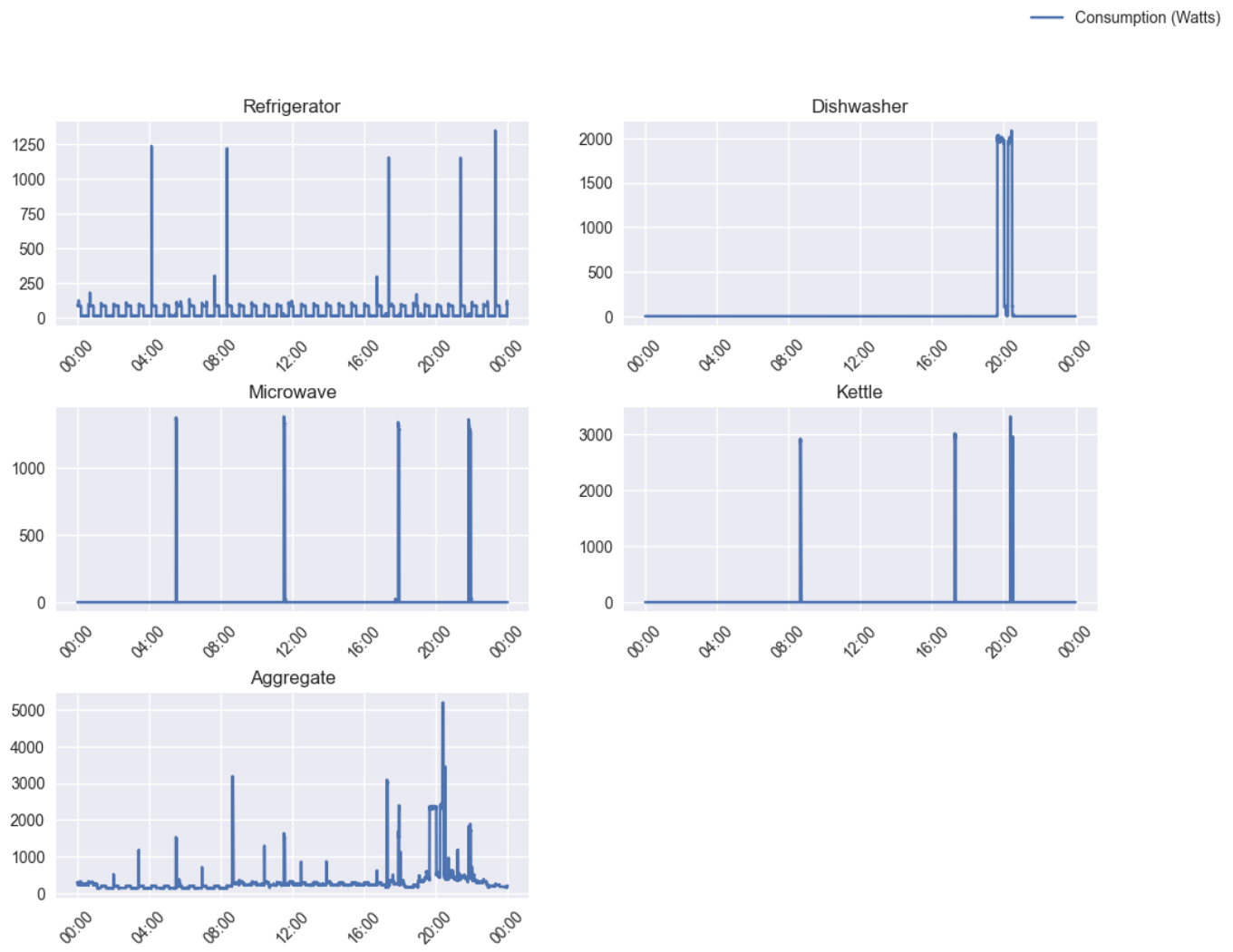


Figure 3: Excerpt of the total and sub-metered power consumptions taken from House 2 of the UK-DALE dataset.

$$\text{F1} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \ , \tag{7}$$

where TP stands for True Positives, FP for False Positives and FN for False Negatives. F1-score is considered to be a reliable metric when it comes to highly imbalanced problems as it balances the precision-recall trade-off. Hence, it is an objective measure of a model's ability to detect appliance activations in imbalanced data and minimize false positives. This is why it can arguably be considered the most important metric in the field of NILM.

Table 3: On state threshold values.

Appliance	On Threshold (watts)
Refrigerator	50
Dishwasher	10
Microwave	200
Kettle	2000
Lighting	20

3.5 Experimental setup

After the data have been preprocessed and brought to the sequence-to-point format they are fed to the CNN model in order to begin the training process. A separate model is trained for each target device of the two datasets, as mentioned in 3.1. This means that eight distinct models are created, each able to perform load disaggregation for its respective appliance. Experimentation has shown that different hyperparameters work better for each device. Moreover, it has also been found that for some devices training on the original values of the data yields better results, hence the scaling step is skipped. Ultimately, this means that the training setup slightly deviates between appliances. Finally, these models are trained on the different time resolutions of the data, as described in 3.2, in order to observe how increasing or decreasing the sampling rate affects their disaggregation capability.

For the cross-domain experiments REDD had to be downsampled to 1/6 Hz in order to be in the same frequency as UK-DALE. The same training and testing houses are used again, as shown in 3.1, in order to have comparable results. Our goal has been to evaluate how features learned by a specific device of a source domain

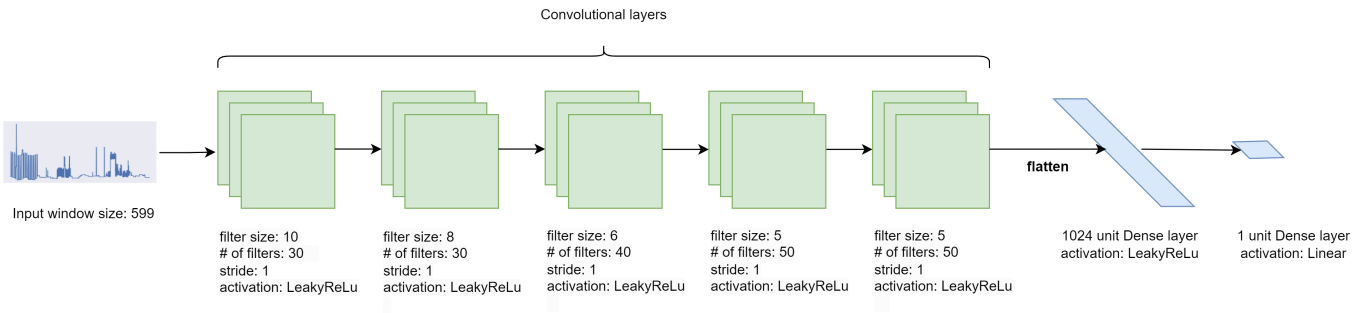


Figure 4: The architecture for the sequence-to-point CNN model.

are transferred to the same device of a target domain, thus only overlapping devices of the two datasets can be used. As a result, lighting and kettle are dropped and the tests are performed only on the refrigerator, dishwasher and microwave. First, the model is trained on each one of the devices on REDD and tested on UK-DALE. The same procedure is repeated but with the two domains reversed.

The results for both parts of the experimentation can be found in Section 4. All experiments have been conducted using the GPU nodes of the Hábrók high performance computing cluster of University of Groningen. The codebase of the experiments can be found on Github [1].

4 Results

This section includes the results of all the experiments. All the performance metrics of the CNN model are averaged over three separate training iterations. Table 4 and Table 5 include the baseline results of the CO model for UK-DALE and REDD respectively.

Table 4: Combinatorial Optimization model performance metrics for House 2 of UK-DALE.

Appliance	MAE	MSE	Acc.	Fscore
Refrigerator	52.04	4609.83	0.411	0.562
Dishwasher	46.14	72153.07	0.961	0.147
Microwave	14.6	17718.445	0.989	0.253
Kettle	44.48	90816.3	0.983	0.474

4.1 Sampling rate experiments

Table 6 and Table 7 include the evaluation metrics of the disaggregation performance of the model on different sampling resolutions of UK-DALE and REDD. Asides from scalar metrics, a visual inspection of the actual predictions is needed to have a complete sense of the quality of the estimated device loads. Thus, Figures 5 and 6 provide plots of the predicted consumptions of each target appliance of the test houses in the original resolutions of the two datasets.

[1] https://github.com/philip-andreadis/Seq2point-for-NILM

Table 5: Combinatorial Optimization model performance metrics for House 1 of REDD.

Appliance	MAE	MSE	Acc.	Fscore
Refrigerator	99.34	17349.89	0.464	0.279
Dishwasher	41.1	35574.21	0.944	0.204
Microwave	52.02	29118.53	0.988	0.0
Lighting	60.22	5322.61	0.325	0.313

4.2 Cross-domain experiments

Table 8 and Table 9 include the results of the model trained on REDD and tested on UK-DALE and vice versa. Specifically, the model is trained on all the train houses of the source domain and tested on the test house of the target domain, which remain the same as in the first part of the experiments. As expected, the results suggest a noticeable performance drop in most cases when compared with Table 6 and Table 7. Interestingly, the refrigerator shows some transferability potential and Figure 7 provides a visual representation of its predicted load on both cases.

5 Discussion

Considering the performance metrics of Table 6 and Table 7, for the majority of the devices of UK-DALE it seems that the model performs better on the original frequency of the data (i.e., sampling rate of 6 seconds). The only exception is the dishwasher, where it benefited from the downsampling. As for REDD, the model is clearly performing the best on the original resolution (i.e., sampling rate of 3 seconds) of the refrigerator and lighting data, while for the rest of the devices the results are ambiguous. The

Table 6: Comparison of model's performance on different sampling rates of UK-DALE.

Appliance	Sampling Rate (sec)	MSE	MAE	Accuracy	Fscore	Training Time (min)
Refrigerator	6	1203.45	**15.856**	**0.896**	**0.876**	33
	10	**1170.08**	17.66	0.868	0.84	23
Dishwasher	6	12895.13	17.19	0.948	0.511	70
	10	**7880.62**	**12.15**	**0.956**	**0.554**	56
Microwave	6	**7407.06**	**12.48**	**0.995**	**0.257**	15
	10	7703.46	14.41	0.993	0.2	11
Kettle	6	**9172.93**	**15.036**	0.997	**0.884**	20
	10	12366.32	20.48	0.997	0.841	13

Table 7: Comparison of model's performance on different sampling rates of REDD.

Appliance	Sampling Rate (sec)	MAE	MSE	Accuracy	Fscore	Training Time (min)
Refrigerator	1	35.05	4421.44	0.857	0.756	151
	3	**28.23**	**3493.51**	**0.882**	**0.805**	43
	10	32.24	3665.42	0.855	0.77	18
Dishwasher	1	12.37	5131.98	0.95	0.264	190
	3	4.95	3501.85	**0.987**	**0.538**	52
	10	**4.68**	**1758.09**	0.944	0.238	18
Microwave	1	**14.65**	28522.23	0.961	**0.678**	208
	3	17.77	**21517.16**	**0.986**	0.517	55
	10	23.3	35527.3	0.984	0.387	38
Lighting	1	42.49	3934.35	0.465	0.03	253
	3	**39.02**	3228.43	**0.578**	**0.545**	45
	10	39.93	**3153.35**	0.506	0.355	29

Table 8: Results of the model trained on REDD and tested on UK-DALE.

Appliance	MSE	MAE	Acc.	Fscore
Refrigerator	3161.62	37.49	0.776	0.758
Dishwasher	1737.31	12.2	0.892	0.233
Microwave	13588.16	12.28	0.991	0.311

Table 9: Results of the model trained on UK-DALE and tested on REDD.

Appliance	MSE	MAE	Acc.	Fscore
Refrigerator	6632.38	47.62	0.824	0.644
Dishwasher	44703.08	39.12	0.927	0.277
Microwave	25789.78	28.79	0.985	0.29

model has benefited from the upsamling of the microwave data to 1 Hz, as the F1-score is noticeably higher in that case plus it yields the lowest MAE. On the other hand, the dishwasher model has a much higher F1-score on the original data resolution, but presents a considerably lower MSE on the downsampled signal. However, the dishwasher consumption pattern consists of infrequent and short activation spikes of high magnitude and by downsampling the signal parts of these spikes can be lost. Given also the fact that MSE punishes larger errors, here it could provide a more favourable outcome as the model is never evaluated on the omitted information of the spikes. This means that the MSE score cannot be considered as reliable as the F1-score, and hence we conclude that the model performs best on the original frequency of the dishwasher signal.

Since one of the main motives of testing a disaggregation system on decreased resolutions of data is to save computation time, it is important to mention how resampling has affected training times, as well. Specifically, the downsampling of UK-DALE to 1/10 Hz results in an average 27.75% decrease in training time. Downsampling REDD to 1/10 Hz yields an average of 47.25% lower training time and after upsampling to 1 Hz there is an average 314% increase compared to the original resolution.

Due to the nature of the NILM problem a visual inspection of the results can provide valuable insights and create a more comprehensive understanding of the final disaggregation quality. Figures 5 and 6 suggest that the model can predict very well the pattern of the refrigerator, which is also evident from the performance metrics. Asides from some instantaneous surges, it can learn to follow the signal well mainly due to its repetitive and constant form. As for the dishwasher the model can adequately locate all the activation spikes, with some false negatives on UK-DALE.

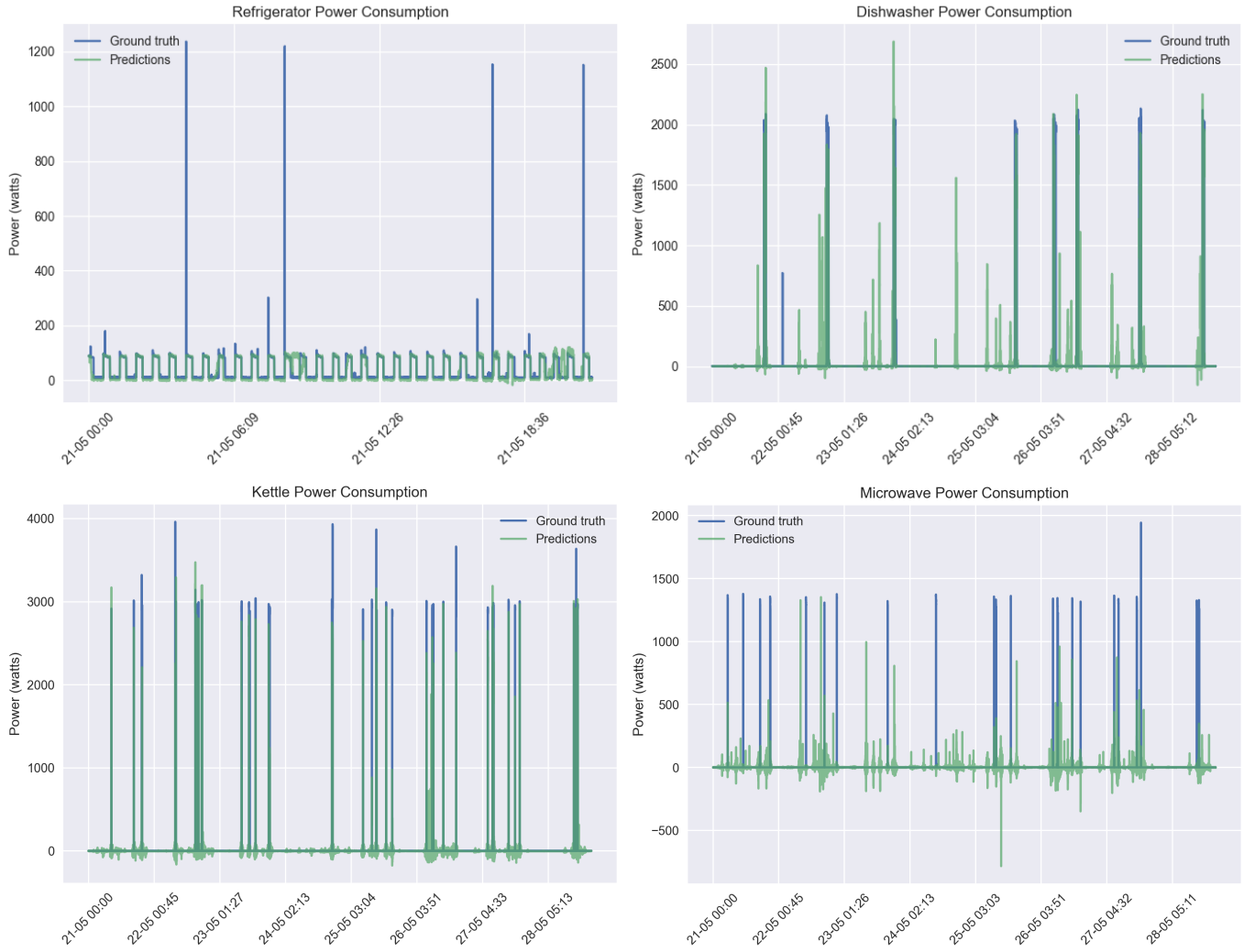


Figure 5: Disaggregation performed on the original resolution of House 2 of UK-DALE. A smaller section of the refrigerator load is used for better visualisation.

There is a clear distinction between the disaggregation quality of the UK-DALE and the REDD microwave. The model can predict many of the activations on REDD with very few false negatives, whereas the UK-DALE plot suggests a much more spasmodic pattern with multiple erroneous activations some of which are even below zero. This observation is inline with the F1-score which is approximately double for REDD. The model seems to be able to perform a very good disaggregation of the kettle load, which can also be supported by the high F1-score. Lastly, it is obvious that the lighting model yields the worst results overall. The reason for that could be multifaceted. First of all, the low magnitude of the activations in the test house suggest that only a subset of the lights has been monitored (e.g., a singe room of the house). Additionally, the consumption pattern can vary greatly between households, even in the same domain, since lights can have a higher and more direct correlation to human activity compared to other appliances. All that can point to the need for a better integrated recording system for lights, along with a more complex disaggregation model.

Regarding the cross-domain disaggregation tests, by observing Table 8 and Table 9 it is obvious that there is a noticeable drop in the overall performance. As expected, on average the F1-score and error rates of the cross-domain models are lower compared to the ones trained and tested on the same dataset. One interesting exception is that of the UK microwave, where it presents a slightly higher F1-score but also a much higher MSE when trained on REDD rather than UK-DALE. The case might be that the network is able to learn the activation pattern of the appliance as it is similar in both domains, however it is not able to accurately predict the magnitude of those activations. This could be attributed to the fact that appliances in the two regions consume different amounts of energy when they are being operated. One can also notice that the performance drop can vary between appliances. For example, the F1-score of the dishwasher is almost halved in both scenarios, whereas this is not true for the refrigerator as the decrease is much smaller in relative terms.

Figure 7 reveals that the quality of the disaggregation performed on the UK refrigerator load by the model trained on the US dataset is better than the opposite case. The reason for this could be that the train set of REDD consists of three houses, whereas only one for UK-DALE. This way the model is able to fit on more diverse con-

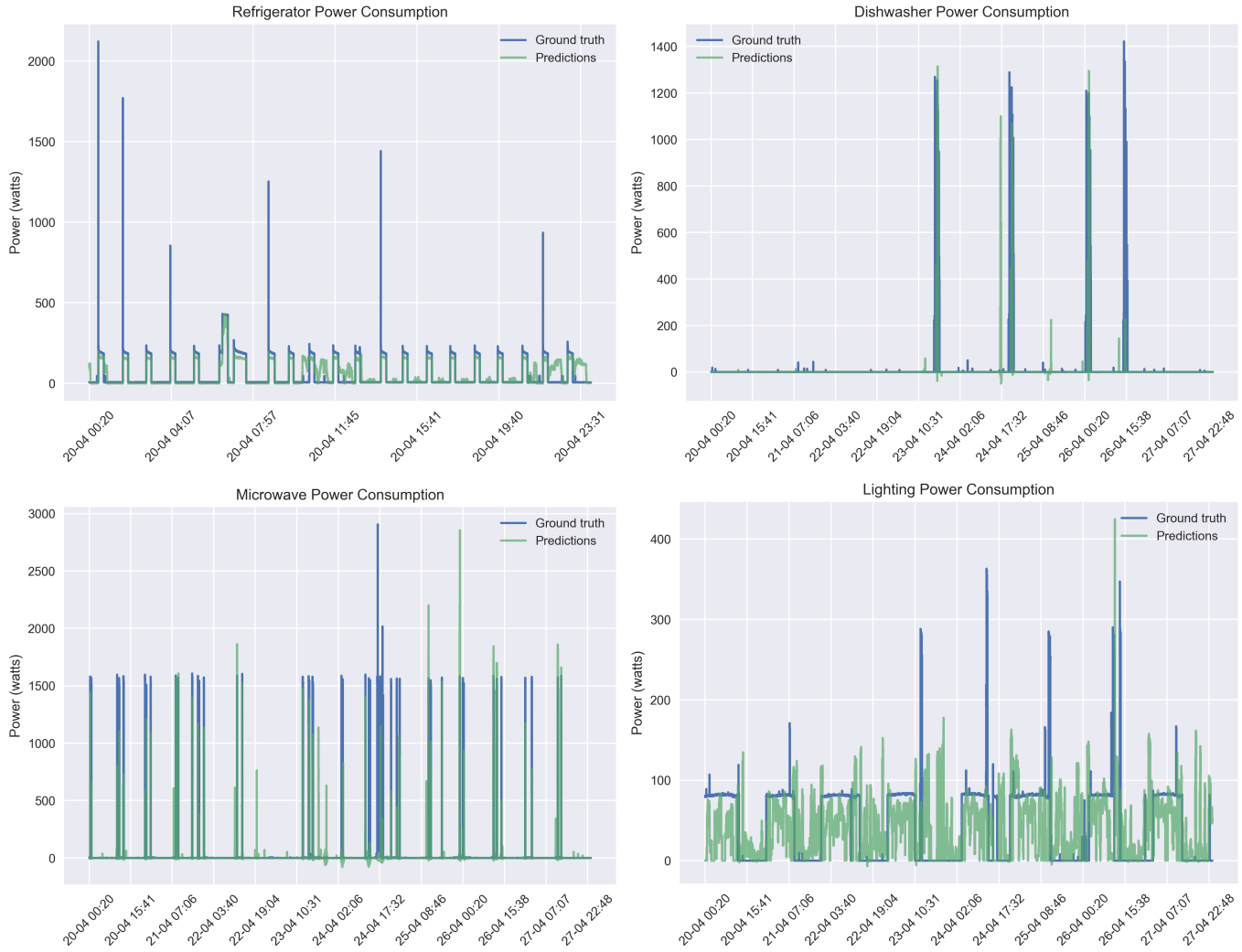
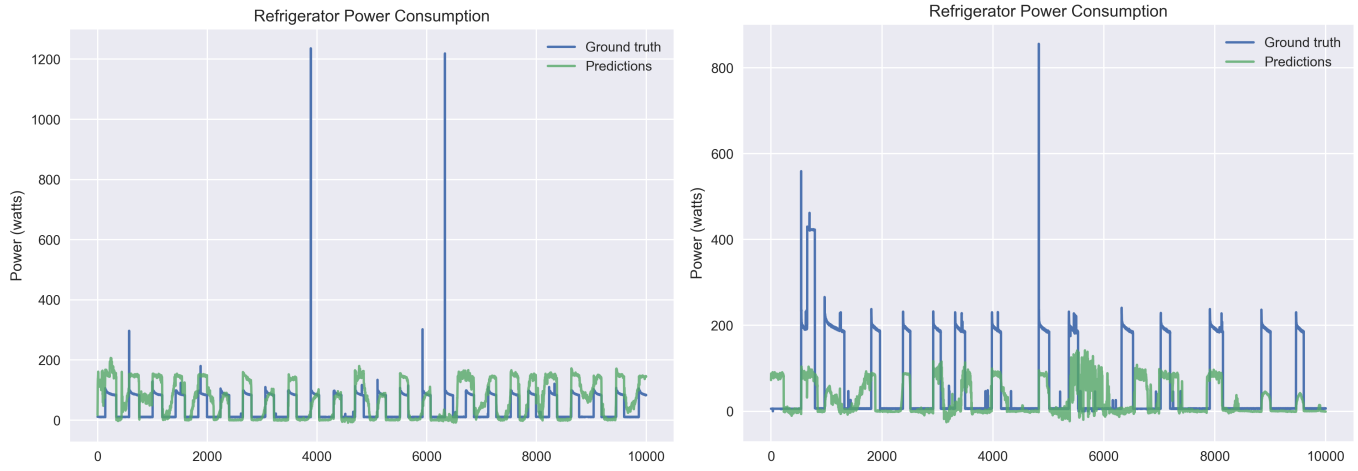


Figure 6: Disaggregation performed on the original resolution of House 1 of REDD. A smaller section of the refrigerator load is used for better visualisation.



(a) (b)

Figure 7: Cross-domain disaggregation of the refrigerator. (a) shows predictions of the model trained on REDD and tested on House 2 of UK-DALE. (b) shows predictions of the model trained on UK-DALE and tested on House 1 of REDD.

sumption patterns and hence, have a better generalization capacity. One more interesting detail is the fact that the magnitude of the activations of the US refrigerator is approximately double of that of the UK appliance. This

discrepancy between the domains is directly reflected on the predictions, as in Figure 7(a) the model overshoots the activations and in (b) it under-estimates them.

6 Conclusion

In this research we study the sequence-to-point learning methodology, coupled with a CNN model. We test the network's disaggregation performance on different resolutions of two NILM datasets. We also evaluate the model's ability to perform appliance detection between two different domains in order to estimate the transferability of learned appliance features.

Overall, maintaining the original sampling rate has yielded the best results for the majority of the tested devices. Nonetheless, it is concluded that downsampling can be used as a viable option, when storage space and computational resources are limited and some potential performance drop can be tolerated. The same cannot be stated for upsampling, as corresponding results do not justify the huge increase in training time.

The initial scope of this research was to include an additional fine tuning step in the cross-domain experiments. Specifically, the model trained on the source domain is also fine tuned on data from a train house of the target domain before it is tested. Due to time restrictions, this part has not been explored to a sufficient extent and the results are not included. The preliminary results presented here suggest that building a cross-domain model with adequate performance is possible for a subset of the available appliances. Hence, this idea can be pursued as promising future work.

One more possible future direction could be to implement and experiment with more complex DL architectures that solve the NILM task. Lastly, it would also be interesting to move even lower on the sampling frequency in order to discover the lowest resolution threshold that allows successful appliance detection and also integrate additional features such as time of day or weather information to see whether they improve performance.

References

[1] I. E. Agency, *Net Zero by 2050*. 2021.

[2] B. Yu, Y. Tian, and J. Zhang, "A dynamic active energy demand management system for evaluating the effect of policy scheme on household energy consumption behavior," *Energy*, vol. 91, pp. 491–506, 2015.

[3] O. Elma and U. S. Selamoğullar, "A survey of a residential load profile for demand side management systems," in *2017 IEEE International Conference on Smart Energy Grid Engineering (SEGE)*, pp. 85–89, 2017.

[4] "Energy statistics - an overview." https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_statistics_-_an_overview#Final_energy_consumption. Accessed: 2023-11-30.

[5] J. Kelly and W. Knottenbelt, "Does disaggregated electricity feedback reduce domestic electricity consumption? a systematic review of the literature," *arXiv preprint arXiv:1605.00962*, 2016.

[6] M. N. Meziane, P. Ravier, G. Lamarque, K. Abed-Meraim, J.-C. Le Bunetel, and Y. Raingeaud, "Modeling and estimation of transient current signals," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1960–1964, IEEE, 2015.

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

[8] Y. Du, L. Du, B. Lu, R. Harley, and T. Habetler, "A review of identification and monitoring methods for electric loads in commercial and residential buildings," in *2010 IEEE Energy Conversion Congress and Exposition*, pp. 4527–4533, IEEE, 2010.

[9] D. Garcia-Perez, D. Pérez-López, I. Diaz-Blanco, A. Gonzalez-Muniz, M. Dominguez-Gonzalez, and A. A. C. Vega, "Fully-convolutional denoising auto-encoders for NILM in large non-residential buildings," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2722–2731, 2020.

[10] H. Çimen, N. Çetinkaya, J. C. Vasquez, and J. M. Guerrero, "A microgrid energy management system based on non-intrusive load monitoring via multitask learning," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 977–987, 2020.

[11] Á. Hernández, A. Ruano, J. Ureña, M. Ruano, and J. Garcia, "Applications of NILM techniques to energy management and assisted living," *IFAC-PapersOnLine*, vol. 52, no. 11, pp. 164–171, 2019.

[12] J. Huchtkoetter and A. Reinhardt, "A study on the impact of data sampling rates on load signature event detection," *Energy Informatics*, vol. 2, pp. 1–12, 2019.

[13] J. Huchtkoetter and A. Reinhardt, "On the impact of temporal data resolution on the accuracy of non-intrusive load monitoring," in *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pp. 270–273, 2020.

[14] C. Zhang, M. Zhong, Z. Wang, N. Goddard, and C. Sutton, "Sequence-to-point learning with neural networks for non-intrusive load monitoring," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[15] H. Kim, M. Marwah, M. Arlitt, G. Lyon, and J. Han, "Unsupervised disaggregation of low frequency power measurements," in *Proceedings of the 2011 SIAM international conference on data mining*, pp. 747–758, SIAM, 2011.

[16] R. Machlev, J. Belikov, Y. Beck, and Y. Levron, "MO-NILM: A multi-objective evolutionary algorithm for NILM classification," *Energy and Buildings*, vol. 199, pp. 134–144, 2019.

[17] K. He, L. Stankovic, J. Liao, and V. Stankovic, "Non-intrusive load disaggregation using graph signal processing," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 1739–1747, 2016.

[18] F. Gong, N. Han, Y. Zhou, S. Chen, D. Li, and S. Tian, "A svm optimized by particle swarm optimization approach to load disaggregation in non-intrusive load monitoring in smart homes," in *2019*

IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2), pp. 1793–1797, IEEE, 2019.

[19] C. C. Yang, C. S. Soh, and V. V. Yap, "A non-intrusive appliance load monitoring for efficient energy consumption based on naive bayes classifier," *Sustainable Computing: Informatics and Systems*, vol. 14, pp. 34–42, 2017.

[20] F. Hidiyanto and A. Halim, "Knn methods with varied k, distance and training data to disaggregate nilm with similar load characteristic," in *Proceedings of the 3rd Asia Pacific Conference on Research in Industrial and Systems Engineering*, pp. 93–99, 2020.

[21] Z. Xiao, W. Gang, J. Yuan, Y. Zhang, and C. Fan, "Cooling load disaggregation using a NILM method based on random forest for smart buildings," *Sustainable Cities and Society*, vol. 74, p. 103202, 2021.

[22] X. Wu, Y. Gao, and D. Jiao, "Multi-label classification based on random forest algorithm for non-intrusive load monitoring system," *Processes*, vol. 7, no. 6, p. 337, 2019.

[23] S. J. Buchhop and P. Ranganathan, "Residential load identification based on load profile using artificial neural network (ann)," in *2019 North American Power Symposium (NAPS)*, pp. 1–6, IEEE, 2019.

[24] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[25] A. Harell, S. Makonin, and I. V. Bajić, "Wavenilm: A causal neural network for power disaggregation from the complex power signal," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8335–8339, IEEE, 2019.

[26] P. Huber, A. Calatroni, A. Rumsch, and A. Paice, "Review on deep neural networks applied to low-frequency NILM," *Energies*, vol. 14, no. 9, p. 2390, 2021.

[27] T. Sirojan, B. T. Phung, and E. Ambikairajah, "Deep neural network based energy disaggregation," in *2018 IEEE International conference on smart energy grid engineering (SEGE)*, pp. 73–77, IEEE, 2018.

[28] S. Sykiotis, M. Kaselimi, A. Doulamis, and N. Doulamis, "Electricity: An efficient transformer for non-intrusive load monitoring," *Sensors*, vol. 22, no. 8, p. 2926, 2022.

[29] 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, pp. 59–62, Citeseer, 2011.

[30] 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, no. 150007, 2015.