Semi-supervised energy disaggregation for real-world adoption

Tobias Albert Kohl  
*Goethe University Frankfurt, tokohl@wiwi.uni-frankfurt.de*

Domenic Kellner  
*Goethe University Frankfurt, kellner@wiwi.uni-frankfurt.de*

Cristina A. Mihale-Wilson  
*Goethe University Frankfurt, mihale-wilson@wiwi.uni-frankfurt.de*

Follow this and additional works at: [https://aisel.aisnet.org/ecis2021_rp](https://aisel.aisnet.org/ecis2021_rp)

**Recommended Citation**

[https://aisel.aisnet.org/ecis2021_rp/68](https://aisel.aisnet.org/ecis2021_rp/68)

This material is brought to you by the ECIS 2021 Proceedings at AIS Electronic Library (AISeL). It has been accepted for inclusion in ECIS 2021 Research Papers by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
SEMI-SUPERVISED ENERGY DISAGGREGATION FOR REAL-WORLD ADOPTION

Research Paper

Tobias Kohl, Goethe-Universität Frankfurt, Frankfurt, Germany, tokohl@wiwi.uni-frankfurt.de

Domenic Kellner, Goethe-Universität Frankfurt, Frankfurt, Germany, kellner@wiwi.uni-frankfurt.de

Cristina Mihale-Wilson, Goethe-Universität Frankfurt, Frankfurt, Germany, mihale-wilson@wiwi.uni-frankfurt.de

Abstract

In this paper, we address the topic of energy disaggregation for real-world implementation. The idea behind energy disaggregation is to utilize the aggregated energy consumption data from smart meters to identify individual appliances in the load by decomposition. However, appliance level usage prediction for residential areas is currently not practical. Yet Smart Living environments might benefit from non-intrusive load monitoring (NILM) adoption. We therefore first lay a theoretical foundation by working out the core challenges for NILM adoption in real-world scenarios. We evaluate the semi-supervised model performance on a publicly available dataset REFIT. By reducing the sampling rate to 32 seconds as well as limiting the available labeled data for training, we give significant insight into the usability of semi-supervised NILM under real-world conditions. The availability of unlabeled observations seems to strengthen model prediction performance, allowing it to outperform a fully trained supervised model for certain appliances.

Keywords: energy disaggregation; non-intrusive load monitoring (NILM); machine learning; semi-supervised learning
1 Introduction

The advancements in technology and the development of new technologies have increased the amount of electricity we consume (He and Chai, 2016). Because much of the energy consumed gets produced with fossil fuels, increased energy consumption has local and global adverse effects such as air pollution and climate change. These in turn have negative consequences not only on individuals' health but also on societies as a whole. As climate change will change temperature and precipitation patterns, it is innately linked to rising sea levels, an increasing number of hurricanes and droughts, and thus changes in agricultural production. Therefore, it is not surprising that the avoidance or slow-down of climate change through the decrease of energy consumption is a top priority in many countries. Following the growing issue of climate change, energy consumption has become a major topic of discussion resulting in a growing political demand to control residential energy usage. Ensuing this demand, Germany validated a new legislature, Smart Meters Operation Act (Messestattenbetriebsgesetz) MsbG 2016, requiring the obligatory installation of smart meters in all houses with a yearly consumption of over 6000KW since 24.02.2020 (BDEW, 2020). Despite such a strong political interest in smart metering, the realization of the actual benefit in the form of energy savings will have to happen on the level of residential households.

One fruitful path for energy consumption reduction is conscious energy use in industrial and private environments. This work focuses on private environments and more specifically on how to leverage Artificial Intelligence to optimize energy consumption patterns in the residential context. A prerequisite for optimizing the energy consumption in the residential context is recognizing the usage of individual appliances. In contrast to the industrial setting, appliances and machines in the residential context are not necessarily equipped with sensors and measuring instrumentation for energy consumption. Hence, the energy consumption of individual appliances cannot be read off directly but rather derived from the household's total energy consumption.

Energy disaggregation, also called “Non-Intrusive Load Monitoring” (NILM), describes the process of analyzing aggregated energy consumption data to predict appliance energy utilization. In simple terms, NILM can produce energy utilization predictions by recognizing so-called appliance-specific load signatures within the aggregated power consumption. Load signatures represent patterns containing information parameters, which can in turn be used to identify appliance activities from the aggregated power consumption (Zoha et al., 2012; Shin, Rho, et al., 2019). Scholars have already harnessed AI techniques such as supervised and unsupervised Machine Learning (ML) for energy data disaggregation. However, the adoption of NILM models on a larger scale faces a series of practical challenges. One central challenge lies in the heterogenic nature of the used appliances in households. Not only do the load signatures for one appliance differ between various manufacturers, but also between different device models or different usage states. Additionally, not only can appliances in the same categories differ strongly, but also just between users and their patterns (Zoha et al., 2012). As a result, most trained NILM models are very limited in their transferability (Shin, Rho, et al., 2019).

Due to the limited transferability of NILM models, another pressing issue is the models' actual training. Most models need to train on the local data of the intended environment. While current supervised models, like subtask gated networks (Shin, Rho, et al., 2019) or WaveNet models (Jiang et al., 2019) have shown remarkable performance and improvements in prediction accuracy, they also require high frequency labeled data for training and prediction. Because high frequency labeled data is typically not available within the residential context, supervised ML for residential energy data disaggregation is not feasible. Within the residential context, one can typically access the total load or total energy consumption via a smart meter. Additionally, the data stemming from smart meters are unlabeled and have a very low frequency that can range between 15 seconds to 1 hour (Chang and Ho, 2019). These challenges make the application of NILM models in real world environments not very straightforward. In theory, the deployment of smart power outlets or smart plug can help the collection of highly granular appliance-specific energy consumptions. These plugs can measure the power consumption of each appliance individually. However, in practice retrofitting, smart plugs to every appliance is often not
Semi-supervised Learning for NILM

economically feasible (Zoha et al., 2012). The corresponding hardware and data management costs are quite substantial.

To address these challenges and advance the implementation of energy disaggregation in the residential context, we follow Barsim and Yang (2015) and Chang and Ho (2019) and argue that semi-supervised learning (SSL) might be a fruitful technique for NILM. Further, we argue that as a combination of supervised and unsupervised learning, SSL methods in practice reflect the real-world residential conditions better than supervised ML. By addressing the economic and technical feasibility through using SSL; Smart Living environments, such as multi-household living complexes, could be particularly interesting on the topic of NILM adoption. The facility’s might feature very similar furniture allowing for similar data structures and observation patterns, making the issues concerning NILM transferability between households more amenable. Accordingly, SSL-based disaggregation models could be easier transferable than supervised ML-based models. In this vein, this paper addresses the following research question:

*Can semi-supervised learning help to advance NILM adoption by addressing the core challenges for energy disaggregation under real-world conditions?*

Formally, this contribution is structured as follows: In the next section, we discuss the challenges to NILM in real-world scenarios. Then, we evaluate the performance of a simple SSL approach, namely self-learning, in comparison to standard supervised learning approaches. To this end, we use a publicly available dataset, called REFIT. We manipulate the amount of labels and the frequencies of the REFIT data to investigate SSL’s prediction power for NILM. In particular, we compare multiple SSL models trained on different amounts of labeled and unlabeled data against a supervised baseline ML model. While we expect to see a decline in prediction accuracy when using SSL, evaluating the SSL’s performance based on various amounts of labels and sampling frequencies allows us to generate significant and practice-relevant insights about the real-world prediction performance of SSL. Section four presents the results of the comparisons between various SSL models with the baseline model – i.e., the supervised ML model. The last section concludes, discusses the limitations of this work and presents fruitful paths for future research.

2 Literature Review

2.1 Non-Intrusive Load Monitoring (NILM)

Managing appliance activities or providing energy and activity feedback for other smart services will empower a diverse set of advanced smart service use cases ranging from domestic energy saving to buildings responding to fluctuating energy prices or smart grid responses by home activity automation (Murray, Stankovic and Stankovic, 2017). However, all this requires context data for energy consumption and usage pattern of each appliance. The smart meter rollout will provide easy access to aggregated energy consumption measurements. Nevertheless, it is not a trivial task to actually utilize these measurements to manage residential energy demand better and conserve energy.

As a solution to this problem, Non-Intrusive Load Monitoring (NILM) was proposed (Hart, 1992). NILM describes a decomposition approach for energy disaggregation that focuses on separating the aggregated energy consumption measurements inside a household to identify individual appliances, without relying on measuring each appliance’s usage via submetering (Shin, Rho, et al., 2019). NILM focuses on recognizing so-called load signatures as measurable parameters, as these contain the information for specific appliance activities (Shin, Rho, et al., 2019) and are thereby used to give information about the operating state of specific appliances inside the aggregated measured consumption (Hart, 1992). A promising solution to realize NILM can be Machine Learning (ML), either in the form of supervised or unsupervised learning (D’Incecco, Squartini and Zhong, 2019). While unsupervised models can be used to decompose aggregated energy consumption data, their usability is very limited. Most unsupervised approaches are based on Hidden Markov models (HMM), which require a lot of custom parameter configuration, making them only applicable in specific circumstances and generally worse considering transferability (Rahimpour et al., 2017). Supervised models have therefore taken the
spotlight, by achieving strong performances while also being more versatile. Especially, deep learning approaches, such as convolutional neural networks (CNN) have been used to great success. These supervised NILM models can broadly divide into optimization or pattern recognition algorithms. Models such as subtask gated networks (Shin, Rho, et al., 2019) or WaveNet models (Jiang et al., 2019) achieved state of the art performances and have shown themselves to be particularly useful for NILM (D'Incecco, Squartini and Zhong, 2019). Even though we see well-performing models in the literature, most NILM approaches currently have three major limitations to be fully useful in real-world scenarios, namely transferability, the necessity of large amounts of data and a required high sampling rate.

2.2 Challenges for NILM adoption

Large amounts of labeled data

Even though supervised learning algorithms show a great performance, they require large amounts of labeled data to train the classifier in order to recognize individual appliance operations from the aggregated load measurements. To acquire the data, based on the local circumstances, each appliance would need to be submetered in an initial setup instrumentation, which would take a lot of human effort on site for the installation and incurs extra cost for data management and preparation (Zoha et al., 2012).

High sampling rate

While most of the aforementioned ML models (given large amounts of labeled training data) can show very promising results, only using the aggregated energy consumption data collected by smart meters might end up being difficult. Current commercially available meters show at least a variation in data measurements of 10% to 20% and are often not designed in their hardware specifications for high frequencies. While current high-end smart meter Network Interface Cards (NIC) can support up to 1KHz, even higher frequencies beyond the 5KHz range, would actually make completely new hardware solutions necessary (Zoha et al., 2012). Realistically, current smart meters will only provide data in the 10-15 seconds frequency (at most) while high end models can push these intervals to around 1 second; which is still far from the required frequency range used in most current approaches.

Transferability

Appliance level usage prediction for residential areas is challenging, as most NILM models have only limited transferability. Each appliance type can feature different load signatures for each device and manufacturer, therefore to use load signatures for pattern recognition in an aggregated load, each appliance would need to be measured individually for its consumption by so-called submetering. This problem gets aggregated as appliances can also have multiple clearly identifiable states, besides just on and off (Zoha et al., 2012). Such multi-state appliances are often problematic as each state could feature a unique load pattern, and then get recognized as an individual appliance by the used model (Najafi, Moaveninejad and Rinaldi, 2018). As each household uses different appliances and usage patterns, each model recognizes appliances only based on their load signature. Currently, there is no standardized way for the aggregated data collection, limiting transferability of models further based on the differences in the observed data structures (Pereira, Velosa and Pereira, 2019). Therefore, the trained NILM models end up being trained on concrete domain knowledge that cannot be transferred easily between households (D'Incecco, Squartini and Zhong, 2019).

In our research, we want to mainly focus on the issue of varying amounts of data in a low sampling rate environment as it is the most pressing challenge for NILM adoption. Moreover, as explained above research has already shown that NILM methods perform relatively well when applied in high sampling rate environments. Regarding transferability, we want to give a first indication of the performance of SSL techniques. Nevertheless, we leave it open for future research as SSL could be more beneficial when applied in a Smart Living environment.
2.3 Semi-Supervised Learning for NILM

To tackle these challenges, we propose to use semi-supervised learning, a combination of supervised and unsupervised learning. SSL is an approach that allows for a minimal amount of labeled data to be used in conjunction with a more extensive set of unlabeled observations for model training. Externally well labeled information is often very scarce and costly to procure. For NILM applications this refers to either actively submetering each relevant appliance or directly hand labelling the necessary data by the inhabitants. Therefore, making a purely supervised approach is not only very time consuming but also cost-intensive. With SSL, the requirements for this high cost labeled data can be minimised. A short setup phase with submetering for specific appliances, combined with a larger set of unlabeled observations from the implemented smart meters, could be enough to allow for a strong model training. This is especially true as compared to a supervised model paradigm providing the extra unlabeled observations to the model for self-training could further strengthen the prediction performance. In order to tackle the necessity for large amounts of data, researchers already actively investigated SSL methods (Barsim and Yang, 2015; Rahimpour et al., 2017). As an example, Barsim and Yang (2015) showed that SSL could actually perform comparable to supervised models, especially in situations with only a very small labeled training dataset, resulting in a NILM model that increases performance over time as it observes more unlabeled data instances. However, the model developed is not very transferable for real-world environments, as the used BLUED dataset only collected data over a period of 8 days and more importantly the sampled data frequency was in excess of 10 KHz (Murray, Stankovic and Stankovic, 2017). When looking at the challenge regarding the necessity of a high sampling rate, there is currently only limited research available in relation to SSL (Miao et al., 2019). For real-world instrumentation, the ideal NILM method would thereby need to focus on both low sampling rate and a reduced labeled training data requirement, while ideally also featuring high transferability. Thus, we propose SSL as solution for the aforementioned challenges. First, SSL can leverage on unlabeled observations for self-training or co-training, which decreases the amount of labeled data. Second, SSL can be applied in low sampling rate environments in order to investigate the advantage compared to supervised models. Third, SSL models can be transferred to another household in order to investigate generalisability. Thus, compared to previous research, our approach focuses on low sampling rate data over a longer time span, artificially reduces the amount of labeled data for training, and then applies the trained model to unseen houses.

3 Data and Research Methodology

3.1 Data

To closely mimic real-world data as it will be available through smart meters in residential areas, we use the REFIT - Electrical Load Measurements Dataset (Murray, Stankovic and Stankovic, 2017). It stands apart from other available datasets as it contains data from a large set of 20 houses over a long time span of roughly two years, spanning from 2013 to 2015. What makes REFIT especially useful for our approach is the relatively low sampling rate. REFIT provides data for nine different appliances with a sampling rate of 8 seconds, but not all appliances appear in each house. This is already much more realistic compared to other datasets like BLUED (Anderson et al., 2012) or EnerTalk (Shin, Lee, et al., 2019), which use a sampling rate of 1Hz or higher (Murray, Stankovic and Stankovic, 2017). However the sampling rate is still higher than what actual targets for residential complexes might realize (Chang and Ho, 2019). Thus, to align the data more closely to real word scenarios we will artificially reduce the sampling rate further. We are doing so by keeping only every second observation or every fourth observation, as REFIT reports the data in watts. Thus, we are looking at three different scenarios: (1) The standard REFIT sampling rate of 8 seconds (2) a scenario where the frequency is artificially lowered to 16 seconds and (3) a scenario where the frequency is artificially lowered to 32 seconds. However, we will mainly focus on the last scenario as it is the most realistic as in real-life applications. The aggregate can mostly be extracted from a smart meter, which can produce a sampling rate of roughly 30 seconds
Semi-supervised Learning for NILM

Twenty-Ninth European Conference on Information Systems (ECIS 2021), Marrakesh, Morocco.

at the most (Chang and Ho, 2019). Nevertheless, in case of the other two we will still mention the differences compared to the lowest sampling rate briefly, as it can give valuable insights into what happens when lowering the sampling rate gradually. While we conducted our research over the data provided by all houses we limited ourself to certain appliances and will focus on house 2 in particular in the next section. After conducting exploratory data analysis we realized that house 2 exhibits the least amount of missing observations while having data concerning all of the chosen appliances. Nevertheless, we also included appliances from other house in our experiment in order to ensure robustness. As NILM is especially useful for appliances, which consume a large amount of energy or are more frequently used within a household (Chang and Ho, 2019), we are focusing on the appliances Fridge-Freezer, Washing Machine, Dishwasher, Television Site, and Kettle. To give some evidence for our particular case, we determined the overall consumption of all the appliances monitored in the house and therefore the fraction of overall power consumption per appliance. Thus for house 2 as an example, the dishwasher makes up the largest portion of household energy usage (40.80%), followed by the fridge freezer (24.19%), Kettle (15.54%), washing machine (12.21%), and television (2.63%). Since power usage of washing machines and dishwashers is usually relatively comparable, we will only show the results for washing machines. Moreover, for the train-test setting, we are limiting the dataset to roughly the first three months (eight weeks training and four weeks testing) of the year 2015.

3.2 Research Methodology

To investigate the advantages of SSL using the REFIT data, we will leverage on a simple algorithm called self-learning (Chapelle, Schölkopf and Zien, 2006). The idea behind self-learning is to combine labeled data as well as unlabeled data by pseudo-labeling the unlabeled data using a Deep Neural Network (c.f., Figure 1). Unlabeled data describes the overall energy consumption for the house, while labelled data includes the submetered real consumption of the appliances. Thus, we can see the training process as follows. In the first stage, we are training the neural network based on the labeled data, which

![Figure 1. Self-Learning-Process on labeled and unlabeled observations with Convolutional Neural Network architecture. (Zhang et. al, 2016).](image-url)
can vary in our case between one and seven weeks. In the next step, the unlabeled data is pseudo-labeled using the neural network trained in the first stage. In the second stage, the neural network with the same architecture is retrained using the labeled data as well as the pseudo-labeled data. Afterwards, the model performance can be tested on a test set, which corresponds to four weeks in our case. Moreover, to see whether adding unlabeled data increases the accuracy, we can compare all the first stage models (as a supervised model benchmark) to the second stage models (as a SSL benchmark). Moreover, a supervised model will be trained over the full training period of eight weeks using labeled data. Thus, when conducting the experiment we will iteratively increase the amount of labeled data to see whether the SSL model accuracy converges against the supervised model accuracy (using full information), while also showing that adding unlabeled observations might increase model performance in comparison to only relying on labeled data. This research methodology will give significant insight into how comparable prediction performance for real environments with only minimal amounts of labeled data can be, without the extra costs for supervised approaches. As already mentioned before, a positive effect of self-learning for NILM was already shown in some papers (Barsim and Yang, 2015; Miao et al., 2019), when using high sampling rates (1Hz or higher).

For both approaches (SSL and supervised) we will rely on the same deep learning architecture in form of a Convolutional Neural Network (CNN) to ensure comparability. The architecture (c.f., Figure 1) follows mainly the model introduced by Zhang et. al (2016) and is only slightly adapted by adding three Dropout layers in order to reduce overfitting. Overall, we rely on this architecture due to the following reasons. First, it shows a decent prediction accuracy when it comes to NILM tasks. As a robustness check, we compared the CNN model’s performance with a Hidden Markov Model (HMM) as well as a Recurrent Neural Network (RNN) using House 2. The CNN outperformed the other models for most of the appliances by slight margins. Second, and more importantly, the computational complexity is lower compared to models like HMM or RNN (Zhang et al., 2017). This is important, as neural networks have to be trained twice in a semi-supervised learning paradigm. This is also the reason why in the train-test setting, we are limiting the dataset to roughly the first three months of the year 2015. Thus, the training set consists of the first two months of the year, namely January and February, while the test set contains the first four weeks of March. Thus, we roughly achieve a train-test split of 80 to 20, which is in line with the literature. On a side note, we do not use a validation set as we are using the hyperparameters provided by Zhang et. al (2016) as the model was already trained for a NILM task, and it seems sufficient for usage in our case. To evaluate our model on the test set we are using the Mean Squared Error (MSE) to investigate how much of the appliance energy consumption is within a specific timeframe the most.

\[
MSE = \frac{1}{n} \sum_{t=1}^{n} (y^t - \hat{y}^t)^2
\]

where \(y^t\) is the true appliance value at time \(t\) and \(\hat{y}^t\) the predicted appliance value at time \(t\). Literature has shown that the MSE is an adequate metric to evaluate Deep Learning models applied to NILM under low sampling rate (Chang and Ho, 2019). As a robustness check, we also evaluated our results based on the Mean Absolute Error (MAE) as well as Root Mean Squared Error (RMSE) and did not obtain any significant differences.

To summarize, the objective of this paper is not to develop and contribute yet another high performance state of the art machine learning model. Instead, we want to investigate if SSL, specifically self-learning, can contribute towards NILM adoption under real-world conditions in a Smart Living environment and solve any of the aforementioned challenges. In our experiment we, therefore, want to show, whether the model performance increases (i.e. the MSE decreases) when including unlabeled data into the training process.
4 Results

In this section we will now present the results after following the research methodology explained in section 3. As discussed, we want to investigate especially the scenario with 32 seconds sampling rate and changing amounts of labeled data as it comes closest to the real-world application and is therefore important for NILM adoption. However, we still want to give a short explanation of the results observed with sampling rates of 8 seconds as well as 16 seconds, as well as transferability capabilities.

Various amounts of labeled data in a low sampling rate setting

In Figure 2, we illustrate the difference in performance of the various models on the defined test set on house 2 at a sampling rate of 32 seconds. The y-axis shows the MSE, while the x-axis shows the amount of labeled data used for model training. The dark blue line represents the MSE of the SSL approach given the amount of labeled data (in weeks), while the light blue line represents the first stage model (i.e. the supervised learning benchmark). Moreover, the dashed line represents the fully trained supervised CNN model on eight weeks of training data. The lower the MSE, the better the presented model performs. Concerning the first appliance, the fridge-freezer, we can see the expected behavior for steady-state appliances. Steady-state appliances are appliances that have a nearly constant operating state, therefore requiring only a couple signal cycles to describe their behaviour (Jimenez et al., 2015). The SSL approach (dark blue) outperforms the supervised approach (light blue) and exhibits less volatility. This behaviour might be explained due to the addition of unlabeled observations and therefore the growing amount of data in general. This can help to stabilize performance if unlabeled data is correctly pseudo-labeled by the first stage model. Moreover, the SSL model performance seems to perform comparable against the fully trained CNN model (dotted line) as more unlabeled observations are available over time, making a SSL model for this appliance only slightly less performant. Nevertheless, using only half the amount of labeled data for the SSL compared to the full CNN model (week 4), the MSE difference is only marginal. This indicates that SSL can already be an alternative for steady-state appliances if marginal losses in performance are irrelevant for the specific tasks like activity tracking.

![Figure 2](image-url)

*Figure 2. House 2: Comparison SSL and supervised model at 32 Second Sampling Rate for Fridge-Freezer, Kettle, Washing Machine and Television Site.*
The Kettle, as a single-state appliance (on/off state), shows a similar result. The performance gains through the unlabeled observations allow the SSL model even to outperform the fully labeled CNN. The reasons for this effect could be the following. As the Kettle is an appliance used by several occupants of the residence and house 2 exhibits four of them, the model will learn different usage patterns. As the overall amount of data for our experiment is relatively small, it might be possible that the labeled data in week 8 changed the general usage pattern of the Kettle as maybe an occupant changed his behavior. Nevertheless, we can show that adding unlabeled observations, even if a large set of labeled data is available, can strengthen model prediction in a low sampling rate environment. Moreover, adding pseudo-labelled data via SSL can help to increase the robustness of the model in general.

The appliances Washing Machine and Television Site show a more complex behavior. The washing machine is a typical multi-state appliance, therefore displaying signature load patterns for each possible operation state. This can make model training particularly difficult and results very volatile. This is exactly, what can be observed when looking at the bottom left chart in Figure 2. Both approaches exhibit a strong volatility when looking at week 4 and onwards. Especially, the SSL models compared to the supervised models show a stronger increase in MSE in week 5 and 6, displaying one of the caveats self-learning exhibits. This phenomenon appears because of the nature of self-learning and pseudo-labeling. We could interpret this as a pattern for a different appliance state emerging in the data at week 4, a state which was not observed beforehand. While the supervised model could react based on the fully labeled data and adjust accordingly at week 6, the SSL approach is unable to adopt the new behavior into its learned pattern (since it never saw it before), therefore taking longer to adjust before outperforming the supervised model again in week 7. As a consequence of this result, we argue that especially for multi-state appliances, SSL models in low rate environments might need more labeled training data to adjust to different observed states, even though unlabeled data helps to increase performance. For the Television Site, the result seems close to those of the Kettle, as it can be understood like a single-state appliance, showing a very stable profile in its load pattern. The SSL model outperforms the supervised model consistently and even surpasses the fully labeled CNN. However, in week seven, it shows again

![Figure 3. Showing SSL Robustness for different houses in the dataset.](image-url)
Semi-supervised Learning for NILM

a very volatile prediction performance. The same reasons stated for the washing machine and Kettle apply here, as more occupants’ means more usage patterns, even though the load pattern might stay the same.

To conduct robustness checks we changed the test period from four to six weeks and mostly achieved similar results and test performance patterns. To get even more robust results we trained models for the aforementioned appliances in all houses in the dataset, that had data concerning these appliances (c.f. Figure 3). In the next section we will thus draw conclusions from the overall observed model behaviour. First, SSL models mostly outperform the supervised benchmark independent of the appliance. This suggests that adding additional information by including pseudo-labeled data can increase performance slightly. However, one has to be careful whether to add unlabeled data or not, as it can also decrease model performance if it is not drawn from the same distribution as the labeled data. Second, in general, SSL model performance converges against or even outperforms to fully labeled supervised CNN model. Both arguments suggest that SSL can help to mitigate the necessity of large amounts of data and can help to tackle one of the core challenge. Third, as we are applying our models to a low sampling environment, we show evidence that SSL can be helpful when looking at a real-world environment. Overall, we can say that our results give a first indication that SSL can help to tackle two of the aforementioned challenges, namely a low sampling environment as well as the necessity of large amounts of labeled data. Nevertheless, as SSL results can be volatile, the experiment should be repeated by using larger amounts of data and different houses. Please note, we are neglecting the dishwasher here, as the test performance pattern is very similar to the washing machine, as both are multi-state appliances.

Varying sampling rate frequencies

Moreover, we conducted an experiment by changing the sampling rate as discussed above. Based on this, we observed the following. Overall, the SSL methods seem to outperform the supervised models for all tested appliances, which indicates that adding unlabeled data is mostly beneficial independent of the sampling rate. Additionally, the same convergence patterns to the performance of the fully trained CNN model can be observed. The main difference to other sampling rates is the following. Mostly, the MSE decreases slightly when looking at higher frequencies, which makes sense as adding more data is mostly beneficial in supervised learning and high performances in a higher sampling rate environment were already shown by related research (Barsim and Yang, 2015). Additionally, model performance was slightly less volatile in a high sampling rate environment compared to the low sampling rate setting. Overall, our results suggest that SSL models can be advantageous when conducting NILM in a low sampling rate environment. Nevertheless, having access to more frequent data can increase model performance as well as robustness.

![Figure 4](image_url)

Transferability of SSL models using House 19 (REFIT)
SSL for Transferability

Even though beyond the scope of this paper, we want to give a first indication of transferability for SSL models. (c.f. Figure 4) In this context, we tested the models trained on house 2 for the appliances Fridge-Freezer and Washing Machine on house 19 using again the REFIT dataset with a sampling rate of 32 seconds. Instead of using the first four weeks of March for the test set, we are now using four weeks of test data over the months January and February, as this was the training period of the models. Comparing the graph for Fridge-Freezer with our original measurements, the prediction performance for all models declines for both the self-supervised as well as the supervised approach. However, comparing the model performances to each other given the new test set, we can see almost the same behavior for the steady-state appliance Fridge-Freezer as in the house 2 test set. The SSL models perform better than the supervised models and even manage to outperform the fully trained CNN slightly. The same explanation given above holds true.

Given that steady-state appliances of the same type might have differences in the load pattern, a lot of the signatures might still look very similar, making learned patterns still applicable to some degree. For the appliance washing machine we actually see a performance improvement; while the SSL Model still shows the same kind of highly volatile predictions, the performance range is actually reduced. The performance improvement might be due to the fact that we are now using the same month for training and testing. Overall, our results give a first indication that SSL models can be beneficial when transferring a model to another environment. Nevertheless, the improvement seems to depend on the type of appliance investigated. Additionally, we have to acknowledge that the data is drawn from different distributions in both houses, making it even more difficult for a supervised algorithm to generalize. Thus, SSL is especially useful when unlabeled data from the same distribution is added to the model to increase its robustness further.

5 Discussion, Future Research and Limitations

AI in Smart Living Environments is about situational awareness and improved decision making based on context information for objects, occupants, and structural arrangements. However, creating and providing high-quality context data, especially for energy consumption on the level of individual appliances for advanced smart living applications, is not only costly but also often inefficient. The current energy disaggregation approaches in the industry, while also using machine learning apply highly upsampled data to train models, which then require environments to provide very high-frequency sampling rates and high quality labeled data for strong predictions. However, this also requires a significant upfront cost for applying NILM in every single household, making NILM adoption a complex challenge in the housing industry. In reality, most energy consumption data will be provided by smart meter in a comparably very low data sampling rate. In addition, while at least some short-term submetering, limited to the most energy significant appliances, in a setup phase might be realistic, longer-term submetering for all appliances in a household for full supervised model training is not. Utilizing energy disaggregation based on a heuristic approach with fewer data requirements by employing semi-supervised learning is not only a contribution to the field but also addresses multiple issues concerning NILM adoption, such as reducing the required labeled training data, and operating on lower frequency sampling rates, with still strong prediction performance.

Future Research

In general, to overcome the limitations shown above and to leverage our results, we suggest using SSL, especially in the context of Smart Living environments, as this combination could produce great value. Currently, NILM adoption in the housing industry is not economically practical, as each individual household would need costly data collection and model training. However, especially Smart Living
environments such as multi-household living complexes might benefit from NILM adoption, as each individual household has similar hardware installed. As an example, in many Smart Living Complexes, kitchens are already furnished and mostly include the same appliance models (i.e. exactly the same fridge and dishwasher), as the apartment layout is standardized. Moreover, the appliance usage and household size might be similar, allowing for similar observed usage patterns and thereby load signatures, together with the ability to standardize the data structures, which helps to address transferability. Thus, the housing industry could leverage on SSL by building the following pipeline. First, by monitoring one apartment of the Smart Living complex, smart meters, smart plugs, and other sensors, could create labeled data for each individual appliance's electricity consumption. After the training of an ML model based on this data, only the aggregated consumption data from other apartments gets added as more extensive unlabeled observations, which can then be pseudo-labeled by the same ML model trained before. Based on this approach for multi-household complexes, the retraining of ML models will use the labeled data from one household as well as the pseudo-labeled data from theoretically multiple other households. Using SSL (i.e. adding a pseudo-labeled apartment to the training set) can be a great advantage (as shown in Figure 2). Adding unlabeled observations is especially useful when drawn from the same distribution as the labeled data. The model could potentially be used to make predictions on all other apartments in the Smart Living environment with no additional sensors instalments needed. This approach could make the transferability of NILM models possible, as the appliances used are the same in each household. Thus, given enough training data in one apartment (i.e. various usage patterns), NILM adoption in a Smart Living environment would be feasible for many other apartments while SSL helps to increase the robustness of the used Deep Learning models.

Limitations
While the models in this paper already show promising results, additional investigations and robustness checks are needed, especially when applied in a Smart Living environment. This is due to the fact that we did not further investigate the SSL approach on a broader set of data and appliances, nor tested in an actual Smart Living environment, as we only used REFIT as a scientific data source. Future research work focusing on an SSL approach as described above would give a more conclusive answer for semi-supervised NILM and entail valuable insights. Moreover, our current approach focused only on a basic CNN, as CNNs had shown remarkable performance in other current NILM approaches. Nevertheless, one could use more advanced techniques like subtask gated networks (Shin, Rho, et al., 2019) or WaveNet models (Jiang et al., 2019) in order to increase the SSL performance even more. Moreover, different SSL techniques like co-training or graph-based methods could be applicable and should be tested. Future work may also explore approaches utilizing semi-supervised disaggregation with context information concerning appliance state, provided by the appliances themselves, enabling accurate disaggregation without requiring any extra submetering.

Acknowledgements
Research funding: this work was funded by the BMWi through the ForeSight Project. The authors state no conflict of interest.
References


