If You Measure It, Can You Improve It? Exploring The Value of Energy Disaggregation

Nipun Batra
Indraprastha Institute of Information Technology Delhi
nipunb@iiitd.ac.in

Amarjeet Singh
Indraprastha Institute of Information Technology Delhi
amarjeet@iiitd.ac.in

Kamin Whitehouse
University of Virginia
Charlottesville, Virginia
whitehouse@virginia.edu

ABSTRACT

Over the past few years, dozens of new techniques have been proposed for more accurate energy disaggregation techniques, but the jury is still out on whether these techniques can actually save energy and, if so, whether higher accuracy translates into higher energy savings. In this paper, we explore both of these questions. First, we develop new techniques that use disaggregated power data to provide actionable feedback to residential users. We evaluate these techniques using power traces from 240 homes and find that they can detect homes that need feedback with as much as 84% accuracy. Second, we evaluate whether existing energy disaggregation techniques provide power traces with sufficient fidelity to support the feedback techniques that we created and whether more accurate disaggregation results translate into more energy savings for the users. Results show that feedback accuracy is very low even while disaggregation accuracy is high. These results indicate a need to revisit the metrics by which disaggregation is evaluated.

1. INTRODUCTION

Over the past few years, dozens of new techniques have been proposed for more accurate energy disaggregation, but the jury is still out on whether these techniques can actually save energy and, if so, whether higher accuracy translates into higher energy savings. In this paper, we explore both of these questions.

Energy disaggregation is the process of estimating the power draw of individual electrical loads based on metering of their aggregate power. Research in this area is broadly motivated by the philosophy of Lord Kelvin: "If you can't measure it, you can't improve it," but is the opposite also true? Disaggregation techniques can be used to help users identify the major energy consumers in their home or business, and some users will pour over this data to achieve significant energy savings [2]. However, several studies have shown that just providing energy data does not necessarily translate to long-term energy savings. After the novelty wears off, users experience a *rebound effect* as unsustainable energy-saving actions unwind themselves and as users tire of sifting through too much data [1]. Other studies have shown more sustainable effects by providing more targeted feedback or recommendations for simple actions [8]. Can disaggregation techniques be used to generate the types of targeted, actionable feedback that would produce sustainable energy savings?

In this paper, we present the exploration of two research questions that are highly relevant to the research community interested in energy disaggregation. First, we explore whether disaggregated power data can be used to provide actionable feedback to residential users, and whether that feedback is likely to save energy. We focus on feedback about refrigerators and HVAC, because they contribute significantly to overall home energy consumption and are available in most homes. We develop a model that breaks the power trace of a refrigerator into three parts: baseline (when no one is using the fridge), defrost (energy consumption when the fridge is in defrost mode) and usage (energy consumption due to fridge usage). Then, we develop techniques to identify users with 1) much more energy due to fridge usage than the norm 2) much more energy due to defrost than the norm, or 3) fridges that are malfunctioning or misconfigured, even during baseline operation. We evaluate our model using a dataset with power traces from 95 refrigerators. Results indicate that our model can break down fridge usage into its three components with only 4% error. Additionally, the three types of feedback could help users save up to 23%, 25% and 26% of their fridge energy usage, respectively. These techniques provide targeted feedback with specific actions, e.g. fix or repair the fridge, and so we expect this energy savings to be sustainable. Similarly, we develop new techniques to differentiate homes with and without setback schedules on the HVAC system based on their HVAC power traces and outdoor weather patterns. This information can be used to give feedback to install a programmable thermostat. We evaluate these techniques with power traces from 58 homes and results indicate that our techniques can classify homes with 84% accuracy. Based on these results, we conclude that disaggregation does indeed have the potential to provide targeted, actionable feedback that could lead to sustainable energy savings.

Second, we explore whether existing energy disaggregation techniques provide power traces with sufficient fidelity to support the feedback techniques that we created, and whether

more accuracy disaggregation results translate into more energy savings for the users. To do this, we re-evaluate the feedback techniques above using power traces produced by disaggregation algorithms instead of those produced by direct submetering. We use three benchmark algorithms provided in an open source toolkit called NILMTK [5]. We verified that these algorithms and the parameters we use produce disaggregation accuracies comparable to or better than the best results published in the literature. Nonetheless, the feedback techniques that we developed become almost completely ineffective when using the disaggregated energy traces. In some cases, they failed to identify over 70% of the homes that should be getting feedback and falsely flagged 14% homes of additional homes that should not receive feedback.

To conclude, we discuss why feedback accuracy is low even while disaggregation accuracy is high: accurate *energy breakdown* feedback (i.e. "Your fridge accounts for 8% of your energy bill") can be given even if the power traces have many errors as long as those errors average out over time. However, more targeted and actionable feedback (i.e. "Your fridge is defrosting too often; fix the seal.") depends on specific features of the power traces. Our results indicate that the disaggregation community needs to revisit the metrics by which it measures progress. Part of this process will necessarily be to look through the lens of applications, including but not limited to the feedback techniques presented in this paper, to find the aspects of power traces that are most important. After all, "what you measure is what you get."

2. RELATED WORK

The field of NILM was found by George Hart in the early 1980s. His early works were motivated towards the development of low cost and easy to use methods for utilities to carry out residential appliance load research [13]. Existing load research methods during that time instrumented the individual branches and appliances. Three potential applications of NILM were proposed in the early works: i) controlling deferrable loads, and ii) providing detailed energy usage to the end user, and iii) identifying faulty appliances.

Between the early 1980s and late 2000s, the field had a steady progress. In the late 2000s, governments started rolling out smart meters and it was easier than ever before to collect data for evaluating NILM. Prior to smart meter rollouts, work in the field primarily consisted of pilot deployments. In 2011, the REDD [16] data set for NILM research was released. Its aim was to mimic the progress made in computer vision research by the availability of public data sets. Since then, the field has shown exponential growth[1]. More than 10 public data sets have been released in the past five years [5]. This exponential growth in the field has also seen a lot of differences in assumptions and evaluation metrics.

A subsection of the research has viewed the disaggregation problem from the perspective of correctly identifying the operational state of an appliance and thus evaluates NILM accuracy using metrics such as precision and recall on appliance states. Such work is often motivated by applications in activity recognition [11]. Other research often looks at the disaggregation problem as providing a fine-grained electricity bill, where the consumers can see how much each appliance costs them per month. Such work often uses metrics such as percentage of energy correctly identified. Such was the variety of metrics used in the literature that it became virtually impossible to compare two papers. To ease the comparison of NILM research, there have been recent efforts with an aim to standardise NILM metrics and provide benchmark algorithms [5, 6].

Recently, there has been an increased focus towards developing NILM applications related to providing energy feedback. In terms of the techniques and evaluation we propose in this paper, there are three works that relate well to ours. Chen et al. [7] did a study on 124 apartments from an apartment complex having same appliances and amenities, where they collected hourly appliance level energy consumption. They explain the variation in fridge energy across homes to be caused by behavioural differences. They estimate the energy savings possible if fridges older than 10 years are replaced by newer efficient fridges. Our work differentiates from their work by evaluating feedback models on disaggregated power traces. Since scaling appliance level metering remains a huge challenge, we believe that there is a lot of value in evaluating the feedback on disaggregated power traces. Further, we evaluate our feedback methods on a wide range of homes that have variable appliances and amenities, unlike the data set used by Chen et al.

Parson et al. [19] also target feedback on the value of shifting to a new fridge across 117 homes from the UK. Our work is similar to theirs as they also give feedback based on disaggregated power trace. A key differentiating factor between our approach and the work by Parson et al. and Chen et al. is that rather than dismissing a high energy consuming fridge as inefficient, our fridge model enables us to answer if high energy is due to high usage, or is the high usage simply due to higher fridge capacity. Importantly, our work proposes feedback methods which are more fine grained than providing feedback just based on appliance energy usage, which can be highly misleading. For instance, when comparing the summer HVAC usage of two homes in a colder and warmer climate, feedback based only on HVAC energy usage may indicate that the home in the warmer climate is doing worse. Instead, the energy feedback needs to consider the climate before providing feedback.

Barker et al. [3] make a case of emphasizing NILM applications over accuracy. Their evaluation deals with the "long" execution times associated with disaggregation using current NILM algorithms, which effectively rule out a host of real-time applications. Our work is in the same vein, but instead does an empirical evaluation of energy feedback methods in an offline fashion. We believe that even before we address the issue of real-time applications, we need to evaluate the accuracy associated with the intended applications. Our work also shows the efficacy of the proposed feedback methods on a large number of homes.

3. DATA SETS

[1] `http://blog.oliverparson.co.uk/2015/03/overview-of-nilm-field.html`

We now describe the two data sets that we will be using throughout the rest of this paper. To assess the value of energy disaggregation, we need a data set containing a large number of homes. We thus use the Dataport[2] data set, which is the largest publicly available dataset containing submetered and aggregate electricity consumption. The first release of the data set contains minutely power readings across different appliances from 240 homes in Austin, Texas from January through July 2014. More recently, a newer version of the data set has been released which contains data from 800 homes for close to 3 years. In addition to power data from different appliances, the data set contains information on energy audits, home surveys, internal temperature and water meter readings for a subset of homes. Since our fridge work predates the latest release, we use the first release made available in NILMTK [5] format consisting of data from 240 homes for our fridge analysis.

The data set contains minutely power data for 172 fridges. Of these, we filtered out 77 fridges that had data collection problems such as missing data and multiple appliances on the same sensor. We use the remaining 95 fridges for evaluation of our proposed techniques. The data set also contains temperature setpoint data from 2013. Since, the initial release does not have electricity data from 2013, we use the 2013 data from the newer release for our HVAC feedback analysis. We use the 58 homes having both the setpoint and power data information in our analysis.

We also collected data from four identical fridges operated in identical ambient conditions across four floors of the computer science building at UVa. We put Hobo loggers[3] to collect power data at 1 Hz frequency from these four fridges. For one of the fridge to which we had easy access to, we collected door status for both doors and the freezer unit and internal temperature data at 1 Hz frequency, in addition to the power data. We collected data under different controlled and uncontrolled settings for two weeks.

4. APPLIANCE ENERGY MODELLING

Having described the data sets that we use, we now discuss energy models for fridge and HVAC, both of which contribute significantly to overall home energy consumption and are available in most homes. These energy models serve as the basis for the energy feedback methods that we later describe in Section 5.

4.1 Fridge energy modelling

A fridge is a compressor based appliance where the motor duty cycles to maintain the fridge at a set temperature. When the compressor is ON, the refrigerant transfers heat from inside the fridge to the outside [9]. The compressor turns ON and OFF at a small offset temperature above and below the set temperature. Since the fridge is operated at a lower temperature than the surroundings, there is always heat leakage from the outside into the inside of the fridge, which is proportional to the temperature difference between the fridge setpoint and ambient temperature.

In the absence of fridge usage (such as opening fridge door), the compressor typically duty cycles at the same rate, shown as the **baseline** compressor usage in Figure 1 which occurs in the early morning hours of the shown fridge. Each time the fridge is opened, the leakage from the ambient environment increases and the compressor has to run longer to remove this extra heat. The addition of items in the fridge also causes the compressor to run longer due to the increased thermal mass. Both these factors cause an increase in the duty percentage of the fridge. The increased compressor ON and decreased compressor OFF durations are shown as **usage** in Figure 1. For efficient running of the fridge, fridges defrost periodically to get rid of frost developed on the cooling coil. **Defrosting** is done via the defrost heater that is a separate circuit from the compressor circuit, which is cut off during the defrost. Since defrosting introduces heat into the system, the next few compressor cycles have higher duty percentage to remove the additional heat introduced into the system. These cycles can be seen in Figure 1.

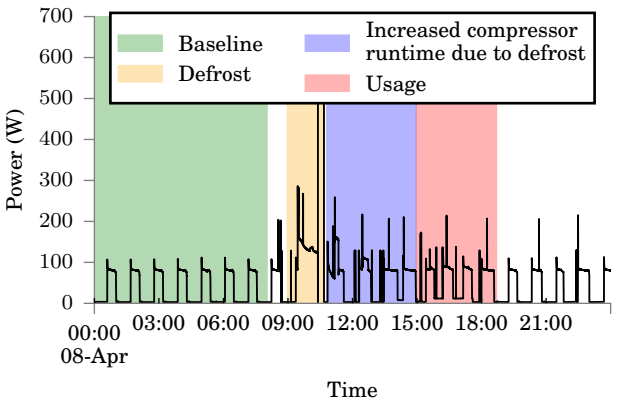

Figure 1: Breakdown of fridge energy consumption into baseline, defrost and usage

Thus, the fridge energy consumption can be broken down into three components: usage, defrost and baseline. We now describe the procedure for breaking down fridge energy into these three components:

1. Finding baseline duty percentage: Duty percentage of a fridge cycle is given by the ratio of the compressor ON duration to the total fridge cycle. Or,

$$\text{Duty percentage (c)} = \frac{\text{ON duration(c)}}{\text{ON duration(c)+OFF duration(c)}}$$

Baseline duty percentage is found as the median of the duty percentage during early morning hours (1 to 5 AM) over the duration of the dataset. Using median overcomes the cases when a home may have high fridge usage on some days.

2. Finding defrost energy: Defrost energy comprises of two parts: energy consumption when the fridge is in the defrost state and the extra energy consumed in the regular compressor cycles that follow the defrost state. We assume that a defrost cycle causes an impact on the next D compressor cycles. For these D cycles, the extra energy consumed is found as:

[2] https://dataport.pecanstreet.org
[3] http://www.onsetcomp.com

Extra compressor energy due to defrost

$$= \sum_{c=1}^{D} (\text{Duty percentage (c) - Baseline duty percentage}) \times (\text{ON duration(c) + OFF duration(c)}) \times \text{Fridge compressor power consumption} \quad (1)$$

Energy consumption when fridge is in the defrost state can be trivially calculated.

3. Finding usage energy: As a prerequisite to finding usage energy, we need to first find *usage cycles*, which we define as fridge cycles that are affected by fridge usage. After removing the defrost cycles and the subsequent D cycles, we look for cycles having duty percentage that is P% more than the baseline duty percentage. The intuition behind choosing a parameter P is that fridges may show some inherent variation in duty cycle percentage independent of usage. We assume that this variation is within P% of the baseline duty percentage. After finding these U usage cycles, the usage energy can be calculated as: Usage energy

$$= \sum_{c=1}^{U} (\text{Duty percentage (c) - Baseline duty percentage}) \times (\text{ON duration(c) + OFF duration(c)}) \times \text{Fridge compressor power consumption} \quad (2)$$

4. Finding baseline energy: All the cycles that are not affected due to defrost or usage contribute towards baseline energy and their energy consumption can be summed to find baseline energy.

4.1.1 *Evaluation of fridge model*

We now evaluate the accuracy of our fridge modelling approach. We use our collected data from the UVa CS building for this evaluation as the Dataport data set does not have labels for fridge usage. Using door sensor data, we manually annotated 3 days for usage cycles from the fridge for which we had instrumented in our data set. We found that the defrost cycle impacts the next 3 cycles, and we thus chose D=3. It should be noted that choosing a slightly different value of D is only going to change marginally the usage and defrost energy numbers since defrost cycles are easily outnumbered by regular cycles. The other parameter in our evaluation, percentage threshold (P) for labelling usage cycles is more important due to the expected high number of usage cycles.

We now define the three metrics used to evaluate our fridge modelling:

1. % Usage energy error for fridge: $\frac{|\text{Predicted fridge usage energy - Actual fridge usage energy}| \times 100\%}{\text{Actual fridge usage energy}}$
2. Precision on fridge usage cycles: $\frac{|\text{Correctly predicted fridge usage cycles}|}{\text{\# Predicted fridge usage cycles}}$
3. Recall on fridge usage cycles: $\frac{|\text{Correctly predicted fridge usage cycles}|}{\text{\# Total fridge usage cycles}}$

Figure 2 shows the usage energy error, precision and recall on usage cycles as they vary with P. At a P of 11-16%, the usage energy error is less than 2%. Usage energy error remains below 4% for P between 9 and 24, showing that the prediction remains useful within a wide percentage threshold. A precision of 1 is not observed until P = 17% due to the presence of a single fridge cycle having a high duty percentage despite being unrelated to usage. This is due to the fact that rare cycles may show an inherent deviation from the regular duty percentage. At P = 11%, the recall drops from 1. This is due to a usage cycle which shows less than 10% deviation from baseline duty percentage.

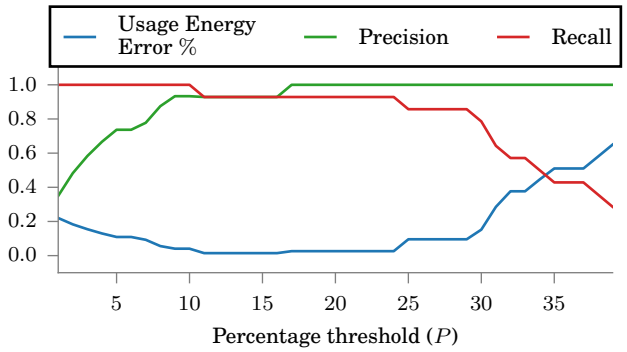

Figure 2: Our model for breaking fridge energy into usage, baseline and defrost is accurate to within 4% energy error for a wide range of percentage threshold above baseline duty percentage.

4.2 HVAC modelling approach

Across the globe, HVAC is the single largest contributor to a home's energy bill [20]. By optimising the HVAC setpoint schedule, upto 30% of HVAC energy can be saved [17]. Giving homes feedback on their setpoint schedule is likely to have a big impact. Thus, we try to build an HVAC model that can help us predict setpoint temperature from HVAC energy data. Since HVAC energy usage is highly dependent on external weather conditions, we incorporate weather data into our HVAC model. While we explain our model for the cooling season (summers, when HVAC is used for cooling), it is equally applicable to the heating season. Our model is based on the following assumptions:

1. HVAC energy is impacted by weather conditions such as humidity, wind speed and temperature.
2. HVAC energy consumption is proportional to the difference in external temperature and home setpoint temperature.
3. Programmable thermostats use the following four setpoint times: night hours from 10 PM to 6 AM; morning hours from 6 AM to 8 AM; work hours from 8 AM to 6 PM; evening hours from 6 PM to 10 PM [10].
4. HVAC energy during an hour is zero if the HVAC was not used during this hour

Based on the first assumption, we have: HVAC energy $\propto$ humidity; HVAC energy $\propto$ wind speed. Based on the second assumption, we have HVAC energy $\propto$ (External temperature-internal temperature setpoint). Based on the third assumption, we have four different temperature setpoints during the day. We use four proportionality constants (a_1 through a_4) corresponding to these four setpoint times, describing how strongly the temperature delta between external and setpoint temperature affects HVAC energy consumption. To

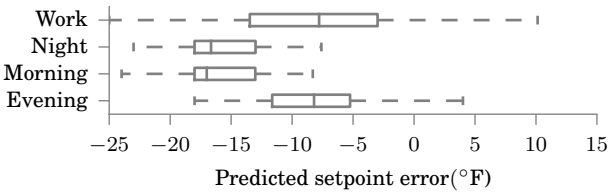

Figure 3: The predicted setpoint temperatures from our HVAC model have a high offset from actual setpoint temperatures.

convert our HVAC model into a regression model, we add a binary variable (is it n^{th} hour) which is 1 if the data is from the $n^{th} hour$ and 0 otherwise. We also use a binary variable indicating if HVAC was used during the n^{th} hour based on the fourth assumption. Combining all of the above, our HVAC models energy consumed in the n^{th} hour of the day as follows:

$$\begin{aligned} HVAC\,energy(n) = a_1 \times [&(External\,temperature(n) \\ &- Night\,hours\,setpoint) \\ &\times Is\,it\,0^{th} hour \times Is\,HVAC\,used(n) \\ &+ \dots \\ &(External\,temperature \\ &- Night\,hours\,setpoint) \\ &\times Is\,it\,5^{th} hour \\ &\times Is\,HVAC\,used\,this\,hour] \\ &+ a_2 \times \dots \\ &+ a_3 \times \dots \\ &+ a_4 \times \dots \\ &+ a_5 \times humidity(n) \\ &+ a_6 \times wind\,speed(n) \end{aligned} \tag{3}$$

Our model has a total of 10 parameters: a_1 through a_6 and four setpoint temperatures.

4.2.1 Evaluation of HVAC model

We now evaluate our HVAC model on its ability to learn the temperature setpoints. We calculate hourly HVAC energy usage for the 58 homes containing both HVAC power and setpoint information. This forms the LHS of Equation 3. We download hourly weather data from Forecast.io web service[4] and use linear interpolation to fill missing readings, similar to the work done by Rogers et al. [21]. Finally, we used non-linear least squares minimisation using the Python lmfit to estimate the 10 parameters in our model. We also constrain learnt setpoints to be within 60 and 90F.

Figure 3 shows that our model is inadequate in accurately predicting setpoint temperatures. This is most likely due to the fact that some of the coefficients in our model are not independent and the fact that our model does consider thermal mass of the building. Our main objective is finding homes which need HVAC setpoint feedback. While an accurate prediction of setpoint temperature would have allowed us to do the same, in section 5.4, we explore machine learning based solutions to use the parameters from our HVAC model to predict homes needing setpoint feedback.

[4] `http://forecast.io`

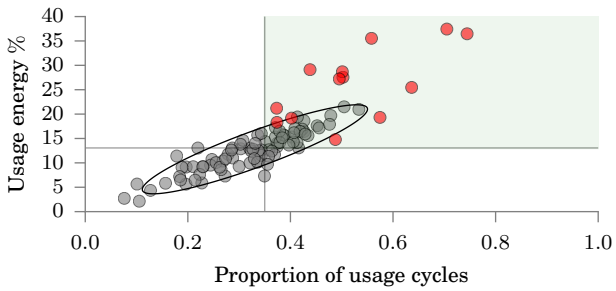

Figure 4: 13 out of 95 homes (shown in red) from the Dataport data set can be given feedback based on their fridge usage, potentially saving up to 23% of fridge energy.

5. ENERGY FEEDBACK METHODS

In this section, we develop and demonstrate some examples of how NILM could be used to provide feedback to users to reduce their energy usage based on the appliance energy modelling we previously discussed. These are only examples, and the analysis presented later in this paper would apply to any applications of NILM.

5.1 Fridge usage feedback

Having shown that we can accurately breakdown fridge energy into usage, defrost and baseline, we now show how we can give feedback to homes based on this breakdown. In this section, we target homes based on fridge usage. We use robust estimator of covariance based outlier detection [12] to detect such homes. The outlier detection method is applied on two dimensions: usage energy% and proportion of usage cycles. We apply this outlier detection method on the 95 homes from the Dataport data set. We divide this two dimensional home data into four quadrants through the medians on usage energy% and proportion of usage cycles. Figure 4 shows the homes that can be given feedback based on their fridge usage energy in red. The black ellipse is the boundary outside which points are predicted to be outliers. Feedback can be given to homes in the first quadrant (shown in green), that have a high proportion of usage cycles and high usage energy. Homes in this category have a lot of cycles affected by usage and thus have high usage energy. 13 homes fall into this category and can save up to 23% of their fridge usage energy. Energy saving potential is calculated as the difference between current energy consumption and median energy consumption. There are no homes in the second quadrant, which denotes homes which have a small proportion of cycles affected by usage and yet having a high usage energy contribution. These homes could possibly have few interactions with the fridge, but, have a high usage energy due to a low fridge internal setpoint, where each interaction with the fridge leads to a lot of heat flow from the outside.

5.2 Fridge defrost feedback

Our method for providing feedback based on defrost is similar to the method of providing feedback based on usage. We use outlier detection methods on two dimensions: defrost

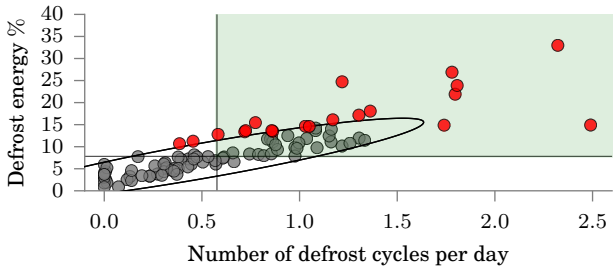

Figure 5: 17 out of 95 homes (shown in red) from the Dataport data set can be given feedback based on their fridge defrost energy, potentially saving up to 25% of fridge energy.

energy% and number of defrost cycles per day and give feedback to the homes lying in the first and the second quadrant. Number of defrost cycles per day is more interpretable and relatable than proportion of defrost cycles (which is going to be a very small floating point number). Figure 5 shows the homes that can be given feedback based on their fridge defrost energy. 15 out of 95 homes fall into the first quadrant, and 2 homes fall into the second quadrant. These 17 homes can save up to 25% of their fridge energy. While homes in the first quadrant have high defrost energy due to high number of defrost cycles, homes in the second quadrant are likely to have a fridge malfunction whereby a fridge remains in the defrost state for a long time.

5.3 Fridge power feedback

We next looked into providing feedback in case we know the make and age of a fridge, and we have data from fridges of the identical make and age. We found four such pairs in the Dataport data set (LG, Frigidaire and two of Samsung) where one of them has a significantly higher fridge steady state and transient power. Transient power is defined as the short duration power when the fridge compressor motor starts. This power is higher than the steady state power, which is defined as the power draw of the fridge once the transient has ended. Figure 6 shows these four fridges and the differences in their steady state and transient powers. In order to eliminate the hypothesis that such differences could arise due to the difference in ambient conditions of these fridges, we also add in this figure the four General Electric fridges from our deployment. 3 of them have a <steady state, transient> power consumption of <80,100> Watts, while the fourth one has <120, 1310> Watts. Since these four fridges were operated under identical ambient conditions, the possibility of ambient conditions causing a power difference between these is ruled out. The arrows in the figure point towards the fridge consuming extra power. These fridges consume upto 26% more energy than their identical counterparts, where extra energy consumption is found by estimating the energy consumption if the fridge operated with lower steady state power. In order to reduce the false positive rate in giving such feedback, we can choose to give feedback when the difference in steady state power is atleast 10%, where we assume that fridges can record upto 10% variation in their power consumption owing to several factors including measurement errors. Thus, LG lfx25960st and Samsung rfg297aars fridge won't be given this feedback.

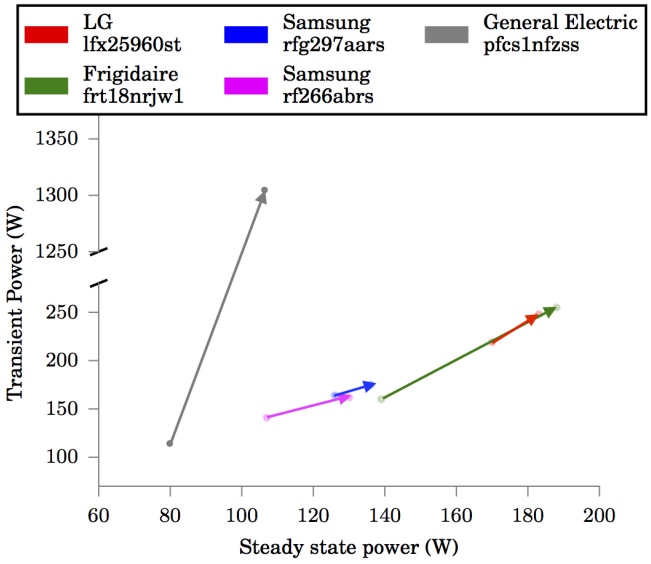

Figure 6: Identical fridges with the same model and age can have differences of 10% or more in steady state power levels. Feedback about failing or misconfigured fridges can save up to 26% energy.

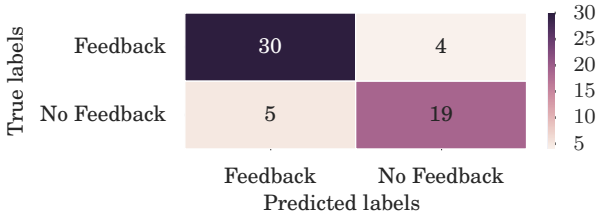

Figure 7: Our techniques correctly classify 84.4% of the homes as either having or not having a setpoint schedule, based on submetered HVAC data.

5.4 HVAC setpoint feedback

We previously saw in section 5.4, that our HVAC model produces an offset in the learnt setpoint temperatures. Instead of using the learnt setpoint temperatures directly to find homes needing HVAC setpoint feedback, we use machine learning methods for the same. We calculate an HVAC efficiency score for the 58 homes in the Dataport data set on a scale of 0 to 4 based on recommended setpoint temperature from EnergyStar [10] as follows: 1)Morning score = 1 if morning setpoint temperature >78F, 0 otherwise; 2) Evening score = 1 if evening setpoint temperature > 78F, 0 otherwise; 3) Work hours score = 1 if work hours setpoint >85 F, 0 if setpoint <=78, (85-setpoint)/7 otherwise; and 4) Night score = 1 if setpoint >82F, 0 if setpoint <=78F, (82-setpoint)/4 otherwise. We decide that 34 homes that have an overall score of 2 or less can be given feedback to optimise their HVAC setpoints.

In addition to the 10 parameters of the HVAC model, we add additional features such as total energy used in work, morning, night and evening hours and the number of minutes HVAC system was on during these times to our machine learning methods We use 2-fold cross validation and a grid search on the feature space to find that the feature $<a_1$,

Authors	Year	Dataset	#Homes	Algorithm	Fridge			HVAC		
					RMSE (W)	Error Energy %	F-score	RMSE (W)	Error Energy%	F-score
Kolter [15]	2012	REDD	6	Additive FHMM	-	62.5 $^{\Delta}$	-	-	-	-
Parson [18]	2012	REDD	6	Difference HMM	83	55	-	-	-	-
Parson [19]	2014	Colden$^{\Psi}$	117	Bayesian HMM		45				
Batra [5]	2014	iAWE	1	FHMM	-	50	**0.8**	-	30	**0.9**
Current work		Data port	240	CO*	85	**19**	0.65	**600**	**15**	0.87
Current work		Data port	240	FHMM*	95	20	0.63	650	18	0.89
Current work		Data port	240	Hart	**82**	21	0.72	890	23	0.76

Table 1: Benchmark algorithms on the Dataport dataset give comparable performance to existing literature.
* Both CO and FHMM achieve best performance for N=2, top-K=3.
$^{\Delta}$ Kolter's paper includes a slightly different metric from which we derived this number.
$^{\Psi}$ Colden data set is not publicly available.

a_3, Energy in evening hours, Mins HVAC usage in morning hours> used by the Random Forest classifier give the optimal accuracy of 84.4% as shown in Figure 7.

6. EVALUATION OF NILM FOR FEEDBACK

Having described our methods for providing energy feedback to homes based on submetered data and showing that these models can give good feedback, we now evaluate how accurately do current NILM approaches match these feedback. We now describe the experimental setup for evaluating NILM performance on the Dataport data set.

6.1 Experimental setup

We use NILMTK [5] to perform our NILM experiments. We use the 3 reference implementations made available in NILMTK, which we describe now.

6.1.1 NILM models

Combinatorial optimisation (CO): CO was proposed by George Hart in his seminal NILM paper [13]. CO models each appliance to consist of a fixed number of states and assigns different power levels to each of these states. The optimisation function involves finding the optimal combination of appliance states for different appliances which minimises the difference between predicted and observed aggregate power.

$$\hat{x}_t^{(n)} = \underset{\hat{x}_t^{(n)}}{\operatorname{argmin}} \left| \bar{y}_t - \sum_{n=1}^{N} \hat{y}_t^{(n)} \right| \quad (4)$$

Here, $\hat{x}_t^{(n)}$ is the state of the n^{th} appliance at time t;$\bar{y}_t$ is the aggregate power consumption at time t and $\hat{y}_t^{(n)}$ is the power consumption of n^{th} appliance at time t. The time complexity of CO is exponential in the number of appliances and hence it becomes intractable for large number of appliances.

Factorial hidden Markov model (FHMM) :FHMM approach is more recent and built upon the finite state machines suggested by Hart [13]. In an FHMM, each appliance is modelled as a hidden Markov model (HMM), where the hidden component is its state, and the observed component is its power draw. Like CO, FHMM has time complexity exponential in the number of appliances and thus become intractable for a large number of appliances.

Hart's steady-state algorithm: Hart in his seminal work presented an event based approach that we here on refer to as Hart's algorithm [13]. This approach finds events in the aggregate power time series and assigns them to appliances. An event is said to occur when the aggregate power changes beyond a threshold. During the training phase, Hart's algorithm pairs rising and falling edges whose magnitude difference fall within a threshold. Next, it clusters these rising-falling pairs where each cluster represents an appliance. Since this algorithm is unsupervised in nature, it requires manual labelling to assign appliance labels to these clusters.

6.1.2 NILM metrics

Having discussed the NILM models, we now discuss the conventional NILM metrics to evaluate these. We use the standard definition of NILM metrics as made available in NILMTK [5].

1. % Error in Energy: $\frac{|\text{Predicted energy - Actual energy}|\times 100\%}{\text{Actual energy}}$
2. Root Mean Squared Error (RMSE) Power: $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{Predicted power}_i - \text{Actual power}_i)^2}$
3. F-score: First, disaggregation is converted to a binary classification problem where an appliance is ON if it consumes more than a threshold and OFF otherwise. Next, the standard definition of F-score is used on this binary classification task.

6.1.3 Parameter optimisation and training strategy

Having discussed the metrics used for evaluating NILM performance, we now discuss the tunable parameters in these NILM models. Since both CO and FHMM are computationally intractable, NILM researchers often select the top-K appliances in terms of energy consumption to reduce the state space. Another parameter in these models is the number of states (N) to use for modelling an appliance (2 states means that an appliance can either be ON or OFF). We vary K from 3 to 6 and N from 2 to 4 and find the accuracy of disaggregation for both fridge and HVAC. We used half of the data for training and the other half for evaluating disaggregation.

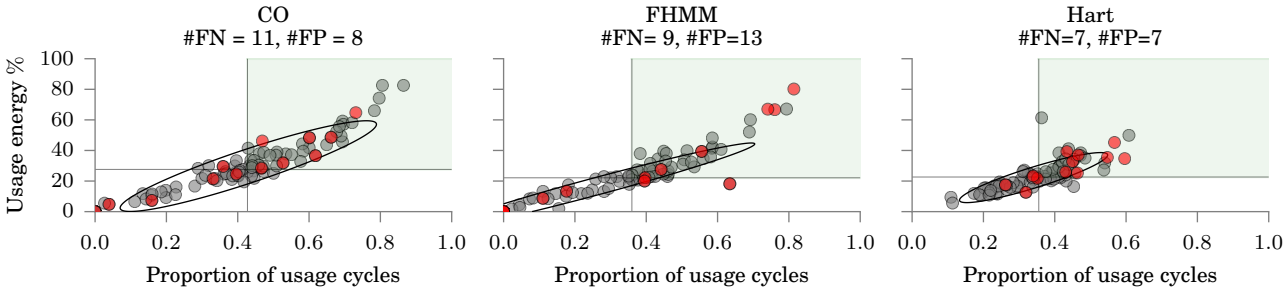

Figure 8: NILM algorithms show poor accuracy in identifying homes which need feedback for high fridge usage energy. Red dots indicate the homes which should be getting feedback based on analysis of submetered fridge data, while these algorithms would give feedback to all homes in the green region outside the elliptical boundary.

6.1.4 NILM accuracy

We now present the results of NILM evaluation on the Dataport data set. We also compare our results with the state of the art. From Table 1, we can see that for both fridge and HVAC, the benchmark algorithms we use are comparable in performance to existing literature. We could not include several recent works due to different reasons. Shao et al. [22] and Kim et al. [14] define precision and recall in terms of identification of appliance power within bounds. It is non-trivial to convert their metrics in terms of ours. Barker et al. [4] show that the performance of their tracking algorithm is comparable to Additive FHMM, which we already consider in our comparison. Kolter et al. [16] do not provide appliance level metrics. Since none of the above-mentioned works gave results on HVAC disaggregation under residential settings, we used the numbers given in the benchmark evaluation accompanying NILMTK [5]. It should be noted that many of the other approaches we compare with in Table 1 make lesser assumptions such as the availability of training data. However, these do not affect our argument since they do not achieve substantially better performance according to conventional NILM metrics.

6.2 Fridge usage feedback

Having established that our NILM performance is at par with the state-of-the-art, we now see how accurate fridge usage feedback we can provide with the disaggregated power trace. Figure 8 shows that all three NILM algorithms have poor accuracy in identifying homes that need feedback for high fridge usage. False negatives (FN) are those homes that should be getting feedback but are not getting, and false positives (FP) are those homes that would wrongly get feedback. We now explain the reasons for the poor accuracy of the used NILM algorithms.

During the night hours when typically only background appliances such as fridge are running, Hart's algorithm has good disaggregation accuracy. Due to this, Hart's algorithm closely matches the baseline duty percentage computed on submetered data as shown in Figure 9. However, Hart's algorithm is susceptible to detection of false events and missing true events, especially during active hours when appliances similar in magnitude to the fridge may be operating. Thus, Hart's algorithms underpredicts and overpredicts fridge compressor cycle durations during the day creating a deviation in fridge usage. While the change in predicted

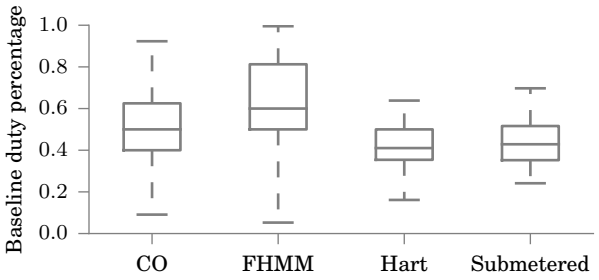

Figure 9: The baseline duty percentage found on Hart's disaggregated power traces matches closely to the submetered one, while CO and FHMM show a wide variation from submetered.

cycle durations has a minimal impact on conventional metrics, it has a significant impact on fridge usage energy metric. The median baseline duty percentage found by CO and FHMM are higher than the median baseline duty percentage on submetered data. Owing to higher baseline duty percentage, usage energy in these homes is lower than submetered, thereby explaining the high false negative rate. The reason behind CO and FHMM finding a high baseline duty percentage is that the objective function in both these algorithms includes minimising the difference between aggregate power and sum of power for predicted appliances. To satisfy this objective, these algorithms predict fridge to be ON longer than actual during the night hours when typically few loads are used. The high false positive rate can be explained by the small number of homes for which the baseline duty percentage learnt is much lower than that for submetered. This causes these homes to have a high usage energy, and thus predicted as candidates to give feedback.

6.3 Fridge defrost feedback

We find that the our approach of breaking down fridge energy into baseline, defrost and usage is unable to find even a single defrost cycle when fed the disaggregated power data. This is due to the inadequacy of the used NILM methods in effectively learning and disaggregating the defrost state. CO and FHMM rely on KMeans and Expectation Maximisation algorithms respectively for learning the different states of an appliance. Due to defrost events being rare in comparison to regular usage, these algorithms are not able to accurately

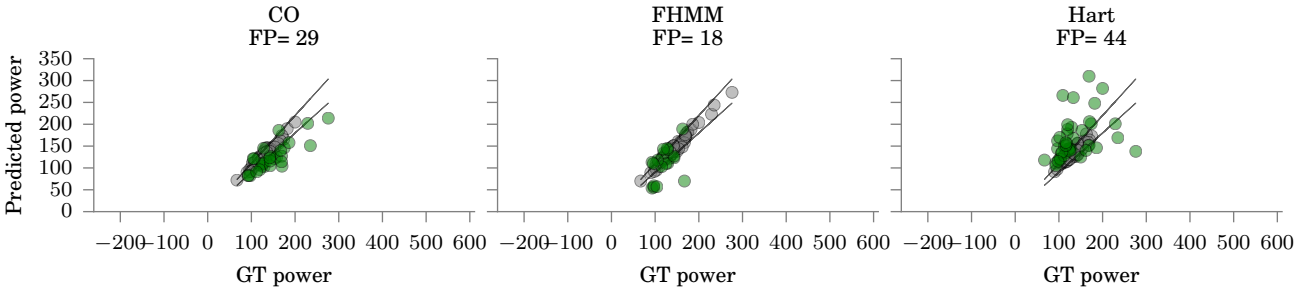

Figure 10: All NILM algorithms estimated the steady state power levels of at least some fridges (shown in green) with errors over 10%, which means that estimates are not accurate enough to reliably detect malfunctioning fridges based on power draw.

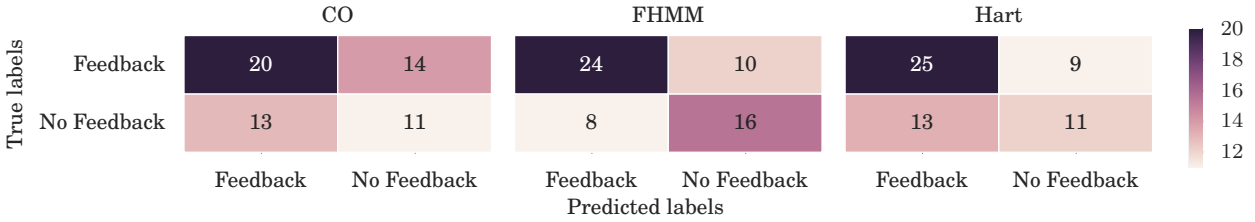

Figure 11: Classification of homes into those with setback schedules decreases from 84% with submetered power traces to 53%, 69%, and 62% respectively with power traces produced by the three NILM algorithms.

associate a cluster with the defrost state. Instead, these algorithms try to find multiple clusters to explain the variation in fridge power when the compressor is ON. Hart's algorithm, which relies on pairing rising and falling edges of similar magnitude in the power signal, is unable to learn the defrost state as the defrost state has a significantly different magnitude of rising and falling edge.

6.4 Fridge power feedback

We now show the efficacy of feedback based on fridge power given NILM power traces. Since there were only 4 homes in the dataset having a corresponding fridge of same make and age, we evaluate this feedback assuming that for each fridge in the data set we had a corresponding identical fridge. For the identical fridge, we use the actual steady state power as its learnt steady state power. Ideally, none of these 95 fridges should be getting feedback based on fridge power. Figure 10 shows that out of these 95 pairs, NILM algorithms produce a high number of false positives due to estimating the steady state power levels with errors over 10%.

Hart's algorithm learns higher than actual steady state power for a large number of fridges. This can be explained by its clustering strategy during the learning stage where pairs of rising-falling edges are clustered. Clustering is susceptible to learning fewer clusters than actual appliances, and thus some of the learnt clusters could span multiple appliances.

For CO and FHMM, the high number of false positives can be explained by the fact that using N=2 states may be optimal for NILM metrics, but is suboptimal for learning fridge steady state power. For N=3, the number of false positives reduces to 17 and 5 respectively for CO and FHMM. Within CO and FHMM, the better performance of FHMM can be attributed to it modelling time relationships between states. Thus, it is more robust to assigning clusters to power values that don't correspond to an actual fridge state, in comparison to CO.

6.5 HVAC setpoint feedback

We now evaluate the efficacy of HVAC feedback based on disaggregated power traces. Figure 11 shows that the classification of homes into those with setback schedules decreases significantly for all NILM algorithms. We now explain the low classification accuracy based on the features used by Random Forest classifier. Of the four features used, a_1 and a_3 are hard to interpret, and thus we provide an explanation based on *Mins HVAC usage during morning hours*. Most of the HVAC usage in the data set occurs during the night hours. Thus, NILM accuracy is likely to be highly dependent on night time HVAC disaggregation. Since, only HVAC and fridge would be typically used in the night, and, HVAC has a distinct much higher power signature than the fridge, NILM accuracy for HVAC is decent (as per Table 1). However, during the morning hours, when typically there is more activity in the home, NILM accuracy for HVAC is expected to be lesser. In Figure 12, we compare the error in prediction of minutes of HVAC usage for different algorithms when compared to submetered. It can be seen that for all algorithms, accuracy is higher in the night. Thus, despite not having a high impact on NILM accuracy, the high error prediction of minutes of HVAC usage affects our classification accuracy.

7. CONCLUSIONS

In this paper, we show that disaggregated power data has the potential to provide targeted, actionable energy feedback to homes. However, we found that the state-of-the-art NILM accuracy isn't effective in enabling such feedback. We believe that the community needs to revisit the metrics for gauging NILM performance, and our work is a step in that

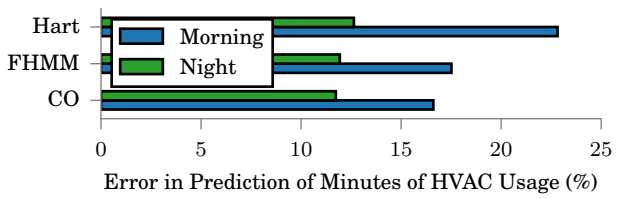

Figure 12: NILM algorithm have high accuracy overall, but have higher error in the morning because other appliances are being used. However, the morning hours are critical to inferring whether a home has a setback schedule.

direction. We finally conclude that- "If you can measure it, you may not necessarily be able to improve it."

8. LIMITATIONS AND FUTURE WORK

Our HVAC energy model is far from perfect. In fact, if it were perfect in predicting HVAC temperature setpoints, we would not need to train a classifier on top. However, it must be noted that our work is tangential to such HVAC modelling and can build upon better HVAC models to provide feedback. In the future, we would like to improve our HVAC model with an aim of more accurate feedback. We plan to develop energy feedback models for some of the other appliances having a high energy impact, such as water heater. We also plan to use long term trends in finding homes needing feedback. For instance, over time a fridge may develop an anomaly causing excessive energy usage.

9. REFERENCES

[1] W. Abrahamse, L. Steg, C. Vlek, and T. Rothengatter. A review of intervention studies aimed at household energy conservation. *Journal of environmental psychology*, 25(3):273–291, 2005.

[2] K. C. Armel, A. Gupta, G. Shrimali, and A. Albert. Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy Policy*, 52:213–234, 2013.

[3] S. Barker, S. Kalra, D. Irwin, and P. Shenoy. Nilm redux: The case for emphasizing applications over accuracy. In *NILM-2014 Workshop*, 2014.

[4] S. Barker, S. Kalra, D. Irwin, and P. Shenoy. Powerplay: creating virtual power meters through online load tracking. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, pages 60–69. ACM, 2014.

[5] 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. In *Proceedings of the 5th international conference on Future energy systems*, pages 265–276. ACM, 2014.

[6] C. Beckel, W. Kleiminger, R. Cicchetti, T. Staake, and S. Santini. The eco data set and the performance of non-intrusive load monitoring algorithms. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*. ACM, 2014.

[7] V. L. Chen, M. A. Delmas, W. J. Kaiser, and S. L. Locke. What can we learn from high-frequency appliance-level energy metering? results from a field experiment. *Energy Policy*, 77:164–175, 2015.

[8] S. Darby. The effectiveness of feedback on energy consumption. *A Review for DEFRA of the Literature on Metering, Billing and direct Displays*, 2006.

[9] R. J. Dossat and T. J. Horan. *Principles of refrigeration*, volume 3. Wiley, 1961.

[10] EnergyStar.gov. Programmable thermostats for consumers, 2015.

[11] S. Gupta, M. S. Reynolds, and S. N. Patel. Electrisense: single-point sensing using emi for electrical event detection and classification in the home. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 139–148. ACM, 2010.

[12] W. J. D. Haan and A. T. Levin. A practitioner's guide to robust covariance matrix estimation, 1996.

[13] G. W. Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, 1992.

[14] H. Kim, M. Marwah, M. F. Arlitt, G. Lyon, and J. Han. Unsupervised disaggregation of low frequency power measurements. In *SDM*, volume 11, pages 747–758. SIAM, 2011.

[15] J. Z. Kolter and T. Jaakkola. Approximate Inference in Additive Factorial HMMs with Application to Energy Disaggregation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 1472–1482, La Palma, Canary Islands, 2012.

[16] J. Z. Kolter and M. J. Johnson. REDD: A public data set for energy disaggregation research. In *Proceedings of 1st KDD Workshop on Data Mining Applications in Sustainability*, San Diego, CA, USA, 2011.

[17] J. Lu, T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field, and K. Whitehouse. The smart thermostat: using occupancy sensors to save energy in homes. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2010.

[18] O. Parson, S. Ghosh, M. Weal, and A. Rogers. Non-intrusive load monitoring using prior models of general appliance types. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 356–362, Toronto, ON, Canada, 2012.

[19] O. Parson, S. Ghosh, M. Weal, and A. Rogers. An unsupervised training method for non-intrusive appliance load monitoring. *Artificial Intelligence*, 217:1–19, 2014.

[20] L. Pérez-Lombard, J. Ortiz, and C. Pout. A review on buildings energy consumption information. *Energy and buildings*, 40(3):394–398, 2008.

[21] A. Rogers, S. Ghosh, R. Wilcock, and N. R. Jennings. A scalable low-cost solution to provide personalised home heating advice to households. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pages 1–8. ACM, 2013.

[22] H. Shao, M. Marwah, and N. Ramakrishnan. A temporal motif mining approach to unsupervised energy disaggregation. In *Proceedings of the 1st International Workshop on Non-Intrusive Load Monitoring, Pittsburgh, PA, USA*, volume 7, 2012.