# Dataset: Device Activity Report with Complete Knowledge (DARCK) for NILM

Justus Breyer
Chair of Communication and Distributed Systems
RWTH Aachen University
Aachen, Germany
breyer@comsys.rwth-aachen.de

Kai Gützlaff
Chair of Communication and Distributed Systems
RWTH Aachen University
Aachen, Germany
kai.guetzlaff@rwth-aachen.de

Leonardo Pompe
Chair of Communication and Distributed Systems
RWTH Aachen University
Aachen, Germany
pompe@comsys.rwth-aachen.de

Klaus Wehrle
Chair of Communication and Distributed Systems
RWTH Aachen University
Aachen, Germany
wehrle@comsys.rwth-aachen.de

## Abstract

Estimating the energy consumption of individual devices and forecasting the total load of end-user apartments have been highly active research areas over the past decades. Many of the proposed approaches and improvements rely on data-driven algorithms, including machine learning, that require large amounts of reliable measurement data for training and evaluation. We identified a severe gap of fully-disaggregated public datasets in research and therefore propose the Device Activity Report with Complete Knowledge (DARCK) dataset, the first dataset monitoring every single appliance in an apartment, including lighting. The dataset was collected in a two-person household in Germany over the span of 6 months and provides power readings of the mains as well as of 51 different appliances at a sampling rate of $1Hz$. The dataset aims to complement the existing range of public research data in Non-Intrusive Load Monitoring (NILM) by offering data at commodity hardware precision that can be fully disaggregated. The paper describes the measurement setup and processing steps for data treatment as well as offering several noteworthy insights about the content of DARCK itself and an exemplary benchmark.

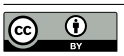
## CCS Concepts

• **Computer systems organization → Sensor networks**; • **Hardware → Energy metering**; *Smart grid*; *Sensor applications and deployments.*

## Keywords

Energy disaggregation, load monitoring, non-intrusive load monitoring, dataset, algorithm performance evaluation

**ACM Reference Format:**
Justus Breyer, Kai Gützlaff, Leonardo Pompe, and Klaus Wehrle. 2025. Dataset: Device Activity Report with Complete Knowledge (DARCK) for NILM. In *The 12th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BUILDSYS '25), November 19–21, 2025, Golden, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3736425.3771959




*BUILDSYS '25, Golden, CO, USA*

ACM ISBN 979-8-4007-1945-5/2025/11
https://doi.org/10.1145/3736425.3771959

## 1 Introduction

Reducing the pace of global warming requires concerted efforts in decreasing greenhouse gas emissions across many sectors of everyday lives, including the energy sector. This challenge is aggravated by the insufficient capacity of renewable energy sources to meet current demands and a global growth in population, calling for more awareness of the impact of everyday activity. Research indicates homeowners being more inclined to save energy when they have access to detailed information about their consumption patterns [4, 6], aiding in the global sustainability effort.

Non-Intrusive Load Monitoring (NILM) as a technique was proposed with these and other use-cases in mind [8] and has since seen a plethora of improvements across the field [1]. It builds on the idea of identifying singular devices in the data of a sensor installed at a house's main power supply based on characteristic consumption patterns (so-called signatures). The foundation of this approach is the additivity of power consumption, resulting in the consumption at the main sensor being the sum of all devices currently active in the building.

In order to evaluate newly proposed algorithms and machine learning models, excessive amounts of data, including measurements of singular devices, are necessary. Ground-truth data about the behavior of inhabitants, activity patterns of devices, concept drift in energy consumption and variations in device signatures are key for adequate model design, training and testing, in order to enable the developed models to be successfully deployed in real-world environments. To increase comparability and decrease the effort of data acquisition for researchers, the NILM community has collected and released a multitude of energy disaggregation datasets [9] over the past decades, each focusing on different aspects to improve. One of the most notable differences in the datasets is the sample rate, being either low-frequency (at line frequency or below, i.e., less than or equal to $50Hz$ in Europe) or high-frequency (above line frequency). These differences stem on the one hand from the typical hardware being used—commodity meters can provide data

only at low frequency—and on the other hand from the two areas of research of NILM algorithms [5]: Event-based algorithms use high-frequency features to classify device changes, whilst eventless algorithms analyze the data over a span of multiple seconds or minutes to infer the current device consumption. Despite the continuous improvements in the range of available datasets in terms of precision and completeness, many of them show missing periods of data, incomplete ground truth, or have been collected in environments that are unrealistic for a deployment scenario (e.g., being collected at a lab or with custom sensors that cannot be attained by the usual consumer).

We propose the Device Activity Report with Complete Knowledge (DARCK) dataset spanning 6 months of recording from a two-person apartment. It contains the aggregate reading from a main smart meter and individual readings from 40 smart plugs, smart relays, and smart power meters monitoring various appliances.

The following novel contributions compared to existing datasets were identified: (1) completeness of device monitoring—to our knowledge, we are the first to release a NILM dataset with every single device in the household being monitored, not only the plug-level loads but including consumption information of lights as well. Previous efforts of completeness (e.g., [2, 14]) included comprehensive plug level measurements but left the consumption of lights open to estimates on very detailed state tracking and device metadata. Multiple reviews over the past years (e.g., [9, 12]) found incompleteness of device surveillance to be the most critical shortcomings of released datasets. (2) usage of off-the-shelf commodity hardware that can be bought and installed with low effort. Researchers of both traditional NILM approaches as well as of hybrid approaches [15] such as Semi-Intrusive Load Monitoring (SILM) [13] can test their algorithms on data of quality that can be expected in real-world deployment scenarios. (3) continuous tracking and curated data over a span of 6 months with 99.3% completeness, increasing utility for more advanced problems such as concept drift in device energy consumption or load forecasting. Missing values were linearly interpolated to increase usability. (4) synchronized aggregated and ground truth data at a uniform sampling rate ($1Hz$). Many of the more popular datasets (e.g., [7, 10, 11]) in NILM feature a higher sampling rate for aggregate readings than for individual appliance meters. Although this drawback is not as prevalent across the range of datasets as a whole, we found research to be far more convenient if no resampling or alignment is necessary.

## 2 Measurement Setup

The DARCK dataset was collected in a two-person apartment of approximately $58m^2$ located in Germany over the span of 6 months (March until September 2025). One of the inhabitants moved out on May $31^{st}$, which resulted in a change in device composition.

### 2.1 Data Collection

Taking into account switched out devices, a total of 51 devices were monitored, using 31 Shelly Plus Plug S, 6 Shelly Plus 1PM and 3 Shelly Plus PM Mini Gen3. The Shelly Plugs were used to monitor individual power outlets, the 1PMs for wired-in devices such as ceiling lights, and the PM Minis for the three phases of the oven. The main meter of the apartment, an eBZ DD3, was monitored using an infrared reading head magnetically attached to the infrared interface of the meter. An ESP8266 flashed with Tasmota decoded the binary datagrams. All measurement devices reported their data via MQTT to Home Assistant running in docker on a Dell OptiPlex 3020M.

### 2.2 Preprocessing

Home Assistant performed a couple of preprocessing steps on the raw data to increase usability: Firstly, the off-the-shelf smart plugs are inexpensive and do not offer industry-level precision. Hence, the individual data might be inaccurate—nevertheless, the meters were calibrated using a pure resistive load. Although a linear behavior, especially considering changing humidity and temperature, cannot be assumed, these calibrations were performed to increase the accuracy of the smart plug data. The Home Assistant instance collecting the data was responsible to perform the resulting adjustment of measurement values.

Further, some smart plugs were not always connected to a socket, e.g., the plug for the vacuum cleaner was instead taped to the device. Plugs disconnecting from the network were indicated by a missing heartbeat signal. Home Assistant was responsible to detect these offline devices or a disconnect of the ESP and notify the inhabitants of potential networking issues. In case one of the measurement devices was unavailable for too long, its last current value was saved and a notification pushed to a mobile app to resolve the potential issues quickly. Outages of the ESP were only sparse and short-term (0.7%), and in case of the Shelly Plugs only occurred when disconnecting the plug from the socket.

Lastly, Home Assistant also added timestamps in *ns* precision to the published measurement values before writing the data into an InfluxDB database.

### 2.3 Postprocessing

The final dataset was generated from the InfluxDB entries using several postprocessing steps that increase the usability of DARCK for researchers in NILM and related areas.

*2.3.1 Aggregate Readings.* The aggregate data was firstly scanned for outliers: Domain knowledge of the measured environment allowed us to exclude values of below $10W$ or above $10,000W$ as definite measurement errors. Occurrences, of which there were only 3 over the complete period of time, were deleted.

Furthermore, since the network connection was not always reliable, we experienced packet bursts of delayed readings, consisting of a large time gap followed by rapid readings. A custom algorithm identified these bursts and back-filled the timestamps to create an evenly spaced time series.

Still, to create a complete series of equidistant data points, the measurement data needed to be further aligned: Since the data was collected with *ns* timestamps by Home Assistant, resampling to $1Hz$ was performed taking the mean of all readings within each second. In 99.5% of cases, this was only one reading, as would be expected under perfect conditions without outages. Any resulting gaps in the data (0.7% outage ratio) were filled using linear interpolation, resulting in a complete time series without the need for further imputation by researchers.

*2.3.2 Individual Readings.* Contrary to the continuous readings of the ESP, the Shelly devices for submetering appliances did not regularly publish readings but only at every change in power consumption. If no power change is observed or the power change is too small (less than a few Watt), the reading is pushed once every minute, together with a heartbeat. On the other hand, in case of a change of device state, the power consumption changes fast enough to trigger several readings per second.

Assuming a near-constant power draw if no new readings are pushed, we used forward-filling to increase the sample rate to $1Hz$. Since forward-filling uses the last seen value, we had to resample sub-second measurements by taking the final value received in each second instead of taking the mean. Using the mean instead would have resulted in a forward-fill of constant power draw in cases of rapid device turn-off transients instead of a power consumption of 0—analogously, steady-state behavior would also have been skewed using mean values for resampling.

To increase usability, readings of multi-purpose plugs were split into separate columns in the final dataset. After merging both the individual readings and the aggregate readings, any remaining `NaN` values, e.g., stemming from a period where the smart plug was not yet connected, were filled with 0, assuming the device did not consume power.

*2.3.3 Manual Correction.* During our post-analysis of measurement data, we compared the sum of all submeters with the aggregate measurements: If the deviation was consistently larger than $80W$ for a duration of at least $90s$, even after subtracting the consumption of the metering devices themselves, we suspected an oversight by the inhabitants in terms of complete device coverage. Two such significant unmetered load events were identified: On March $10^{th}$, an unmetered $107W$ bulb was active, and on May $31^{st}$, an unmetered $101W$ pump for an air mattress. Both of these incidents have been manually corrected in the load data: The unmetered bulb was subtracted from the main readings as if it never occurred. The air mattress, as it was monitored before at a different occasion, was manually added to the respective plug's data as if it was monitored by the plug.

Although other unmonitored device activities might still exist, it is highly unlikely, since special attention was given by the inhabitants of the apartment to not plug devices directly into an outlet without an intermediary smart plug. We chose to not use a lower threshold for identifying unmetered loads as previous research on a different dataset, collected with highly accurate sensors [14], did reveal significant discrepancies between the measured values at submeters and the mains meter, questioning the general assumption of additivity to hold in public datasets [3]. Indeed, similar to the reported tendency of isolated fridge events deviating up to $50W$ compared to the power change reported at the mains [3], we similarly found numerous events to deviate to this order of magnitude, even for always-monitored devices that could not have been accidentally plugged-in without a meter. A more in-depth analysis of these deviations and their causes, across multiple datasets and to a larger, more detailed extent is warranted but outside the scope of our contributions.

## 3 Dataset Characteristics

The DARCK dataset contains the active power measurements measured at the apartment's mains as well as of 51 different appliances. Plugs that were collecting data from different appliances were being documented and their respective appliances afterwards split up into separate columns. This procedure omits the need for additional annotation files and increases the reusability of the data. The collection period was a span of 6 months, between the $5^{th}$ of March and $4^{th}$ of September 2025. The data has been resampled to a sampling rate of $1Hz$, with no gaps or missing values remaining after the postprocessing (Section 2). The data is organized in a `csv` file, which has been compressed to 3% of its initial file size. Decompressing increases file size from about $90MB$ to $4GB$.

Additionally to the timestamp, the mains and the sensor readings, a couple of summarizing columns have been provided in the dataset: Aggregated consumption has been calculated for the chargers, the stove plates and the lighting of the apartment. This facilitates the analysis of the consumption of mobile devices and the stove respectively, and enables a quick overview of the share of energy consumed by lighting the apartment. Furthermore, since every device within the apartment has been monitored, this enables us to calculate the total measurement error, stemming from inaccuracies of the off-the-shelf hardware. We added a separate column in the dataset detailing the measurement inaccuracy after subtracting a $30W$ offset for the measurement devices themselves, which have been benchmarked before the data collection was started. The advantages of including these columns to the dataset are demonstrated in Section 4 in a short exemplary analysis.

The complete set of appliances as well as their location and a description of how to use the dataset can be found alongside DARCK itself here: 10.5281/zenodo.17159850.

## 4 Analysis

To illustrate some of the characteristics of DARCK, we performed a couple of statistical analyses.

Firstly, we investigated whether there was a change in total power consumption after one of the inhabitants moved out. For this purpose, we analyzed the mean daily consumption of the periods until May 31 and afterwards. Surprisingly, the consumption rises slightly (0.2%) after the apartment was inhabited by less persons. A deeper analysis revealed, that the fridge had an increase in the overall share of power consumption that compensated for the loss of the second inhabitant's devices, as can be seen in Figure 1. Further, the overall share of lights and the use of the oven increased.

The drastic increase in the share of the fridge triggered a second investigation into potential concept drift of device signatures: Since the measurement period spanned multiple seasons, we assumed that devices with temperature-based feedback loops might be affected by the outside temperature and draw a correlated amount of power. We researched the temperature of the inhabitants' town, which serves as a rough estimate of the actual location's outside temperature, and compared its trend with the mean weekly power consumption of the fridge. As shown in Figure 2, the power consumption of the fridge is correlated to the temperature, both showing a steady increase towards the summer season. In fact, the correlation between the mean daily power consumption of the refrigerator and the mean

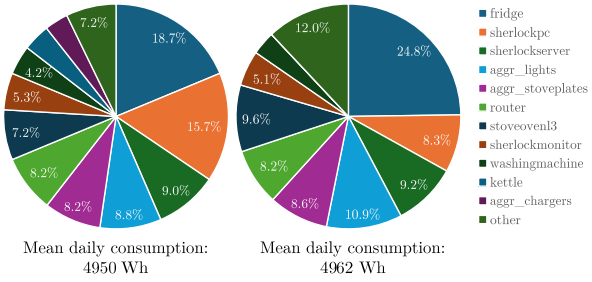



**Figure 1: The overall share of power consumption in % of different household devices with two inhabitants (left side) and one inhabitant (right side).**

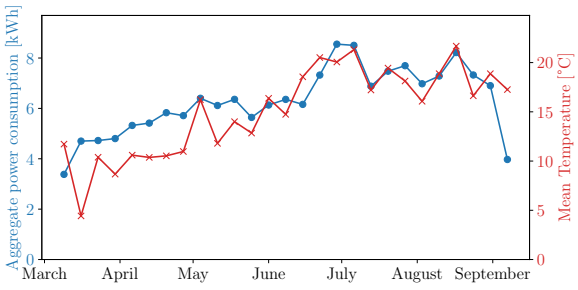



**Figure 2: The mean weekly power consumption of the fridge compared to the outside temperature. The first and last measurement week are incomplete (5 and 4 days respectively), hence their aggregate consumption is lower.**

daily ambient temperature is $r = 0.82$. A further look into the fridge's behavior revealed that the durations of its cooling phases increased drastically, whilst the signature itself did not change in amplitude.

Lastly, we performed a sanity check on the measurement inaccuracy, i.e., the difference between the sum of all measured devices (including $30W$ for the measurement devices themselves) and the mains power, which can be found in the `inaccuracy` column of the dataset. We found that the inaccuracy lies between 0.0 and $2,182.9W$, with a median of $2.7W$ and a mean of $3.5W$. Hence, the vast majority of data points lies within an acceptable error of a few Watts, and the previously discussed occasions of event deviations below $80W$ (Section 2.3.3) appear seldom compared to the overall duration of the collection period. A deeper analysis of the very high inaccuracy values above $80W$ revealed that they exclusively appeared in isolated cases, i.e., only a single value showed a high inaccuracy, after which the discrepancy dropped back to expected values. The reason behind these patterns lies in a slight desynchronization between mains and smart plug readings, stemming from the resampling: For some device state changes, e.g., turning on the oven, the jump in power consumption was registered with slight delay, resulting in an offset by one second. We did not adjust these discrepancies between smart plug and mains data, since a systematic behavior could not be identified.

**Table 1: Disaggregation Performance of Seq2Point on DARCK**

| Device | MAE |
|---|---|
| Aggr. Chargers | 6.7488 |
| Printerscanner | 1.3613 |
| Aggr. Lights | 8.0164 |
| Router | 0.3211 |
| Aggr. Stoveplates | 6.9314 |
| Oven (Stove L3) | 4.7880 |
| Fridge | 10.6103 |
| Vacuum | 2.2499 |
| Washing Machine | 4.0811 |
| Hairdryer (Sherlock) | 6.5189 |
| Monitor (Sherlock) | 5.7129 |
| PC (Sherlock) | 10.4927 |
| Server (Sherlock) | 0.2569 |
| TV (Sherlock) | 3.8718 |
| Kettle | 8.6775 |
| Microwave | 2.1201 |

## 4.1 Benchmark

We performed an exemplary disaggregation test using Seq2Point models, a type of neural networks (NNs) that have shown recent popularity in NILM [16]. Since DARCK is a low-frequency dataset, eventless approaches, e.g., using NNs, are more suitable than event-based algorithms, which typically rely on high-frequency features based on harmonics.

We trained Seq2Point models for 16 different columns of the dataset, including only devices that were active more than once per week with more than $5W$, using a 75/12.5/12.5 train/validation/test split and a batch size of $1,000$. The model architecture was designed according to related work [16] and the optimization performed using the Adam optimizer and early stopping with a patience of 3. The respective Mean Absolute Error (MAE) in $W$ is reported in Table 1 and scores on the DARCK dataset in the same orders of magnitude as reported results on other datasets such as UK-DALE [16].

## 4.2 Discussion

A couple of limitations are known to us, as the DARCK dataset was not designed to fit every research purpose in the area of energy consumption analysis. Firstly, the dataset contains only measurements of a singular household—the effort to maintain an extensive infrastructure capable of monitoring every single electrical load is hard to scale to multiple buildings. However, since the number of inhabitants changed over time, opportunities for generalization analyses are given. Secondly, the sample rate of the dataset is $1Hz$, which excludes high-frequency analyses in NILM from taking advantage of harmonics related features. Lastly, the measurement infrastructure is commodity hardware with the sensor inaccuracies that would be expected from low-cost devices. This results partly in non-negligible measurement discrepancies that will increase the difficulty for some machine learning models to correctly disaggregate the consumed energy. On the flip side, the given inaccuracy provides a realistic scenario for researchers that aim to deploy their models in actual household settings, as similar discrepancies in

collected data will very likely be experienced. In a similar vein of limitations lies the measured quantity, which is only the active power. Although more data, e.g., on temperature, humidity, water consumption, or even reactive power, might be helpful to increase the precision of the trained models, ultimately we do not expect customers in real—world use cases to be capable of providing this additional information to the energy disaggregation models. Hence, the provided data serves as an opportunity for researchers aiming to build models for real-world deployments in realistic scenarios.

## 5 Conclusion

We present the DARCK dataset, spanning mains and individual device measurements of a two-person apartment in Germany over a period of 6 months, where after half of the measurement period the number of inhabitants changes. Measurements have been synchronized to a sample rate of $1Hz$, enabling research on low-frequency energy consumption data. DARCK is the first dataset to include submetering of every appliance in a household, including lights, allowing for research in the areas of NILM, energy forecast and related approaches such as SILM. A short analysis of the collected data confirms the presence of concept drift for at least one appliance (the fridge), offering the opportunity for researchers to test their algorithm's adaptability. Finally, we offer some baseline results performed with Seq2Point models for other researchers to compare against. By providing this dataset, we help researchers improve their algorithms on more advanced problems such as changing device constellations and concept drift in device activations. Since our measurement is still ongoing, we hope to extend DARCK in the future with at least one update.

## References


[1] Georgios-Fotios Angelis, Christos Timplalexis, Stelios Krinidis, Dimosthenis Ioannidis, and Dimitrios Tzovaras. 2022. NILM Applications: Literature Review of Learning Approaches, Recent Developments and Challenges. *Energy and Buildings* 261 (2022), 111951.

[2] Nipun Batra, Manoj Gulati, Amarjeet Singh, and Mani B Srivastava. 2013. It's Different: Insights into Home Energy Consumption in India. In *Proceedings of the 5th ACM workshop on embedded systems for energy-efficient buildings*. 1–8.

[3] Justus Breyer, Jonas Koerhuis, Muhammad Hamad Alizai, and Klaus Wehrle. 2025. Practical Insights from Implementing Event-Based NILM Systems. In *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*. 751–756.

[4] K. Carrie Armel, Abhay Gupta, Gireesh Shrimali, and Adrian Albert. 2013. Is Disaggregation the Holy Grail of Energy Efficiency? The Case of Electricity. *Energy Policy* 52 (2013), 213–234.

[5] Suryalok Dash and NC Sahoo. 2022. Electric Energy Disaggregation via Non-intrusive Load Monitoring: A State-of-the-Art Systematic Review. *Electric Power Systems Research* 213 (2022), 108673.

[6] Karen Ehrhardt-Martinez, Kat A Donnelly, Skip Laitner, et al. 2010. Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities. American Council for an Energy-Efficient Economy Washington, DC.

[7] Adrian Filip et al. 2011. BLUED: A Fully Labeled Public Dataset for Event-Based Non-Intrusive Load Monitoring Research. In *2nd Workshop on Data Mining Applications in Sustainability (SustKDD)*, Vol. 2012. 5.

[8] George William Hart. 1992. Nonintrusive Appliance Load Monitoring. *Proc. IEEE* 80, 12 (1992), 1870–1891.

[9] Hafiz Khurram Iqbal, Farhan Hassan Malik, Aoun Muhammad, Muhammad Ali Qureshi, Muhammad Nawaz Abbasi, and Abdul Rehman Chishti. 2021. A Critical Review of State-of-the-Art Non-Intrusive Load Monitoring Datasets. *Electric Power Systems Research* 192 (2021), 106921.

[10] Jack Kelly and William Knottenbelt. 2015. The UK-DALE Dataset, Domestic Appliance-Level Electricity Demand and Whole-House Demand from Five UK Homes. *Scientific data* 2, 1 (2015), 1–14.

[11] J Zico Kolter and Matthew J Johnson. 2011. REDD: A Public Data Set for Energy Disaggregation Research. In *Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA*, Vol. 25. Citeseer, 59–62.

[12] Lucas Pereira and Nuno Nunes. 2018. Performance Evaluation in Non-Intrusive Load Monitoring: Datasets, Metrics, and Tools—A Review. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery* 8, 6 (2018), e1265.

[13] Guoming Tang, Kui Wu, and Jingsheng Lei. 2015. A Distributed and Scalable Approach to Semi-Intrusive Load Monitoring. *IEEE Transactions on Parallel and Distributed Systems* 27, 6 (2015), 1553–1565.

[14] Benjamin Völker, Marc Pfeifer, Philipp M Scholl, and Bernd Becker. 2020. FIRED: A Fully-labeled hIgh-fRequency Electricity Disaggregation Dataset. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 294–297.

[15] Benjamin Völker, Philipp M Scholls, Tobias Schubert, and Bernd Becker. 2018. Towards the Fusion of Intrusive and Non-Intrusive Load Monitoring: A Hybrid Approach. In *Proceedings of the Ninth International Conference on Future Energy Systems*. 436–438.

[16] Chaoyun Zhang, Mingjun Zhong, Zongzuo Wang, Nigel Goddard, and Charles Sutton. 2018. Sequence-to-Point Learning with Neural Networks for Non-Intrusive Load Monitoring. In *Proc. AAAI*, Vol. 32.