Renewable Energy, Environment, and Power Quality Journal (REE&PQJ)
ISSN:3020-531X
Volume 26, July 2026

Integration of High-Frequency and Harmonic Measurements from Open-Source Meters into NILMTK Datasets

C. Rodriguez-Navarro[1], F. Portillo[1], I. Robalo[1], L. Castro-Santos[2], A. Filgueira-Vizoso[3], F. G. Montoya[1] and A. Alcayde[1]

[1] Departamento de Ingeniería
Escuela Superior de Ingeniería, Universidad de Almería
Carretera Sacramento s/n, La Cañada de San Urbano, 04120 Almería (Spain)
e-mail: crn565@inlumine.ual.es, portillo@ual.es, isabel.robalo.cabrera@ual.es, pagilm@ual.es, aalcayde@ual.es

[2] Departamento de Enxeñaría Naval e Industrial
Centro de Investigación en Tecnoloxías Navais e Industriais (CITENI), Escola Politécnica de Enxeñaría de Ferrol, Universidade da Coruña
Campus Industrial de Ferrol, Esteiro, 15471 Ferrol (Spain)
e-mail: laura.castro.santos@udc.es

[3] Departamento de Química
Centro de Investigación en Tecnoloxías Navais e Industriais (CITENI), Escola Politécnica de Enxeñaría de Ferrol, Universidade da Coruña
Campus Industrial de Ferrol, Esteiro, 15471 Ferrol (Spain)
e-mail: almudena.filgueira.vizoso@udc.es

Abstract. This work presents a methodology for generating datasets compatible with the Non-Intrusive Load Monitoring Toolkit (NILMTK) using measurements from open-source smart meters: OpenZmeter (oZm) and Open Multi Power Meter (OMPM). The purpose is to advance research on Non-Intrusive Load Monitoring by enabling detailed analysis of energy disaggregation and power quality. The main challenge is integrating data sources with extended timestamps and high-frequency electrical harmonics into the standardised NILMTK format. To solve this, custom data converters were developed for different versions of oZm (with and without harmonic support) and for OMPM (without harmonics). These converters transform HDF5/CSV data into NILMTK-compatible HDF5 structures, including up to ten electrical variables and conversion of 13-digit timestamps. Datasets with harmonic support provide 150 additional measurements of harmonics up to the 50th order. Modifications to NILMTK components enable recognition of these new formats. Results demonstrate the feasibility of extending NILMTK for advanced analysis in power-quality and energy-efficiency research.

Key words. Dataset generation, Non-Intrusive Load Monitoring, NILMTK framework, Open-source smart meters, Power quality analysis.

Acronyms. **CSV:** Comma-Separated Values, **HDF5:** Hierarchical Data Format version 5, **JSON:** JavaScript Object Notation, **NILM:** Non-Intrusive Load Monitoring, **NILMTK:** Non-Intrusive Load Monitoring Toolkit, **NILMTK-DF:** NILMTK Data Format, **OMPM:** Open Multi Power Meter, **oZm:** OpenZmeter,·**REST API:** Representational State Transfer Application Programming Interface,·**THD:** Total Harmonic Distortion,·**YAML:** YAML Ain't Markup Language.

1. Introduction

The Non-Intrusive Load Monitoring (NILM) approach has become an essential tool for promoting energy efficiency in both residential and industrial environments [1]. It enables disaggregating total energy consumption into appliance-level profiles using only aggregate electrical measurements. Among the software frameworks developed for this purpose, the Non-Intrusive Load Monitoring Toolkit (NILMTK) [2] stands out as one of the most widely adopted due to its capacity to manage standardised datasets and train accurate disaggregation models efficiently.

However, NILMTK presents limitations when integrating data from non-standard or high-resolution measurement sources [3], such as those produced by the new Open Zmeter (oZm) [4] and the Open Multi Power Meter (OMPM) [5] devices, designed for advanced electrical monitoring applications. These meters record high-resolution electrical variables using 13-digit timestamps that represent milliseconds since the UNIX epoch. In the case of the oZm, it also includes voltage, current, and power harmonics up to the 50th order. The resulting data structures can contain up to 160 parameters per record: 150 correspond to harmonic measurements (50 orders × 3 types: voltage, current, and power), while the remaining 10 correspond to fundamental electrical variables (timestamp, active power, reactive power, apparent power, frequency, voltage, power factor, and current). This complexity exceeds the capabilities of the conventional NILMTK format and, therefore, cannot be integrated directly into the framework.

To overcome this limitation, a set of custom data converters was developed to adapt oZm and OMPM measurements into the NILMTK Data Format (NILMTK-DF). These converters manage temporal transformation, reorganise the essential electrical fields, and incorporate harmonic analysis support, ensuring full compatibility with the NILMTK ecosystem.

This contribution facilitates the inclusion of high-resolution data sources in NILM studies, enabling simultaneous assessment of power quality and energy efficiency. Consequently, it strengthens NILMTK's capabilities for advanced analysis of renewable energy and hybrid generation systems, supporting ongoing efforts to develop intelligent, sustainable energy-monitoring infrastructure.

The remainder of this paper is structured as follows. Section 2 reviews related work on NILMTK datasets and existing converters. Section 3 describes the methodology, including environment preparation, data capture, preprocessing, and the development of custom converters for oZm and OMPM devices. Section 4 presents the resulting datasets and the results obtained from their analysis. Finally, Section 5 concludes the paper and outlines directions for future work.

2. Related Work

NILMTK is an open-source tool widely used in energy disaggregation research. It standardises the workflow for breaking down total energy consumption into appliance-level data, facilitating dataset sharing, result comparison, and algorithm evaluation. The toolkit defines a common data format and detailed metadata for interoperability and reproducibility. It processes energy measurements through conversion and preprocessing, producing datasets in HDF5 format for efficient storage and analysis. Custom converters have been created to handle high-precision timestamps and generate new datasets. Despite its benefits, NILMTK has limitations in integrating advanced or high-resolution data, restricting its use to more conventional formats.

NILMTK includes several dataset converters, also known as data importers, that allow different public energy consumption datasets to be transformed into the standard HDF5 format. Until the most up-to-date version, which includes both the main toolkit and its NILMTK-Contrib [6] community extension, converters are implemented for more than 10 datasets widely used in NILM.

The main official repository (nilmtk/dataset_converters/) includes converters for the following datasets: Reference Energy Disaggregation Dataset (REDD) [7], UK Domestic Appliance-Level Electricity (UK-DALE), Dataport (Pecan Street) [8], Almanac of Minutely Power Dataset (AMPds) [9], Indian Ambient Water and Energy (iAWE), Electricity Consumption & Occupancy (ECO) [10], Residential Electricity Usage Data from the Innovative Technology (REFIT), and the Green Energy Consumption Dataset (GREEND) [11]. In addition, the NILMTK-Contrib module, developed by the research community, complements this list with additional converters that support datasets such as Commercial Building Dataset (COMBED), Smart Star, ENERTALK, UK-DALE REDUCED (processed version), Dataport-REDUX (adaptation to new Pecan Street formats), and REDD-LowFreq (low-frequency variant of REDD).

Among the most notable datasets, iAWE, [12] was developed in India to analyse residential energy use. It uses 33 sensors to record real-time electricity and gas data continuously for 73 days, including aggregate and appliance-specific measurements. iAWE supports supervised NILM techniques but highlights challenges in disaggregating nonlinear or multistage loads. Stable appliances, such as air conditioners, perform well with classic algorithms, whereas complex devices, such as washing machines, require more advanced models.

Most public datasets focus on low-frequency or conventional data structures, limiting their capacity to represent high-resolution or harmonic measurements. This limitation motivated the development of advanced data sources such as oZm and OMPM, along with specific NILMTK converters, expanding the toolkit's ability to analyse power quality and energy efficiency in sophisticated smart monitoring scenarios.

3. Methodology

The methodology for generating new NILMTK-compatible datasets is articulated into several fundamental phases: environment preparation, data capture, pre-processing, conversion software, and post-processing, all aimed at transforming electrical data from the new devices (oZm v1, oZm v2, and OMPM) into NILMTK-compatible datasets.

A. Environment Preparation and Data Capture

The environment for generating a NILM dataset is based on open-source hardware and an adapted version of NILMTK. It uses precision open meters like oZm, capable of capturing fundamental electrical variables (up to 10 parameters) plus voltage, current, and power harmonics up to the 50th order, resulting in up to 160 electrical fields per record. The system architecture includes a main meter for aggregate consumption and multiple submeters for individual appliances, with configurations ranging from six to 12 channels to capture detailed operational records over several hours. This setup enables comprehensive monitoring while prioritising commonly used appliances due to meter limitations. The structure distinguishes

between aggregate and individual measurements of appliances.
To simulate real consumption scenarios in a controlled, replicable way, a sequencer module can be used: either based on an Arduino Uno board and a relay board for OMPM or using Tasmota devices controlled by a Python script. This mechanism enables automated control of the application's on/off state, generating operation records with different sequences and time intervals.

In the case of oZm or other meters, to obtain the measurements it is necessary to request the oZm API, passing through the POST method, the URL of the device, the header where it is indicated that an array of data in Json format is to be retrieved by adding the cookie, and in the body, the measurement interval (200 ms), the parser ID, the source and target values, as well as the list of characteristics to be analysed.

Once the response of each oZm has been obtained, the results can be stored in a file for processing in the subsequent phase of dataset generation. When other meters are used, this step is not required because certain devices (such as the OMPM) can directly provide the measurement files for each application, together with the aggregate, in CSV format.

B. Data Preprocessing

Once the meter operation records have been acquired in the previous phases (through an API or CSV/HDFS files), the process focuses on adapting that data to the NILMTK structure, which requires a pre-processing step (only for oZm). i.e., the raw data files, which can contain up to 160 data fields when harmonic support is enabled, must be processed to select the relevant fields and adapt to the format required by NILMTK-DF.

At this stage, the obtained files are collected into files with up to 160 fields per record, in HDFS or CSV format. Since many of these fields are not relevant to NILM analysis, a selection must be made based on the meter type, harmonic support, and the number of connected applications. This selection reduces the data's dimensionality and facilitates its adaptation to the NILMTK-DF format.

The pretreatment process can be implemented using Python scripts that execute the following operations on each measurement file:

- *Loading the file into a DataFrame (Dataset):* Using the Pandas library.
- *Calculation of the apparent power:* From the active and reactive power, using square-root operations on the sum of the squares of both components.
- *Renaming columns:* To fit the NILMTK nomenclature (e.g., 'Active_Power' to 'W', 'Reactive_Power' to 'VAR', 'Voltage' to 'VLN').
- *Elimination of non-electrical or redundant fields:* Such as internal indexes, labels, flags, and complex fields not processable by NILMTK.
- *Reordering columns:* To establish a coherent structure and final export of the measure set in CSV format.

For datasets with harmonics, a modulus function computes the absolute value of complex pairs (real and imaginary parts), generating 150 additional measurements (voltage, current, and power harmonics up to the 50th order). In addition, a detailed temporal analysis of each measurement file can be performed by counting the number of samples and retrieving the original timestamps of the first and last rows.

These values, expressed in 13-digit Epoch UNIX format, must be converted to a readable format using the 'datetime' library's functions to create a coherent time series.

Precisely this change in the timestamp format, which is not supported by NILMTK's standard converters, is one of the main motivations for developing new, specific converters.

C. Creating New Datasets

Since NILMTK uses standardised formats for NILM data processing, it is also necessary to develop new converters that adapt the new data to the NILMTK-DF format (especially regarding the extended 13-digit timestamp format and harmonic support, if the meter supports it). These constitute the three phases required for the creation of a new dataset:

1) *Development of new converters:* Due to the novelty of the data offered by the new meters, the development of new converter functions in Python is required. These converters (such as 'convert_ualm','convert_ualmt','convert_ualm10H ', or 'convert_ualm2') are based on the IAWE converter and support a 13-digit timestamp (including milliseconds), which is vital for time accuracy. These converters must be registered in the NILMTK structure (e.g., in the '__init__.py' file) for use.
2) *Generation of metadata files:* Metadata files are created in YAML format that describe the data set, the building configuration, and the specifications of the meters.
3) *HDF5 format generation:* The final process converts the pre-processed CSV files and YAML metadata into a single optimised file in HDF5 format (NILMTK-DF), which is the standard format for storing large time series measurements and metadata.

Given the small number of converters in meter-specific NILMTKs, it is necessary to develop new converters specific to the selected meters after the measurement files have been pre-treated. These converters are based on the iAWE model but incorporate substantial modifications to support extended timestamps, new electrical measurements, and differentiated structures by meter type.

The new converter code, based on the iAWE converter, includes a column mapping that links the original names of

the electric fields (W, A, PF, VA, VAR, VLN, f) with the measurement categories used by NILMTK (active power, current, power factor, etc.).
In cases with harmonic support, 150 additional fields are added corresponding to the voltage (VH), current (IH), and power (PH) harmonics, from order one to 50. This mapping is implemented using loops that dynamically generate the keys needed for processing. Also, a key operation in the converter is converting the 13-digit timestamp field to a date format.

Each converter performs a series of operations to transform raw measurements into the NILMTK-DF format. First, the 13-digit timestamp field is converted to a standard datetime format. Subsequently, the electrical variables (active, reactive, and apparent power; voltage, current, power factor, and frequency) are loaded and mapped to the corresponding NILMTK measurement categories. The data is then sorted, resampled, and reindexed to ensure temporal consistency. This process is repeated for all measurement files, after which the metadata is incorporated, and the final dataset is saved in HDF5 format.

As for the metadata structure required by NILMTK, it is defined in the subdirectory '/metadata/', a directory that must contain the following three files in YAML:

1) *'dataset.yaml':* Includes general information about the dataset, such as name, authors, publication date, institutional contact, geographical location (e.g., Almeria), time zone (Europe/Berlin), etc.
2) *'building1.yaml':* Configure the electric meters used in the experiment. For the oZm v1, six devices are defined (one main and five submeters), while for the oZm v2, it is extended to 11 devices, reflecting a greater diversity of connected appliances.
3) *'meter_devices.yaml':* Describes the physical characteristics of the type of meter, including the supported measurements (active, reactive, apparent power, voltage, current, frequency, power factor) and, in cases with harmonics, the 150 additional measurements corresponding to harmonics up to the order 50.

Each converter accesses the measurement files in the '/electricity/' folder and processes the numbered files (1.csv, 2.csv, etc.) according to the contents of 'labels.dat', which links each file to the corresponding appliance. For example, in the case of oZm v1, the 1.csv file corresponds to the main meter (mains), while the files 2.csv to 6.csv are associated with devices such as fan, freezer, television, vacuum cleaner, and boiler (Fig. 1).

In the case of the oZm v2, the number of applications is expanded to 11, including electric oven, microwave, and shower heater. For both the metadata files and the Python code for each converter to be processed by the toolkit, new lines must be added to the __init__ file to support the new converters.

Once the measurement files have been located and the new converter has been saved (including the corresponding entry in the init file), the specific converter can be invoked by calling the new function (e.g., convert_ualm). The metadata path and the new name of the dataset file to be generated in H5 format are provided as arguments.

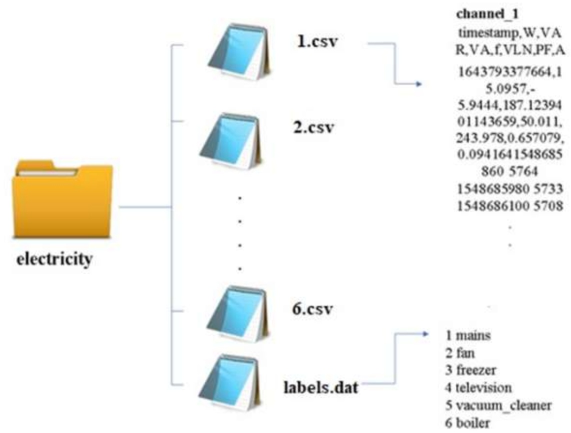

Fig. 1. Structure of the measurement files for five applications.

D.-Analysis and post-processing.

As an example, five new datasets were generated using the methodology described (DSUALM, DSUALMH, DSUALM10, DSUALM10H, and UALM2). The main differences between them, excluding the specific types of devices, lie in the complexity of the measurements:

- *DSUALM / DSUALMH:* Generated with the oZm v1, with five applications plus the aggregate. DSUALMH includes voltage, current, and power harmonics up to order 50, while DSUALM does not.
- *DSUALM10/DSUALM10H:* Extended versions generated with the oZm v2/v3, which include ten applications plus the aggregate, increasing the complexity of disaggregation tasks. DSUALM10H includes harmonics.
- *UALM2:* Created with the OMPM, this dataset focuses on monitoring low-power devices (five applications plus the aggregate) to increase the difficulty of disaggregation, without including harmonics.

Regarding metadata and loaded devices, it presents information on the meters (ElecMeter) loaded into the dataset, including whether they are main meters (site_meter) or submeters associated with household appliances (e.g., fan, freezer, television, kettle, etc.).

To improve dataset quality and optimise NILM model performance, once the dataset has been generated, it is advisable to apply additional statistical and analytical techniques. Particularly useful is the good_sections() method, which identifies continuous time intervals with valid data, allowing incomplete or corrupted segments to be filtered out. It is also recommended to implement resampling strategies using the resample_kwargs parameter, which defines the arguments passed to the pandas resample () function. This makes it possible to adjust the sampling frequency through the sample_period or rule parameter and specify the aggregation or interpolation method (e.g., mean, sum, first, or forward fill). In addition, temporal analysis can be complemented by evaluating the overall time coverage (get_timeframe()) and identifying valid sections (good_sections()) that contain no significant measurement gaps. Missing data or temporal discontinuities are determined from each meter type's metadata, specifically via the max_sample_period parameter, which defines the maximum allowable time interval between consecutive measurements before the meter is considered to have a data gap or be inactive. In the example in Fig. 2, the source result is "0.0", indicating that there are no missing samples.

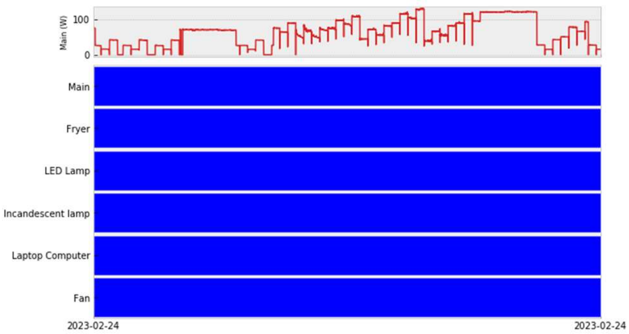

Fig. 2. Missing data for UALM2.

To reduce noise and measurement uncertainty, samples with low electrical values are discarded using a threshold-based filter. Active power values below 10 W are excluded due to their high relative uncertainty. For harmonics, components below 1% of the fundamental are omitted, as their influence on power quality analysis is negligible. This strategy maintains a high signal-to-noise ratio while preserving relevant electrical events.

These steps are integrated into the conversion process via functions that reindex data frames, handle missing values, sort indices, and save the results to HDF5 with metadata. This creates a structured, complete, and NILMTK-compliant dataset that enables detailed, disaggregated energy-consumption analysis with high temporal and spectral resolution.

E. Validation of Harmonic Data Transfer

The validation was performed through:

- *Metadata verification:* Harmonic fields are explicitly declared in meter_devices.yaml.
- *Data accessibility:* Harmonic data can be retrieved using the meter.get_measurements ('current', 'harmonic3').
- *Content verification:* Raw oZm API values match exactly those in the converted NILMTK-DF dataset.

4. Results

Energy consumption distribution (Fig. 3) shows clear differences among appliances, with incandescent lamps exhibiting higher consumption than laptops. Activity logs (Fig. 4) record events above a 10 W threshold, capturing appliance operation peaks. Electrical variable plots (Fig. 5) reveal characteristic patterns, such as motor-starting peaks in vacuum cleaners and freezers, and purely resistive behaviour in kettles.

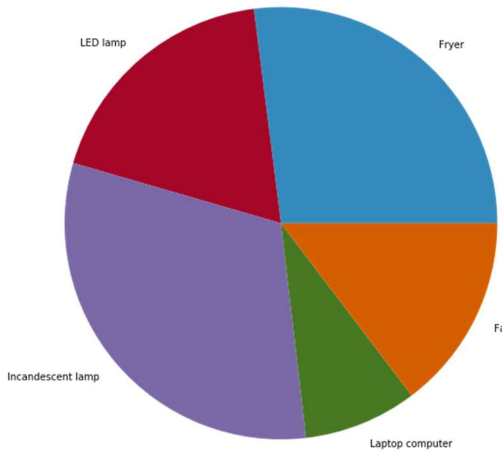

Fig. 3. Plot the fraction of energy consumption of each appliance.

Fig. 4. Activity log for using a 10 W threshold for UALM2.

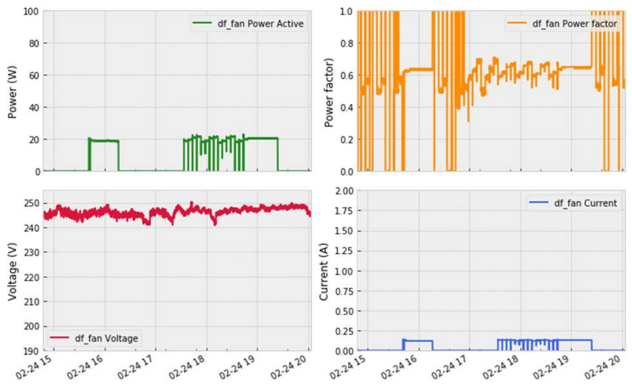

Fig. 5. Graph of the power, frequency, voltage, and current for meter number 2.

A summary of the UALM2 dataset shows six meters (one main, five submeters), a high correlation (0.996) between the aggregate and submeter sums, 95.2% of energy submetered, no data gaps, and 60-second sample periods.

To evaluate the impact of harmonic inclusion on NILM disaggregation performance, Fig. 6 compares key metrics across multiple public datasets generated with commercial hardware versus the open-source hardware datasets proposed in this work. Among the latter, datasets that include harmonic measurements (DSUALMH and

DSUALM10H) consistently achieve lower metric values, indicating better disaggregation accuracy than their non-harmonic counterparts (DSUALM and DSUALM10). This improvement is particularly evident in metrics such as MNEAP, RMSE, and EAE, where the inclusion of harmonic information reduces prediction error by capturing finer electrical signatures of nonlinear loads. Lower values in these metrics correspond to better disaggregation performance.

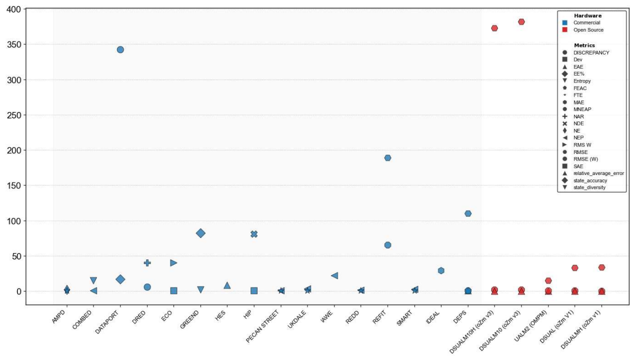

Fig. 6. Comparison of different NILM metrics in datasets that use commercial or non-commercial hw and in open datasets, those that include or do not include harmonics.

5-Conclusions

This work developed new data converters to generate NILMTK-compatible datasets from advanced open-source smart meters (oZm and OMPM). The converters enable processing of high-resolution measurements with extended timestamps and harmonic support, overcoming limitations of existing NILMTK tools. Validation of the harmonic data transfer was performed by verifying that the 150 harmonic fields (voltage, current, and power harmonics up to the 50th order) are correctly declared in the metadata, accessible via NILMTK's standard functions, and retain their original values after conversion.

Unlike existing NILMTK datasets, which are limited to fundamental electrical variables at low sampling rates, the proposed converters enable the integration of high-resolution measurements with millisecond-precision timestamps and harmonic content up to the 50th order, extending NILMTK's applicability to power quality analysis and advanced energy disaggregation.

The resulting datasets provide detailed, standardised resources for non-intrusive load monitoring research, enhancing analysis of energy disaggregation and power quality. This contribution supports reproducible research and advances in intelligent energy monitoring systems.

Future work should focus on expanding device support and optimising data processing to improve efficiency and support additional electrical variables.

References

[1] Juliano Souza, ‘Non-Intrusive Load Monitoring (NILM): An Introduction and practical example.’, https://www.linkedin.com/pulse/non-intrusive-load-monitoring-nilm-introduction-practical-souza-rpy4f/. Accessed: Mar. 11, 2025. [Online]. Available: https://www.linkedin.com/pulse/non-intrusive-load-monitoring-nilm-introduction-practical-souza-rpy4f/

[2] B. Nipun, K. Jack, P. Oliver, D. Haimonti, and K. William, ‘NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring’, Apr. 2014, doi: 10.1145/2602044.2602051.

[3] J. Kelly *et al.*, ‘NILMTK v0.2: A non-intrusive load monitoring toolkit for large scale data sets’, in *BuildSys 2014 - Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, Association for Computing Machinery, Nov. 2014, pp. 182–183. doi: 10.1145/2674061.2675024.

[4] E. Viciana, A. Alcayde, F. G. Montoya, R. Baños, and F. Manzano-Agugliaro, ‘An open hardware design for internet of things power quality and energy saving solutions’, *Sensors* , vol. 19, no. 3, Feb. 2019, doi: 10.3390/s19030627.

[5] C. Rodríguez-Navarro, F. Portillo, F. Martínez, F. Manzano-Agugliaro, and A. Alcayde, ‘Development and Application of an Open Power Meter Suitable for NILM’, *Inventions*, vol. 9, no. 1, p. 2, Dec. 2023, doi: 10.3390/inventions9010002.

[6] R. Kukunuri *et al.*, ‘NILMTK-Contrib: Towards reproducible state-of-the-art energy disaggregation’. [Online]. Available: https://github.com/nilmtk/buildsys2019-paper-notebooks

[7] J. Z. Kolter and M. J. Johnson, ‘REDD: A Public Data Set for Energy Disaggregation Research’, Cambridge, 2011. Accessed: Aug. 29, 2025. [Online]. Available: http://redd.csail.mit.edu/

[8] P. Oliver, F. Grant, H. April, B. Nipun, and K. Jack, ‘Dataport and NILMTK: A building data set designed for non-intrusive load monitoring.’, *IEEE*, 2015, Accessed: Apr. 21, 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7418187/keywords

[9] S. Makonin, F. Popowich, L. Bartram, B. Gill, and I. V. Bajic, ‘AMPds: A Public Dataset for Load Disaggregation and Eco-Feedback Research.’, *IEEE Electrical Power*.

[10] 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 *BuildSys 2014 - Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, Association for Computing Machinery, Nov. 2014, pp. 80–89. doi: 10.1145/2674061.2674064.

[11] A. Monacchi, D. Egarter, W. Elmenreich, S. D’Alessandro, and A. M. Tonello, ‘GREEND: An energy consumption dataset of households in Italy and Austria’, in *2014 IEEE International Conference on Smart Grid Communications, SmartGridComm 2014*, 2015. doi: 10.1109/SmartGridComm.2014.7007698.

[12] C. Rodriguez-Navarro, ‘PhD thesis Explore the horizons of unbundling electricity consumption and emissions: innovations in hardware (OZM and OMPM) and open-source software’, Almeria, Oct. 2025. Accessed: Oct. 30, 2025. [Online]. Available: https://dialnet.unirioja.es/servlet/tesis?codigo=394849