DATA MANAGEMENT PLAN

0. Proposal name

Project: Transfer Learning Method for Energy Disaggregation

Author: Rohit Kumar	*Version: 1*	*Date: 12 June 2023*

1. Description of the data

1.1 Type of study

This work seeks to understand the effective NILM system frameworks and review the performance of the benchmark algorithms. Following this, the paper proposes the approach of using transfer learning methods to achieve high-accuracy load disaggregation and tackles the problem of generalizability by investigating domain adaptation methods for energy disaggregation.

1.2 Types of data

This dataset records the power demand from five houses. In each house, we record both the whole-house mains power demand every six seconds as well as the power demand from individual appliances every six seconds. In three of the five houses (houses 1, 2, and 5) we also record the whole-house voltage and current at 16 kHz.

1.3 Format and scale of the data

High speed (16 kHz) whole-house voltage and current data for 1 house. Data is stored in individual directories for each week. This dataset continues UK-DALE-2015: UK-DALE-16kHz.

Description: High speed (16 kHz) whole-house voltage and current data for 1 house. Data is stored in individual directories for each week. This dataset continues

Data Type: Time Series

Number of Records: 19500

Parameter Names: Timestamp (UNIX epoch = number of seconds since 1970/01/01 00:00 UTC) Appliance power consumption or Aggregate household power (N.B. some meters record active power (W) and some meters record apparent power (VA).

An HDF5 version of the 1-second and 6-second data (for use with NILMTK) is available on the UKERC EDC. The complete April 2017 version of the 16kHz dataset occupies 7.6 TBytes. The 16 kHz data are stored as a sequence of stereo FLAC files ("FLAC" stands for "Free Lossless Audio Codec"). Each FLAC file is about 200 MBytes. One channel is whole-house voltage, the other is whole-house current.

2. Data collection/generation

2.1 Methodologies for data collection/generation

The dataset was first collected by Jack Kelly and the team and is now present on the UKERC EDC website. The dataset comprises recordings from five houses, capturing detailed information every six seconds. Specifically, recorded the active power consumption of individual appliances and the overall apparent power demand in all houses. Furthermore, in three houses, collected data at a sampling rate of 44.1 kHz for whole-house voltage and current, which was down-sampled to 16 kHz for storage. Additionally, calculated the active power, apparent power, and RMS voltage at a lower frequency of 1 Hz. In House 1, conducted recordings for a period of 655 days and meticulously captured data from nearly every appliance in the house. This resulted in a total of 54 separate channels of recorded information, although fewer channels were recorded during the initial stages of the dataset. Notably, the recordings in House 1 will continue for the foreseeable future, enabling long-term *observations. For the other four houses, recorded data for several months. Each of these houses recorded between 5 and 26 channels of individual appliance data, providing valuable insights into their energy consumption patterns.*

2.2 Data Quality and Standards

When using the UK-DALE dataset for transfer learning in Non-Intrusive Load Monitoring (NILM), data quality and standards are crucial. Through data cleaning techniques like interpolation, smoothing, and outlier detection, it is crucial to resolve missing values, outliers, and inconsistencies in order to assure dependable and accurate findings. The dataset can be made consistent and comparable between various attributes and occurrences by normalizing it. The efficiency of transfer learning is increased by identifying the most instructive features for NILM using methods like correlation analysis or mutual information. The dataset's accurate annotation and labeling provide training and evaluation with ground truth data. Reproducibility is ensured by maintaining thorough metadata documentation, which includes information about data collection, preprocessing procedures, sensor specifications, and variable definitions.

3. Data management, documentation and curation

3.1 Managing storing, and curating data.

For use with the greater computing capacity provided by Google Colab Pro, the UK-DALE dataset has been downloaded and moved to a Google Drive account. There are no issues with its backup because the training data is openly accessible. However, training logs will contain information about the trained networks, including their designs and weights. These logs will be locally backed up and stored in the GitHub repository along with the project source.

3.2 Metadata Standards and data documentation

The network designs and hyperparameter information needed to train the models will be outlined in the main report's appendices. Specific hyperparameter values will be explicitly mentioned in a README file within the project's GitHub repository, together with training schemas and algorithms, to ensure reproducibility. Each training and validation cycle's duration, machine characteristics, and training environment will all be noted down. These specifics will be helpful for comparing the models' performance, particularly in situations where speedy execution is essential, like segmenting live images for autonomous vehicle applications.

4. Data security and confidentiality of potentially disclosive information

4.1 Formal information/data security standards

Since training data is entirely public data, security issues with third-party storage solutions (raised by ncl.ac.uk [Working | University Library]) are not a concern. Indeed, even any trained networks we create push towards a philanthropic cause meaning public distribution is encouraged.

5. Data sharing and access

5.1 Suitability for sharing

Yes, all of the information I'll be using is openly accessible, has received a lot of peer review, and has been extensively cited in numerous academic works. It will be encouraged to share code and experiments since this project intends to apply transfer learning methods to datasets that are currently available in the wild.

5.2 Discovery by potential users of the research data

All code libraries will exist within the public project GitHub repository (https://github.com/mustang-raven999). If the project successfully manages to fulfill its aim, the publication will be sought due to the universal application across domains for fair AI, whence a DOI will be generated to facilitate discovery.

5.3 The study team's exclusive use of the data

According to government guidelines, the research data will be stored and maintained indefinitely. The precise retention duration will be chosen and adhered to, ensuring the data's ongoing accessibility and availability. Adhering to specified preservation standards guarantees the study data's safekeeping and continuing upkeep, permitting its future reference, analysis, and prospective use in additional research projects.

5.4 Restrictions or delays to sharing, with planned actions to limit such restrictions

There are no substantial restrictions or delays in sharing the data as it is publicly available. Researchers can freely access and utilize these datasets without the need for additional licenses or approvals. The data from the sensors maintained by the UKERC EDC website can be easily accessed by anyone, enabling broad access and use for research purposes.

6. Responsibilities

Are there any resources (e.g. storage/ training) that you will require to fulfil the plan?

Google Colab Pro+ ≈ £45 / month

Google Drive Storage (100GB) ≈ £1.59 / month

7. Relevant institutional, departmental or study policies on data sharing and data security

Policy	URL or Reference
Data Management Policy & Procedures	https://www.ncl.ac.uk/media/wwwnclacuk/research/files/ResearchDataManagementPolicy.pdf
Data Security Policy	https://services.ncl.ac.uk/itservice/help-services/security/
Institutional Information Security Policy	https://services.ncl.ac.uk/itservice/policies/InformationSecurityPolicy-v2_1%20SJ%20v0.1%20amended%202022-08-05.pdf