An Energy Activity Dataset for Smart Homes

A flexible OCR-based energy data acquisition approach using subsample convolutional neural networks

Chen, C. L., Li
Foshan University

Meilan, M. L. C., Chen∗
Foshan University

ABSTRACT

A smart home energy dataset that records miscellaneous energy consumption data is provided. The proposed energy activity dataset (EAD) has a high data type diversity in contrast to existing load monitoring datasets. In EAD, a simple data point is labeled with the appliance, brand, and event information, whereas a complex data point has an extra application label. Several discoveries have been made on the energy consumption patterns of many appliances. Load curves of the appliances are measured when different events and applications are triggered and launched. A revised longest-common-subsequence (LCS) similarity measurement algorithm is proposed to calculate energy dataset similarities. Thus, the data quality prior information is available before training machine learning models. In addition, a subsample convolutional neural network (SCNN) is put forward. It serves as a non-intrusive optical character recognition (OCR) approach to obtain energy data directly from monitors of power meters. The link for the EAD dataset and the LED digit image dataset is: https://drive.google.com/drive/folders/1zn0V6Q8eXXKsKgcs8ZRVa1L5VEn3anD

CCS CONCEPTS

• Information Systems: • Information Retrieval; • Retrieval Tasks and Goals;

KEYWORDS

Dataset, Smart Home, Demand-side Management, Load Monitoring, Longest Common Subsequence, Convolutional Neural Network, Optical Character Recognition

ACM Reference Format:

Chen, C. L., Li and Meilan, M. L. C., Chen. 2023. An Energy Activity Dataset for Smart Homes: A flexible OCR-based energy data acquisition approach using subsample convolutional neural networks. In 2023 8th International Conference on Intelligent Information Technology (ICIIT 2023), February 24–26, 2023, Da Nang, Vietnam. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3591569.3591599

1 INTRODUCTION

Deep learning (DL) plays a vital role in smart home energy projects. State-of-the-art achievements in load identification [1], load event detection [2], and load forecasting [3] are often implemented by DL. The purpose of DL-powered energy projects lies primarily in three aspects: Firstly, a complete energy consumption profile fosters homeowners’ energy-saving habits. An energy recommender system creates optimal energy usage schedules for electricity consumers. Secondly, home automation programs based on energy utilization data and Internet of Things (IoT) technologies improve occupant comfort. Such programs include automatic cooling, heating, lighting adjustment, and home security surveillance. Thirdly, as an increasing number of intelligent devices enter millions of households, privacy and security protections have become crucial in smart home research [4]. DL-powered energy projects can monitor appliance usage data and detect malicious attacks in time, e.g., illegal cryptocurrency mining [5].

DL-powered energy projects heavily rely on energy datasets, i.e., a neural network without training data is analogous to an engine without fuel. With the evolution of hardware acceleration technologies, deeper neural networks are designed to tackle challenging load identification, disaggregation, and event detection tasks. However, complex DL models require more training data to resolve overfitting issues. Generated artificial data may not represent the real distribution, e.g., although random flipping and rotation are viable data augmentation techniques for image classification tasks, reversing time-series-based energy data is impractical. Thus, collecting authentic and labeled energy data is crucial for training complex DL models. The DL models will become increasingly accurate as new samples are obtained from the real distribution during the cumulative data collection process.

Although public energy datasets are available [6], [7], [8], [9], [10], [11], several problems exist in these datasets. First, the number of datasets in the energy sector is significantly less than the one in other research disciplines, such as computer vision and natural language processing. Second, the data type diversity of existing energy datasets is low because the appliance, brand, event, and language processing. Second, the data type diversity of existing energy datasets is low because the appliance, brand, event, and application/software labels are unavailable. For instance, there are only eleven types of appliances in the PLAID dataset [10], six of which are conspicuously distinguishable. The COOLL dataset [9] only records the on-to-off events. Third, the low sampling frequency makes capturing instant energy events almost infeasible, especially for power-on and power-off events that contain ample appliance identity information. Part of the UK-DALE [11] data is collected by 1Hz meters, which makes load identification tasks challenging. Fourth, physical quantities are incomplete in some energy datasets. For example, the PLAID dataset offers only the voltage and current data. However, voltage, current, apparent power, active power, and power factor quantities should be provided to depict an energy event comprehensively. Other energy dataset issues include 1) update discontinuation; 2) reluctance to disclose user behavior data; 3) only the aggregated energy data are available; iv) inaccuracy; iv) disproportional data points in terms of category.
Traditional energy data collection approaches are inflexible and cumbersome. For power meters without network functionalities, exporting energy data to workstations requires complicated hardware configurations and embedded programming expertise. Unfortunately, the programs for hardware-level data retrieval are not platform-independent, e.g., programs may require revision when the power meter is replaced. Besides, conventional data communication technologies, e.g., RS-232 and RS-485, are subject to environmental electromagnetic disturbance. For power meters with network functionalities, heterogeneous network communication protocols make power meter replacement and upgrade difficult. The inconvenient energy data collection procedure undermines the volume and diversity of energy datasets. Therefore, a non-intrusive solution is required to collect energy data.

This paper proposes a novel method for creating an energy activity dataset. Detailed labels, including appliance type, appliance brands, applications of appliances, and energy usage events, are prepared to analyze energy consumption patterns. The application label is available if the appliance has applications or software installed, e.g., a cell phone or laptop. With all four labels combined, an "energy activity" describes a homeowner’s behavior of using an appliance of a particular brand to perform a task, possibly through an application or software of the appliance. Besides, sub-datasets for different DL tasks can be easily created using permutation. Theoretically, \(2^4 - 1\) types of sub-datasets can be created, given that the application label exists. Otherwise, \(2^3 - 1\) kinds of sub-datasets can be created. The EAD dataset contains plenty of data points for low-powered appliances so that homeowner behaviors are better described. A revised LCS similarity measurement algorithm is proposed to calculate data point similarities of an energy dataset. This algorithm offers data quality prior information before training machine learning models.

A non-intrusive OCR-based energy data collection procedure is proposed, allowing data collection directly from the power meter monitor. Contrary to prior works on OCR-based meter readings that rely on conventional CNNs [12], [13], [14], the SCNN is designed to improve digit image recognition accuracy. An automatic correction mechanism is also proposed to fine-tune energy data according to physics rules. The rules formulate constraints for voltage, current, active power, apparent power, power factor quantities.

This paper makes the following contributions:

1. An energy activity dataset with detailed labeling and high data type diversity is publicly provided. Energy consumption patterns of applications and events are measured based on appliances from various brands. At the time this manuscript was submitted, 900 data points were collected. The dataset is provided in both JSON and MongoDB BSON formats.

2. A revised LCS similarity measurement algorithm is proposed to provide prior information on energy dataset quality before training machine learning models.

3. A non-intrusive OCR-based energy data collection procedure is put forward based on the proposed SCNN and the physics quantity auto-correction mechanism.

This paper is organized as follows: Section 2 illustrates the creation of the energy activity dataset. Section 3 describes the energy data collection procedure. Section 4 summarizes the paper.

## 2 ENERGY ACTIVITY DATASET CREATION

The energy activity dataset provides training data for DL-based energy projects. The dataset can be applied to energy conservation tasks such as energy usage profile creation and identification of energy-consuming appliances. The dataset can also be utilized by home automation applications to analyze homeowner behaviors. DL-based appliance attack detection models can be trained using the dataset because malicious activities impact energy consumption. This section offers an overview of the energy activity dataset, proposes an energy dataset similarity measurement algorithm based on LCS, and discusses energy consumption patterns of complex and simple appliances.

### 2.1 Energy Activity Dataset

Six physical quantities are provided in EAD to describe energy activities: voltage \(u\), current \(i\), apparent power \(s\), active power \(p\), power factor \((\cos \phi)\), and frequency \(f\). \(i, s, p, \text{ and } \cos \phi\) are directly influenced by energy activities, while \(u\) and \(f\) are indicators of grid stability. Each data point stores all six quantities within a short duration to describe an energy activity. The OCR sampling frequency is 5 Hz, i.e., five frames are captured by the camera in a second. Figure 1 illustrates a data point in EAD. For the rest of the paper, "power-on" and "open application" events are denoted by purple triangle symbols; "power-off" and "close application" events are denoted by purple square symbols; "level increment" and "level decrement" events are denoted by yellow triangle and square symbols; self-defined starting and ending events are denoted by lake blue triangle and square symbols.

EAD has two labeling systems: appliance-brand-application-event and appliance-brand-event. Applications can be installed on a complex appliance, e.g., computers and smartphones, while a simple appliance has no application, e.g., fans and heaters. Thus, the first labeling system is for complex appliances, and the second is for simple appliances. The appliance label represents a set of devices with similar functionalities, e.g., a refrigerator or television. The brand label combines the manufacturer name and the type of an appliance, e.g., Dell-Inspiron-15-7547. The application label denotes the software installed on an appliance, e.g., YouTube. The event label corresponds to an activity associated with human actions or automated programs, e.g., powering on a laptop or scheduled launch of anti-virus software. An energy activity is defined as a person triggering an event using an appliance of a particular brand, possibly through an application.

### 2.2 Energy Dataset Similarity Measure

Solving DL-based load identification and load event detection problems requires high-quality energy datasets. The similarity is a quality indicator of energy datasets. In supervised learning, homogeneous data points come from the same distribution, whereas heterogeneous data points originate from different distributions. In EAD, labels of all homogeneous data points are the same, while at least one label differs for any two heterogeneous data points. Consequently, homogeneous data points are expected to have a higher similarity than heterogeneous data points. A dataset containing homogeneous data points has limited research value if its similarity is too low. Likewise, a dataset with heterogeneous data points has
An Energy Activity Dataset for Smart Homes ICIIT 2023, February 24–26, 2023, Da Nang, Vietnam

The lower and upper bounds of \( \text{sim}(x, y) \) are:

\[
\text{sim}(x, y) \in \left[ \frac{\text{lcs}(x, y)}{\max \{\text{len}(x), \text{len}(y)\}}, \frac{\text{lcs}(x, y)}{\min \{\text{len}(x), \text{len}(y)\}} \right]
\] (4)

Although \( \text{sim}(x, y) \) is the de-facto LCS similarity, the upper bound of \( \text{sim}(x, y) \) denoted by \( \text{usm}(x, y) \) is more appropriate for measuring energy dataset similarities because \( \text{usm} \) is less sensitive to time series length difference, i.e., even if \( \text{len}(x) \ll \text{len}(y) \), as long as \( x \) is the sub-time-series of \( y \), then \( \text{usm}(x, y) = 1 \).

In the second step, the self-similarity matrix and cross-similarity matrix are created. In the first case, let \( N \) be the number of homogenous data points in dataset \( D \); \( x^{(i)} \) be the \( i^{th} \) data point in \( D \);

\[
\text{usm}(D) = N^{-2} ||S(N)||_1
\] (6)

The quality of \( D \) is unsatisfactory when \( \text{usm}(D) \) is too low. In the second case, the dataset \( D \) consists of several sub-datasets: \( D = D_1 \cup D_2 \cup \ldots \cup D_Q \). Each sub-dataset contains homogenous data points, where labels of all homogenous data points are the same. Let \( D_a \) and \( D_b \) be any of the two sub-datasets of \( D \); \( M \) and \( N \) be the number of data points in \( D_a \) and \( D_b \); \( x^{(i)} \) be the \( i^{th} \) data point in \( D_a \); \( y^{(j)} \) be the \( j^{th} \) data point in \( D_b \). The \( M \times N \) cross-similarity matrix \( S(M, N) \) is defined as:

\[
S(M, N) = \begin{bmatrix}
\text{usm}(x^{(1)}, y^{(1)}) & \text{usm}(x^{(1)}, y^{(2)}) & \ldots & \text{usm}(x^{(1)}, y^{(N)}) \\
\text{usm}(x^{(2)}, y^{(1)}) & \text{usm}(x^{(2)}, y^{(2)}) & \ldots & \text{usm}(x^{(2)}, y^{(N)}) \\
\vdots & \vdots & \ddots & \vdots \\
\text{usm}(x^{(M)}, y^{(1)}) & \text{usm}(x^{(M)}, y^{(2)}) & \ldots & \text{usm}(x^{(M)}, y^{(N)}) 
\end{bmatrix}
\] (7)

The similarity between \( D_a \) and \( D_b \) is:

\[
\text{usm}(D_a, D_b) = (MN)^{-1} ||S(M, N)||_1
\] (8)

The qualities of \( D_a \) and \( D_b \) are unsatisfactory when \( \text{usm}(D_a, D_b) \) is too high. The similarity of the dataset \( D \) is a \( C_2^N \) dimensional vector:

\[
\text{usm}(D) = \text{usm}(D_1, D_2), \text{usm}(D_1, D_3), \ldots, \text{usm}(D_{Q-1}, D_Q)
\] (9)

2.3 Complex Appliance Patterns

Four scenarios are discussed for the complex appliances according to Table 1.

In the first scenario, identical measurements are conducted independently. The appliance, brand, application, and event labels are the same for each measurement. According to Figure 2, highly similar energy consumption patterns are observable. Table 3 provides the dataset description and self-similarities for complex appliances.

In the second scenario, different applications are launched. The appliance, brand, and event labels are the same for each measurement. Figure 3 shows that different applications have conspicuous private energy consumption for the same appliance instance within the same time period. The quality of \( D \) is unsatisfactory when \( \text{usm}(D) \) is too high. The similarity of the dataset \( D \) is a \( C_2^N \) dimensional vector:

\[
\text{usm}(D) = \text{usm}(D_1, D_2), \text{usm}(D_1, D_3), \ldots, \text{usm}(D_{Q-1}, D_Q)
\] (9)

Little research significance if its similarity is too high. Thus, a practical similarity measurement approach is essential for evaluating energy dataset qualities. This paper proposes a two-step approach to measure energy dataset similarity: 1) measure the similarity of two time series; 2) create similarity matrices.

Figure 1: An energy activity of opening and closing the Chrome browser using a Dell-G3-3590 laptop computer.
Table 1: Energy Activity Scenarios for Complex Appliances

| Scenario | Appliance | Brand | Application | Event |
|----------|-----------|-------|-------------|-------|
| I        | Same      | Same  | Same        | Same  |
| II       | Same      | Same  | Different   | Same  |
| III      | Same      | Different | Same        | Same  |
| IV       | Same      | Same  | Same        | Different |

Figure 2: Six current time series depict energy activities of opening and closing the Chrome browser using the Dell-G3-3590 laptop.

Figure 3: Six active power time series depict energy activities of opening and closing the Chrome, Edge, Excel, OneNote, PowerPoint, and Word applications using a Dell-G3-3590 laptop.

distinctions in energy consumption patterns. Nevertheless, applications from the same product family share similar patterns, e.g., Microsoft Office applications.

Figure 4 offers an example of the cross-similarity bar chart in the second scenario, e.g., the finance application AliPay and the e-commerce application JD are only 23% similar, making them highly distinguishable.

In the third scenario, the same application is run on appliances from different brands. The appliance, application, and event labels are the same for each measurement. Figure 5 shows that hardware differences significantly impact energy consumption patterns. However, similar features in terms of peaks and valleys are still noticeable.

Figure 6 offers an example of the cross-similarity bar chart in the third scenario. All three devices are highly distinguishable.

In the fourth scenario, different events are triggered. The appliance, brand, and application labels are the same for each measurement. Figure 7 shows that each event can have a distinctive energy consumption pattern.

Figure 8 offers an example of the cross-similarity bar chart in the fourth scenario. Events such as playing videos and scanning QR codes are highly distinguishable because they consume more energy than other events.
This paper offers an OCR-based non-intrusive energy data collection method. The method is based on the proposed subsample convolutional neural network and an auto-correction mechanism. The proposed method is more convenient and less intrusive than traditional data collection approaches that require sophisticated network or hardware configurations.

### 3.1 Overall Procedure

Appliance energy consumption data are first displayed on a power meter’s monitor. The camera then captures video frames of the meter and sends them to the server. The server is responsible for three consecutive tasks: 1) recognize the energy data in image format; 2) fine-tune the energy data according to physics rules; 3) store the energy data into the EAD dataset. Figure 14 offers a graphical illustration of the energy data collection procedure.

### 3.2 Subsample Convolutional Neural Network

Inspired by the human retina structure that the fovea region has a higher resolution than its surroundings, an SCNN is proposed to extract additional features from a given sub-region of an input tensor, e.g., the central area of an image can be a sub-region of the image. In this work, energy data from the power meter monitor are recognized by SCNN. The sub-region lies at the lower right corner of a digit where a decimal point exists. Given that a **decimal point** occupies a tiny area of a digit image, SCNN improves decimal point recognition accuracy because it processes important information emphasized by the sub-region. Figure 15 shows the positions of the digit image sub-regions.

To clarify illustrations, define $C$ as both the number of filters and the number of classification categories; $D$ as the height and width of a filter; channel count, height, and width as the dimension order; $(\cdot)$ as the neural network layer index. The SCNN has three convolution layers, two fully connected layers, and a softmax regression layer. There are $C$ filters in the first convolution layer (Conv1), each of which has a dimension of $3 \times D \times D$. There are $C$ filters in the second convolution layer (Conv2), each of which has a dimension of $C \times D \times D$. Let $A^{(2)}$ with a dimension of $C \times P \times Q$ be the output tensor of Conv2; $S^{(2)}$ with a dimension of $C \times [y_kP] \times [y_wQ]$ , $y_k, y_w \in (0, 1)$ be the sub-tensor of $A^{(2)}$ and also the input tensor for the sub-sampling convolution layer (SConv), where $[x]$ is the largest integer no greater than $x$. The outputs of Conv2 and the SConv are flattened to a fully connected (FC) layer and a sub-fully connected (SFC) layer. Then FC and SFC are combined to form a softmax regression layer with $C$ outputs. Layer normalization [17] and leaky-relu activation function [18] are applied to all convolution layers. Figure 16 illustrates the architecture of the subsample CNN.

In the energy data collection scenario, hyperparameters settings are as follows: $C = 21$, where categories 1~10 represent digits 0~9, categories 11~20 represent digits with decimal points 0~9,
Figure 5: Three apparent power time series on the top and bottom rows depict energy activities of opening and closing the Word, Excel, and PowerPoint applications using a Dell-Inspiron-15-7547 laptop computer and a Surface-1 tablet computer.

Table 3: Self-similarities for Some Complex Appliances

| Appliance | Brand          | Application | Event             | Similarity |
|-----------|----------------|-------------|-------------------|------------|
| cell phone| iPhone-10      | AliPay      | open→close        | 93.34%     |
| cell phone| iPhone-10      | call        | hang-on→hang-off  | 94.83%     |
| cell phone| iPhone-10      | screen      | on→off            | 93.29%     |
| cell phone| iPhone-10      | text message| send→receive      | 91.21%     |
| cell phone| iPhone-10      | WeChat      | open→close        | 92.99%     |
| laptop    | Dell-G3-3590   | Acrobat Reader| open→close       | 94.25%     |
| laptop    | Dell-G3-3590   | Chrome      | open→close        | 92.08%     |
| laptop    | Dell-G3-3590   | PyCharm     | open→close        | 94.03%     |
| laptop    | Dell-G3-3590   | Teams       | open→close        | 96.59%     |
| laptop    | Dell-G3-3590   | Word        | open-document→close-document | 95.90% |
| laptop    | Dell-G3-3590   | Word        | save-as-start→save-as-end | 90.06% |
| laptop    | Dell-Inspiron-15-7547 | Chrome | open→close        | 87.99%     |
| laptop    | Dell-Inspiron-15-7547 | Edge   | open→close        | 89.92%     |
| laptop    | Dell-Inspiron-15-7547 | Excel  | open→close        | 88.11%     |
| laptop    | Dell-Inspiron-15-7547 | PowerPoint | open→close       | 82.50%     |
| laptop    | Dell-Inspiron-15-7547 | Word   | open→close        | 92.33%     |
| laptop    | MacBook-Air-A1466 | Chrome | open→close        | 96.32%     |
| laptop    | MacBook-Air-A1466 | Excel   | open→close        | 96.83%     |
| laptop    | MacBook-Air-A1466 | Map      | open→close        | 97.63%     |
| laptop    | MacBook-Air-A1466 | OneNote  | open→close        | 98.15%     |
| laptop    | MacBook-Air-A1466 | Word     | open→close        | 96.12%     |
| pad       | Surface-1      | Camera     | open→close        | 90.62%     |
| pad       | Surface-1      | OneDrive   | open→close        | 83.24%     |
| pad       | Surface-1      | OneNote    | open→close        | 85.53%     |
| pad       | Surface-1      | PowerPoint | open→close       | 91.96%     |
| pad       | Surface-1      | Word       | open→close        | 88.25%     |

Notes: ~ is the event transition symbol.

category 21 represents a blank area; $D = 3$; $y_h = y_w = 0.5$, where $S^{(2)}$ is the sub-tensor on the lower right corner of $A^{(2)}$. The height and width of the digit image are 57 and 42 pixels. A padding of 4 pixels is added to each digit image. When training SCNN, the learning rate is 0.01, 105,000 digit images are collected, 90% and 10% of which are used for training and testing. The test set accuracy is 100%.

SCNN is proven to be more accurate and may converge faster than a conventional convolutional neural network (CNN). Let $a^{(2)} \in A^{(2)}$ be a scalar from the subsampled region of $A^{(2)}$. Since only reshaping occurs between Conv2 and FC, there exists a unique
Table 4: Self-similarities for Some Simple Appliances

| Appliance          | Brand               | Event                                    | Similarity |
|--------------------|---------------------|------------------------------------------|------------|
| cell phone         | iPhone-10           | power-on-charging-on~power-on-charging-off| 92.32%     |
| eye massager       | Desleep-DE-F09      | power-off-charging-on~power-off-charging-off| 94.96%     |
| fan                | Midea-KYT2-25       | on-off+increase~decrease                  | 98.16%     |
| fan                | Midea-KYT2-25       | on-off+rotate~halt                        | 100.00%    |
| fluorescent lamp   | Osram-STL-T412W-03WT| on-off                                   | 97.10%     |
| hair dryer         | Desleep-DE-F09      | power-off-charging-on~power-off-charging-off| 94.96%     |
| massager           | Luyao-LY-518A       | on-off+heat~cool                          | 99.70%     |
| massager           | Luyao-LY-518A       | on-off+press~release                      | 98.48%     |
| heater             | Xianfeng-DYT-Z2     | on-off                                    | 97.58%     |
| monitor            | AOC-27B2H           | on-off                                    | 98.98%     |
| mosquito repeller  | Zhuangchen-SCJ-IC-169| on-off                                  | 93.37%     |
| pad                | Huawei Honor-X2-Gem-703L | power-off-charging-on~power-off-charging-off | 97.95%     |
| pad                | Kindle-D01100       | power-off-charging-on~power-off-charging-off | 95.89%     |
| pad                | Surface-1           | on-off                                    | 93.39%     |
| presnet            | Knorvay-N75c        | power-off-charging-on~power-off-charging-off | 99.70%     |
| razor              | Philips-RQ310       | power-off-charging-on~power-off-charging-off | 98.23%     |
| razor              | Xiaoshi-FB-BK       | power-off-charging-on~power-off-charging-off | 96.19%     |
| router             | Xiaomi-4A-1200M     | on-off                                    | 98.24%     |
| toothbrush         | Panasonic-Doltz-EW-DM71 | charging-on~charging-off               | 89.19%     |
| toothbrush         | Philips-Sonicare-HX6530 | charging-on~charging-off               | 74.45%     |

Notes: ~ is the event transition symbol, and + is the event separation symbol.

Figure 6: Cross-similarities for opening and closing the Excel application using different laptops.

When \( \frac{\partial J}{\partial a^{(2)}} > 0 \), a consensus is reached by FC and SConv regarding the output category. Thus, \( |\frac{\partial J}{\partial a^{(2)}}| \) will be larger in SCNN than in CNN. The model will converge faster.

When \( \frac{\partial J}{\partial x^{(FC)}} \frac{\partial J}{\partial x^{(SConv)}} < 0 \), a discrepancy occurs between FC and SConv regarding the output category. Therefore, \( |\frac{\partial J}{\partial a^{(2)}}| \) will be smaller in SCNN than in CNN. The likelihood of wrong classifications will be lower.

When \( \frac{\partial J}{\partial x^{(FC)}} \frac{\partial J}{\partial x^{(SConv)}} = 0 \), either \( \frac{\partial J}{\partial x^{(FC)}} \) or \( \frac{\partial J}{\partial x^{(SConv)}} \) is 0. Thus, vanishing gradient problems are less likely to happen, making the model easier to converge.

\( x^{(FC)} \) scalar in FC such that \( a^{(2)} = x^{(FC)} \). Likewise, since only sub-sampling occurs between Conv2 and SConv, there exists a unique \( x^{(SConv)} \) scalar in SConv such that \( a^{(2)} = x^{(SConv)} \). According to the chain rule, the partial derivative of the loss function \( J \) with respect to \( a^{(2)} \) is:

\[
\frac{\partial J}{\partial a^{(2)}} = \frac{\partial J}{\partial x^{(FC)}} \frac{\partial x^{(FC)}}{\partial a^{(2)}} + \frac{\partial J}{\partial x^{(SConv)}} \frac{\partial x^{(SConv)}}{\partial a^{(2)}} \tag{10}
\]
Figure 7: Six current time series depict energy activities when the Word application is used on a Dell-G3-3590 laptop: opening and closing a document, document saving, typing, image insertion, word count, and paragraph navigation.

Figure 8: Cross-similarities for triggering different events using the WeChat application on iPhone-10.

Figure 9: Six current time series depict energy activities of powering on and off a XiaoMi-4A-1200M router.

Although SCNN is only applied to energy data collection in this paper, SCNN can be applied to other computer vision tasks when sub-regions contain critical information.

3.3 Auto-correction

An auto-correction mechanism is introduced to fix potential measurement errors of power meters by making physics quantities consistent. Let the energy vector be \( \mathbf{e} = (u, i, s, p, \cos(\varphi)) \), where scalars in \( \mathbf{e} \) denote voltage, current, apparent power, active power, and power factor. The energy vector should satisfy the following constraints, where \( \epsilon_1 \) and \( \epsilon_2 \) are maximum errors allowed:

\[
|s \cdot u| < \epsilon_1 \\
|\cos \varphi - p/s| < \epsilon_2
\]
Likewise, when only the second constraint is satisfied:

$$p$$

and $$p$$ candidates exist when at least one of $$p_i$$, $$i = 1, 2$$ is assumed to be correct.

$$u_i, s_i$$ on-to-off events are triggered.

Similarly, when both constraints are incorrect, there are six energy vector candidates:

$$e_5 = (u', i', s', p', \cos \varphi'), s' = u'/s'$$
$$e_6 = (u', i', s', p', \cos \varphi'), s' = u'/s'$$
$$e_7 = (u', i', s', p', \cos \varphi'), i' = s'/u'$$
$$e_8 = (u', i', s', p', \cos \varphi'), i' = s'/u'$$
$$e_9 = (u', i', s', p', \cos \varphi'), u' = s'$$
$$e_{10} = (u', i', s', p', \cos \varphi'), u' = s'$$

(14)

The best energy vector candidate $$e_k$$ is closest to the original energy vector $$e$$:

$$\arg \min_k ||e_k - e||$$

(15)

4 CONCLUSIONS

This paper offers a public energy activity dataset EAD. EAD is designed for smart home energy research and has overcome various shortcomings of existing energy datasets, e.g., insufficient labeling and lack of data type variety. All data points in EAD are fully labeled and have a high data diversity. Specifically, the appliance, brand, and event are labels of a simple appliance, whereas a complex appliance has an additional application/software label. EAD is designed to be the data source of DL-based energy projects so that homeowner energy consumption behaviors can be better understood. This paper provides graphs of sample data points as an overview of the dataset. Besides, energy dataset similarities are offered as a dataset quality indicator which is helpful for training machine learning models. All energy data are collected using the non-intrusive OCR approach. The approach is based on SCNN and the auto-correction mechanism. In the future, EAD can be used for DL-based tasks such as energy-based cyberattack detection and intelligent demand-side management.

REFERENCES

[1] C. Li, G. Chen, G. Liang, and Z. Y. Dong, "A Novel High-Performance Deep Learning Framework for Load Recognition: Deep-Shallow Model Based on Fast Backpropagation," IEEE Transactions on Power Systems, vol. 37, no. 3, pp. 1718-1729, Sep. 2021.

[2] C. Li, G. Liang, H. Zhao, and G. Chen, "A Demand-Side Load Event Detection Algorithm Based on Wide-Deep Neural Networks and Randomized Sparse Backpropagation," Frontiers in Energy Research, vol. 9, p. 856, Dec. 2021.
Figure 12: Six apparent power time series depict energy activities when using a massager gun: power-on-to-power-off, intensity increment, heating, relax mode, refresh mode, and excite mode.

Figure 13: Cross-similarities for different events using a Luyao-LY-518A massager. ~ is the event transition symbol, and + is the event separation symbol.

Figure 14: Energy data displayed on the power meter are recognized, fine-tined, and stored in EAD. The red and green arrows represent the hardware and software processes of the energy data collection procedure.

Figure 15: The green bounding boxes are sub-regions of digit images.

Figure 16: SCNN improves recognition accuracy by processing information emphasized by the sub-region.
[3] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," IEEE Transactions on Smart Grid, vol. 10, no. 1, pp. 841-851, 2017.

[4] H. Touqeer, S. Zaman, R. Amin, M. Hussain, F. Al-Turjman, and M. Bilal, "Smart home security: challenges, issues and solutions at different IoT layers," Journal of Supercomputing, vol. 77, no. 12, pp. 14053-14089, 2021.

[5] M. Russo, N. Sendić, and P. Laskov, "Detection of illicit cryptomining using network metadata," EURASIP Journal on Information Security, no. 1, pp. 1-20, 2021.

[6] K. Anderson, A. Ormeau, D. Benitez, D. Carlson, A. Rowe, and M. Berges, "BLUED: A fully labeled public dataset for event-based non-intrusive load monitoring research," Proceedings of the 2nd KDD workshop on data mining applications in sustainability, pp. 1-5, 2012.

[7] J. Z. Kolter and M. J. Johnson, "REDD: A public data set for energy disaggregation research," Workshop on data mining applications in sustainability (SIGKDD), vol. 25, no. Citeseer, pp. 59-62, 2011.

[8] D. P. B. Renaux, F. Pottker, H. C. Ancelmo, A. E. Lazzaretti, C. R. E. Lima, R. R. Linares, E. Oroski, et al., "A dataset for non-intrusive load monitoring: Design and implementation," Energies, vol. 13, no. 29, p. 5371, 2020.

[9] T. Picon, M. N. Meziane, P. Ravier, G. Lamarque, C. Novello, J. C. L. Bunetel, and Y. Raingeaud, "COOLL: Controlled on/off loads library, a public dataset of high-sampled electrical signals for appliance identification," arXiv:1611.05803, 2016.

[10] J. Gao, S. Giri, E. C. Kara, and M. Bergès, "Plaid: a public dataset of high-resolution electrical appliance measurements for load identification research: demo abstract," Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings, pp. 198-199, 2014.

[11] J. Kelly, and W. Knottenbelt, "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes," Scientific Data, vol. 2, no. 1, pp. 1-14, 2015.

[12] K. Koščević, and M. Subalić, "Automatic visual reading of meters using deep learning," Croatian Computer Vision Workshop, pp. 1-6, 2018.

[13] M. K. Prakruthi, K. V. Vismaya, and S. Suhas, "Application of convolutional neural networks in mobile devices for inferring readings from medical apparatus," International Journal of Research and Scientific Innovation, vol. 4, pp. 2721-2765, 2017.

[14] R. Laroca, V. Barroso, M. A. Diniz, G. R. Gonçalves, W. R. Schwartz, and D. Menotti, "Convolutional neural networks for automatic meter reading," Journal of Electronic Imaging, vol. 28, no. 1, pp. 13-23, 2019.

[15] C. Cassisi, P. Montalto, M. Alotta, A. Cannata, and A. Pulvirenti, "Similarity measures and dimensionality reduction techniques for time series data mining," Advances in Data Mining Knowledge Discovery and Applications, pp. 71-96, 2012.

[16] M. Vlachos, G. Kolios, and D. Gunopulos, "Discovering similar multidimensional trajectories," Proceedings of 18th International Conference on Data Engineering, pp. 673-684, 2002.

[17] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," arXiv:1607.06450, 2016.

[18] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical Evaluation of Rectified Activations in Convolution Network," arXiv:1505.00853, 2015.

[19] M. Nijim and D. Ballampalli, "The design of a novel smart home control system using a smart grid based on edge and cloud computing," International Journal of Smart Grid and Clean Energy, vol. 11, no. 2, pp. 57-71, 2022.