FEDERAL UNIVERSITY OF TECHNOLOGY, AKURE
JOURNAL OF ENGINEERING AND ENGINEERING TECHNOLOGY

A REVIEW ON ELECTRICAL ENERGY DATASET: MEASUREMENT, FEATURES, AND APPLICATIONS

[1]Dahunsi, F. M., [2]Olawumi, A. O., [3]Sarumi, O. A., [2]Ponnle, A. A. and [2]Melodi, A .O.

[1]Department of Computer Engineering, Federal University of Technology, Akure, Nigeria
[2]Department of Electrical and Electronics Engineering, Federal University of Technology, Akure, Nigeria
[3]Department of Computer Science, Federal University of Technology, Akure, Nigeria
fmdahunsi@futa.edu.ng; olawumiabayomi@gmail.com; oasarumi@futa.edu.ng; aaponnle@futa.edu.ng; aomelodi@futa.edu.ng

Dahunsi, F. M., [2]Olawumi, A. O., [3]Sarumi, O. A., [2]Ponnle, A. A. and [2]Melodi, A .O. (2023): A Review on Electrical Energy Dataset: Measurement, Features, and Applications. *Journal of Engineering and Engineering Technology /17(1), 22-37*

Recieved Date- 05-06-22 **Accepted Date- 23-10-22**

Abstract
Electrical energy datasets provide information about industrial and residential energy use. They contain parameters such as voltage, current, real power, reactive power, apparent power, and energy which provide vital information about the electrical power supplied by the utility provider and energy consumed by the consumers. Utility companies and researchers collect energy data; data collected by the former via smart meters are mainly for billing purposes. The energy data collected by utilities is further applied in some energy management practices, such as load profiling, forecasting, and analysis. Researchers have also shown keen interest in acquiring and applying energy data to develop algorithms to aid various energy management practices. Different energy-related parameters acquired depend on the purpose in view; thus, the hardware and software architectures adopted during the data acquisition differ. This paper reviewed energy datasets to identify the various components and technologies employed during data acquisition across existing datasets. Furthermore, significant applications of energy data, such as load profiling, load forecasting, and energy theft, were discussed. These include techniques widely utilized, challenges faced during energy dataset usage, and various ways to mitigate the challenges.

Keywords: *Energy datasets, load profiling, load forecasting, smart meters, machine learning*

Introduction

The worldwide energy demand is expected to rise approximately 30% from 2015 to 2035, following the increasing worldwide economic and population growth (BP, 2021). Over the years, the continual increase in energy has resulted in the environmental and climatic impact of burning fossil fuels. This menace's ongoing and predicted effects led to the incessant urgency to regulate energy production and consumption globally (Solangi et al., 2011). Having relevant information about the power supplied and energy consumed can help make informed decisions and develop programs; hence, the idea of Load Monitoring. Load Monitoring is an approach to performing detailed energy sensing and consumption statistics of energy spent (Danilo, 2015; Zoha et al., 2012).

The energy dataset is a set of energy-related characteristics of industrial, commercial, and residential buildings with various electrical parameters such as voltage, current, real power, reactive power, apparent power, and energy. These parameters show the pattern of electricity consumption in a domain. They are either collected over a period to form historical data or collected in real-time and have been found helpful for industrial purposes, academic research, and many start-ups (Wang et al., 2019). Start-ups often attempt to collect and analyze smart meter data as it assists in making decisions on value-added services for consumers and other stakeholders.

The need to measure electricity data, coupled with the recent advances in sensing and communication technologies, led to the development of smart metering infrastructure that can collect, measure, and communicate electricity consumption (Grolinger et al., 2016). Research teams and regulatory bodies design various software and hardware architectures for data collection and monitoring. The data are usually collected through a wireless medium from consumers' houses to the backend server through the Network Interface Card (NIC) (Zoha et al., 2012). Control actions can be taken remotely based on the data collected via the communication infrastructure, whereas, in some cases, the data are processed inside the meter, and decisions are taken automatically. These sensor data's availability subsequently modified load monitoring's primary objectives, engendering numerous methodologies such as energy forecasting, load disaggregation, and many more that can revolutionize the energy industry globally.

Klemenjak et al. (2019) classified existing energy datasets into two (2) primary forms: macroscopic and microscopic datasets. The macroscopic datasets contain sampling rates around 1 Hz, while microscopic datasets are sampled at several kHz and beyond. Most energy meters often capture macroscopic datasets containing power metrics such as real power, reactive power, and Root Mean Square (RMS), while microscopic energy datasets are usually collected for research purposes (Zoha et al., 2012). In addition, the kind of energy dataset collected is influenced by sensing and reporting infrastructures (hardware and software) that make up the measuring device, which includes subjects of communication technology, storage facility, and many more. After capturing the desired data, these parameters are paired with the UNIX timestamps and are saved in various file formats such as Comma-separated values (CSV), txt, FLAC, and Hierarchical Data Format 5 (HDF5). The CSV file format is widely used to hold macroscopic data, and FLAC holds microscopic data (Zoha et al., 2012).

The acquisition of energy data plays a strategic role in energy management. Models and critical policies are formulated from the accumulated data using various mining techniques such as classification, clustering, and deep learning. Remarkable features such as changes of real power and reactive power, and harmonics of the transient waveforms can be extracted from the data acquired and used to train and test models that can be adopted for energy management practices such as disaggregation algorithms, as discussed in (Zoha et al., 2012). Having identified the possibilities of energy data, Grolinger et al. (2016) noted that algorithms learn better from bigger datasets. Therefore, the accuracy and efficiency of a developed model are a function of the quantity of data available. However, acquiring such big data can be challenging, but the significance cannot be overlooked (Wang et al., 2019).

This study aims to review the nature of the energy dataset, the factors influencing the choice of collection technologies, application areas, challenges, and mitigation methods. The rest of the paper is organized as follows: Section 2 gives a thorough explanation of the make-up of energy data- the parameters and features of energy data; Section 3 provides the typical applications of energy data; Section 4 provides a review of works of literature on existing energy data samples; Section 5 discusses the various challenges of quality data collection. Section 6 explains multiple ways to mitigate the mentioned challenges, and Section 7 presents conclusions.

Energy Data

Energy Data Parameters are electrical quantities that represent energy consumption. The two main energy data parameters are voltage and current. Every other energy data parameter, such as power, energy, total harmonic distortion (THD), and other AC power components, are derived and calculated using standard mathematical formulae (Haq and Jacobsen, 2016). Existing datasets differ in the number of parameters they provide. Some datasets contain separate traces for electrical voltages and currents. At the same time, some only report the apparent power that does not give room for robust analyses compared to those that capture and log voltage and current waveform. Moreover, some meters are configured to collect only current data and assume a constant RMS voltage, Sayed et al. (2019) noted that such limits the accuracy of measurements as it does not represent the real consumption, noting that the voltage varies right from the grid. These basic and derived parameters are further explained in this section.

Basic Measured Data: Instantaneous Current and Voltage

Current i(t) and voltage v(t) are the electrical current and voltage passing through the sensors and are measured instantaneously. These instantaneous current values vary with the loads (which can be resistive, inductive, or capacitive in make-up), and hence cannot be ignored in energy management. They are defined by Bernard (2018) as stated in Equations (1-2):

$$i(t) = \hat{i}\sin(\omega t + \varphi_i) \quad (1)$$

$$u(t) = \hat{u}\sin(\omega t + \varphi_u) \quad (2)$$

With iˆand uˆrepresenting the peak current and voltage and φ_i and φ_u. The current's phase shift and the voltage's phase shift regarding the start of the measurement.

The overall phase shift between the voltage and the current is, therefore, = φ_u- φ_i ;

2.2 Derived Parameters

These are estimated from the current i(t) and voltage v(t) measured by the sensors.

a. Root Mean Square Values Estimation

The mean values of $i(t)$ and $u(t)$ are zero in an AC power network. Therefore, the root mean square values or effective values are estimated. The Root Mean Square (RMS) value of a signal is the amplitude of a DC signal, which will transfer the same energy as the applied signal into an ohmic resistor simultaneously.

The Root Mean Square of the Current (I_{RMS}) of an AC signal is the steady direct current that dissipates an equal amount of power as the average power dissipated by the AC signal. I_{RMS} is what is usually calculated and stored in some datasets. The formula as given by Haq and Jacobsen, (2016) is presented in Equation (3):

$$I_{RMS} = \frac{\hat{i}}{\sqrt{2}} \text{ or } \sqrt{\frac{1}{T}\int_0^T i(t)^2\,dt} \quad (3)$$

where, i(t) is the current at instant t, T is the samples per cycle and $\hat{i}$ is the peak current.

Similarly, the Root Mean Square Voltage (V_{RMS}) is calculated from the measured instantaneous voltage values as stated in Equation (4):

$$V_{RMS} = \frac{\hat{v}}{\sqrt{2}} \text{ or } \sqrt{\frac{1}{T}\int_0^T v(t)^2 dt} \quad (4)$$

$v\ (t)$ = voltage at instant t, and T is the number of samples per cycle, $\hat{v}$ = Peak voltage.

b. Electrical Power Estimation

Other energy parameters include voltage, current, and root mean square values. These additional parameters give further definition to the energy consumed, allowing for a greater extent of analysis. Among the parameters are real power, apparent power, displacement power factor, and many others. These parameters are often calculated using mathematical formulae. The formulae for these parameters are derived from the primary electrical parameters.

i. Instantaneous Power p(t)

Instantaneous power is the power at one point in time. It is calculated by multiplying the instantaneous voltage and current value at this point (Equation 5). A sequence of instantaneous power values creates the oscillation of power. The letter for this power is "p (t)".

$$Instantaneous\ Power\ (p(t)) = u(t) \times i\,(t) \quad (5)$$

ii. **Apparent power (S)**

Apparent power is the product of the RMS values of voltage (V_{RMS}) and current (I_{RMS}) over the same period without reference to the phase angle, presented in Equation (6). It is measured in the unit Volts-Amps (VA). A phase-shifting or a distortion of these signals is not considered. The energy that elapsed by Apparent Power over a period is called Apparent energy; expressed as KVAhr.

$$\text{Apparent Power(S)} = V_{RMS} * I_{RMS} \quad (6)$$

iii. Active or Real power (P)

Active power is the power transferred to the load and does not return within a defined time. The active power is the average value of the power oscillation. The formula is given in Equation (7):

$$Active\ Power\ (P) =$$

$$\frac{1}{T}\int_{t=0}^{T}\big(v(t) \times i(t)\big)dt \text{ or } I_{RMS} * V_{RMS}\cos\emptyset \quad (7)$$

The sign of P is "+" or "-, depending on the direction of power flow, and cos Ø represents the power factor. Active power can only be generated by the same-frequent voltage and current components, expressed in a Watt unit. The energy corresponding to Real Power is called Real energy (P_r); expressed as Wh. Active or real energy can be absorbed or converted to other forms of energy, such as heat, motion, sound, and many more.

iv. Reactive power (Q)

This is the power that is transferred to the load and does return within a defined time. The reactive power is calculated as

presented in Equations (8 - 9)

Reactive Power (Q)=

$$I_{RMS} \times V_{RMS} \sin\emptyset \quad (8)$$

Reactive Power

$$(Q^2) = S^2 - P^2 \quad (9)$$

Although Inductive and capacitive loads dissipate zero power, the product of the voltage dropped and current drawn during the usage is represented as the Reactive Power, which is measured in a unit called Volt-Amps-Reactive (VAR). The energy corresponding to Reactive Power is called Reactive energy (Q), expressed as KVARhr. Reactive energy flows back and forth from the source to the load.

v) Displacement Power Factor (DPF): Displacement Power Factor is the power factor resulting from the phase shift between voltage and current at the fundamental line frequency. The harmonics generated are excluded during calculations. It is used to measure the efficiency of power delivery. It is calculated as the cosine of the angle between the fundamental voltage and fundamental current signals in degree (Sayed et al., 2019).

vi) Apparent Power Factor (APF): This is the ratio of real power to apparent power, considering harmonics (Equation

10). It is used to quantify a system's overall efficiency, including the effects of harmonics.

$$Apparent\ Power\ Factor =$$

$$= \frac{Total\ Real\ Power\ (W)}{Total\ Apparent\ Power\ (VA)} \quad (10)$$

Other Non-Energy-Related Parameters

Non-energy-related additional parameters are instrumental during energy data acquisition, pre-processing, and application. These parameters give more detailed information about the energy data and their patterns. Among them are individual householder demographics, data timestamps, and meters' locations. Other supplementary data, such as geographical data, meteorological effects, per capita growth, and electricity prices, are collected using energy data for sophisticated applications such as energy forecasting (Singh et al., 2012).

These parameters collected are usually stored in a separate CSV file. Discussed in this section are some standard non-energy-related parameters.

a) **Meter ID:** This unique ID is used to distinguish each meter; it provides a means of identifying and mapping datasets to their sources (Quilumba et al., 2014).

b) **Timestamps:** Note that the energy samples collected do not carry information about the time sequence. However, to analyze power consumption, the time and date each sample is collected

are significant ((Beliaeva et al., 2013). They are needed to match measurements with contextual data for analysis(Zanella et al., 2014). The time for each sample is collected as a UNIX timestamp (a discrete value). A UNIX timestamp is the number of seconds that have elapsed since the UNIX epoch, that is, the time 00:00:00 UTC on January 1 1970. These values are decoded to the date and time the data gets to the destination or point of use. A critical observation of open datasets reveals that the timestamp column occupies the first column in every data table, as seen in Makonin et al. (2018) for easy matching up contextual data collected.

c) **Location data:** The location data entails the geographical details of the data source. One of the geographical details is the Local Time Zone of the house monitored.

d) **Meteorological data:** Environmental data such as weather phenomena of energy production, temperature, and humidity are also essential when energy data is applied in areas such as energy forecasting. These data are often extracted from an online environmental data repository, as implemented in Pereira et al. (2014).

e) **Demographics:** This data record describes the household where the meter is installed Pereira et al. (2014). These data include the period at which the user consumption data is available, the family size, the type of residence, and many more.

2.4 Features of Energy Data

Energy data collected by utilities or researchers using smart meters vary in features. These differentiations have thus prompted various data-gathering campaigns. These features include the quality, size, sampling frequency, number of data points captured, the method of collection, and many other unique properties that qualify an energy dataset.

2.4.1 Size

Existing energy datasets vary in size for different reasons. Makonin et al. (2018), who intended to carry out Non-Intrusive Load Monitoring (NILM), prioritized the need for accuracy and reliability; hence, the authors captured 11 electrical parameters from 2 houses at 1Hz, which resulted in 11.3 million power readings. The UK-DALE dataset, presently considered one of the most enormous energy time series datasets, with approximately half a billion records data, captured the raw ADC and stored these readings locally in the UK-DALE; hence, requiring 4.8GBytes per day (Kelly and Knottenbelt, 2015). Therefore, the size of energy data is a function of the sampling frequency, the number of parameters collected, and the number of households considered –primarily influenced by the intended applications (Shin, Lee, et al., 2019). Conversely, a small data size is required for basic applications such as billing.

2.4.2 Number of Parameters

The number of parameters measured at every point of energy data collection is also a function of the collection purposes, i.e., the application in views such as household applications, industrial applications, research applications, and utility applications (Sayed et al., 2019). Traditionally, energy meters only record household electricity usage by actual power consumption. However, Pitì et al. (2017) noted that the new generation of smart meters is now collecting more than one physical quantity. For salient applications such as Non-Intrusive Load Monitoring (NILM), occupancy detection, appliance usage modelling, and many more, many parameters for the whole house and appliance-by-appliance demand are required. The GREEND and HES datasets captured only the active power (Monacchi et al., 2015), while the likes of the RAE dataset and REDD dataset captured 11 electrical parameters and four electrical parameters for more sensitive applications, respectively (Kolter and Johnson, 2011; Makonin et al., 2013).

2.4.3 Method of Collection

Energy data are collected by developing a system that houses hardware, communication, compression, etc. In most cases, the underlying primary electrical

parameters (current and voltage) are collected using current and voltage sensor transducers. The connection between the sensing circuit and the terminal to be measured can be either invasive or non-invasive. A sensor is invasive if it requires physical contact with the household circuit to take measurements. In contrast, it is non-invasive if it does not need physical contact with the electrical terminal in the household for measurement. The ZMPT101B (voltage sensor) and ACS712 (current sensor) are prominient invasive transducers requiring a direct connection with the main circuit. The conventional current sensors often used are the Current Transformer, Rogowski coil, Hall Effect Sensors, and Shunt (Haq and Jacobsen, 2016). Among the current sensors, the Current Transformer is the most used among existing open datasets as it requires no external power supply (Pereira et al., 2014; Kelly and Knottenbelt, 2015; Kahl et al., 2016). Many datasets, such as the UK DALE, and SustData dataset, have employed the 'YHDC SCT-013 000' - a current transformer clamp because it is non-invasive as it provides a more convenient and safe way of measuring (Pereira et al., 2014; Kelly and Knottenbelt, 2015).

Moreover, the standard voltage measurement methods are voltage dividers, optocouplers, and voltage transformers (AC-AC adapters). The voltage transformer is the most widely used; it captures the instantaneous voltage of a line. The method was adopted while collecting UK-DALE and WHITED datasets (Kelly and Knottenbelt, 2015; Kahl et al., 2016). However, the current transducers vary in performance, response time, accuracy, sensitivity, and resolution (Sayed et al., 2019).

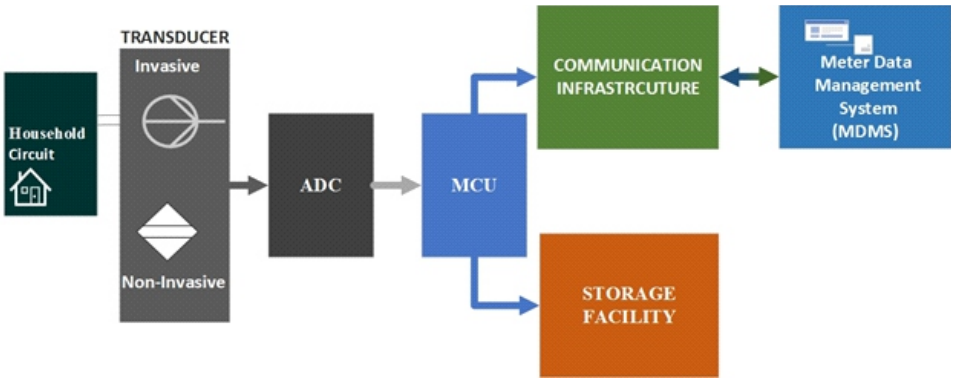

Figure 1: Functional Block Diagram of Energy Data Collection

The measurements from the sensors are passed to an Analog Digital Conversion(ADC) module that does the data digitization at a given rate, as seen in Figure 1 ((Rodrigues et al., 2017; Sayed et al., 2019), the sampling frequency is determined by the number of channels present and configured in this module. The Microcontroller Unit further processes discrete values obtained (MCU). The MCU provides interfaces to connect to the transducers and can accommodate various communication protocols(Sayed et al., 2019). Many works, such as Kelly and Knottenbelt (2015) and Monacchi et al. (2015), used ARM-based platforms (e.g., Raspberry Pi or BeagleBone) and Nanode, respectively. Moreover, derived electrical parameters such as power consumption computation are carried out on the MCU from the processed voltage and current signals (Lua et al., 2009; Sayed et al., 2019).

Required data are then compressed to a format that can be stored in the device's internal memory or transmitted through a communication module. However, some applications display relevant measured parameters on the meter's screen (Sayed et al., 2019). The communication technologies via which the acquired data can be collected and controlled are Zig-Bee, Power Line Communication (PLC), Z-wave, and Wi-Fi. Comparing them, Lua et al.(2009) and Rodrigues et al. (2017) noted that the ZigBee wireless communication technology is the most commonly used network in Advanced Metering Infrastructure. For many smart meter applications, the total energy consumed in the household is calculated at the point of collection using some mathematical formulae and transmitted to a central point. In the case of highly sampled data, the parameters are usually down-sampled before being transmitted or stored (Kolter and Johnson, 2011).

However, in capturing appliance load demand for applications such as NILM, an auxiliary device (a plug with a wireless transmitter) is often employed to capture and transmit the active power consumed by the home appliances (Shin, Lee et al., 2019). This application is found in Kelly and Knottenbelt (2015) and Shin et al.(2019), where an ENERTALK plug and an Ecomanger Transmitter Plug were adopted, respectively.

2.4.4 Frequency of Data collection

According to Lu et al. (2012), a smart meter collects data by the minute, while the old mechanical meter

collects data hourly or monthly because the meter only needs the aggregated household demand over a period. The frequency can be increased for high resolution when the purpose of data from the meter transcends billing to handling insightful applications (Shin, Rho, et al., 2019). Smart meters with high sampling resolution are already found in many new generations of smart meters installed in Italy (Pitì et al., 2017). Highly sampled data is essential for applications such as NILM. Shin, Rho, et al. (2019) study on data sampling rate showed that at least a 1-3Hz sampling rate is required to prevent NILM performance from deteriorating, noting that the signatures are destroyed when the sampling rate is too low (Shin, Rho, et al., 2019). The data was highly sampled in Kelly and Knottenbelt (2015) as signals were sampled at 1Hz,6Hz, and 44.1kHz.

2.4.5 Storage

Data collected are either saved locally at the point of generation or transmitted to a central server; the storage facility depends on the system's architecture—the collected data is written on files and saved in different formats. Kelly and Knottenbelt (2015) compressed the highly sampled raw ADC data using FLAC and saved to a disk, while the active power, apparent power, and RMS voltage were saved to a CSV file once every second. The CSV file format is the most common format widely adopted by many existing datasets (Kelly and Knottenbelt, 2015). However, Sophisticated file formats such as The Hierarchical Data Format 5 (HDF5), WAVE (Pereira et al., 2022), and FLAC (Kelly and Knottenbelt, 2015) found in some datasets are mainly used to store microscopic data. Klemenjak et al. (2019) noted that the HDF5 is more viable among the existing file formats as it supports metadata annotations, efficient data storage, data transformations, and libraries for most scientific computing frameworks. The 15Hz power readings collected by the smart meters built by Encored Technologies (a private organization that provides energy AI services to customers) were sent to cloud data collection servers via SSL/TCP. The ENERTALK dataset was collected from various houses where the metering devices were installed (Shin, Lee, et al., 2019). Some works, such as Monacchi et al. (2015), employed the two approaches, saving the data acquired locally on a CSV file daily and remotely on cloud storage (a MySQL server in the cloud). In the case of Smart Meters, Pitì et al. (2017) noted that electrical parameters are also stored in Smart Meter Registers for many days to cope with the possibility of unavailability of the communications channel.

3. Application of Energy Data

Due to the possibility of collecting energy at high frequency, the majority of the recent applications of energy data now employ machine learning techniques such as Inferring Rules, Statistical Modeling and Naıve Bayes, K-Means, Support Vector Machines, and many more. These techniques automatically learn the energy users' behaviour using the acquired consumption data. These techniques can either be supervised learning or non-supervised learning. Supervised learning (classification) uses labelled data to assign objects to classes. Unsupervised learning uses unlabeled data to group instances into classes.

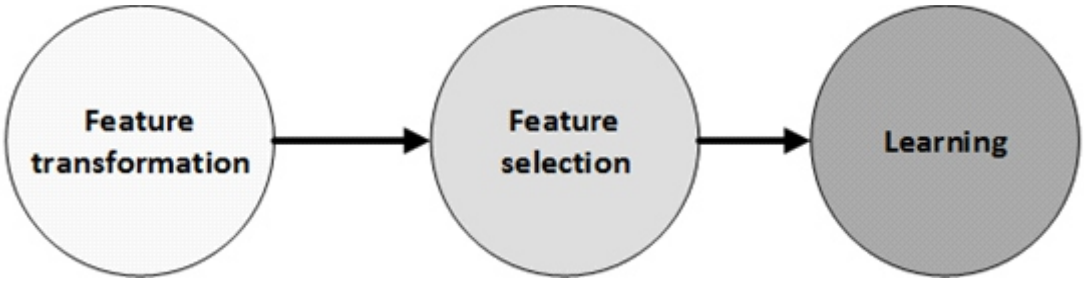

Figure 2: Classical Machine Learning Process

The collected parameters are pre-processed and filtered from outliers. Afterward, features are extracted from the pre-processed data and further subjected to learning for any application (Bernard, 2018). According to Dahunsi et al. (2021), data processing involves data cleaning, transformation, and integration; it addresses some underlying issues, such as missing data in a data frame (Quilumba et al., 2014). Intensive pre-processing improves the quality of energy data. Depending on their impacts on the application, the processed data stream is further streamlined to some key features.

In some cases, the extracted features are normalized and scaled before subjection to learning to reduce the impact of variance on the training (Bernard, 2018; Dahunsi et al., 2021). The hidden pattern revealed through these intelligent techniques on energy parameters is usually maximized for an amalgam of applications such as load profiling, load forecasting, and many more. For instance, Encored, a private company, uses energy data to provide energy artificial intelligent services such as forecasting, diagnostics, and many more to end-users (Shin, Rho et al., 2019).

3.1 Load Profiling

Load profiling refers to estimating the total energy consumed by end-users over a period. Wang et al.(2019) defined it as the classification of load curves according to consumers' behaviours. Having an insight into the consumption behaviour of energy consumers is pivotal in developing pertinent policies and programs that are cogent to managing energy consumption, power system operations, and many more. For instance, in Tariff formulation, one of the applications of load profiling, references are made to specific customer categories, and technical and economic attributes are used to design Tariff plans (Chicco, 2012).

3.1.1 Individual Load Profiling

The traditional energy meters record household electricity usage for billing. However, Lu et al. (2012) opined that close monitoring is needed to be aware of abnormalities in energy consumption. Energy consumed can be collected daily or hourly as opposed to the usual practice of the existing meters. The various signatures of the Individual Load Profile are explained below.

a) **Monthly Household Energy Consumption Profile**: This is the usual practice of many utilities; this profile can be visualized by plotting the monthly demand against the months in the year. This monthly data reduces the expected communication between the meters and the control center Lu et al. (2012) and is limited to only billing. Utility companies are not informed about abnormalities in the energy supplied.

b) **Daily Household Energy Consumption Profile**: This approach embraces the collection of energy consumed daily, having a data point for each day; a curve generated from this is called the daily load curve. The daily load consumption pattern is employed in the Home Energy Management System, which gives complete information about electricity usage to consumers; this can thus be used for better controlling their consumption. At this stage, from the utility end, the time series data of the energy consumed begins to infringe on the consumers' privacy; in such case, Lu et al. (2012) recommended the use of statistical tools that hide time information, such as mean the utilities can use the standard deviation to detect abnormalities.

c) **Hourly Household Energy Consumption Profile**: The hourly load pattern generates 24 data points daily, capturing the energy consumed in households every hour. This form of profiling gives more detailed energy consumption than daily energy consumption. However, customers' privacy remains a concern Lu et al. (2012).

3.1.2 Load Pattern Categorization

Load pattern categorization is one of the techniques utilities adopt, having obtained power consumption of electricity supplied to consumers. It is the classification of consumers based on the similarity of their different power consumption behaviours. Utilities use this to develop various programs such as payment programs; the end consumer can choose suitable payment programs offered by the utilities based on the categories (customer classes) created from Load pattern categorization. Some of the applications of load classification include Bad data identification, Load forecasting based on load classification, and Tariff setting based on load classification(Zhou et al., 2013).
Chicco (2012) opined that load pattern categorization based on only aggregated residential load data (considering residential consumers as individual entities) is inefficient for several applications except for billing purposes. This assertion was established because the consumption pattern of individual residential loads varies with many dynamic factors, such as the number, activity, age, and lifestyle of occupants. He noted that detailed statistical analyses based on several factors would need to be carried out (Chicco, 2012).

However, the massive amount of information gathered from an individual customer being monitored by smart meters can be overwhelming, requiring extensive communication channels and memory storage and placing a high computational demand on the hardware and software infrastructures. However, a smart meter with high resolution infers that only a limited number of customers need to be monitored for effective load pattern categorization (Chicco, 2012).

Jiang et al. (2017) noted that categorizing consumers based on load patterns without further analysis is inefficient; therefore, their research proposed two-stage clustering and category identification for Load Pattern Categorization. Clustering is a more efficient and intelligent approach and the most adopted method lately. It is one of the several methods used to group houses of similar energy use patterns together for monitoring, modelling, or control purposes. Clustering techniques are classified into direct and indirect clustering (Wang et al., 2019).

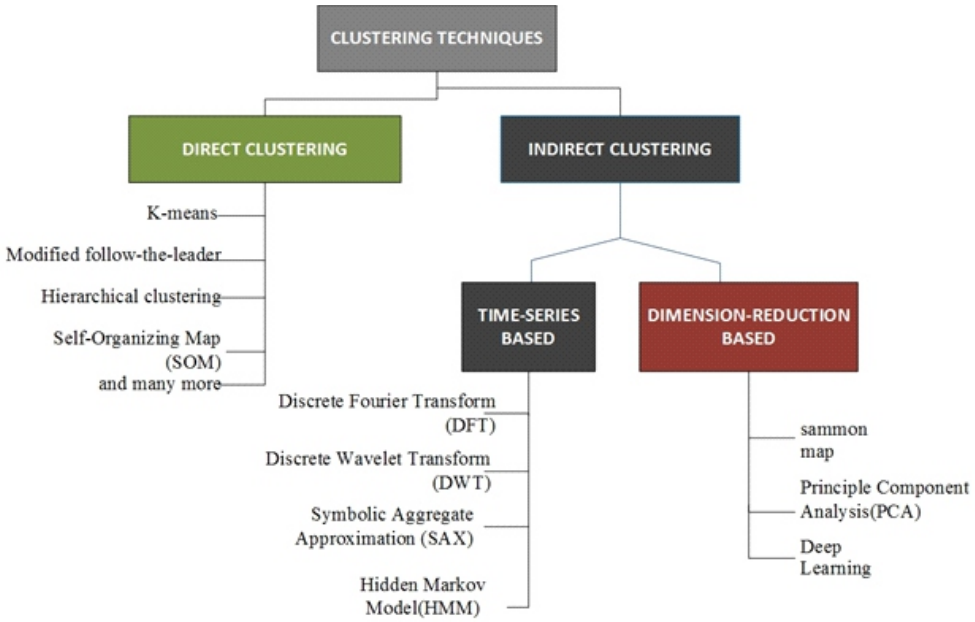

Figure 3: Clustering Techniques for Load Profiling

According to Meshram et al. (2012), some processes occur before load curve clustering, including data selection, reduction, and pre-processing. It was observed that different normalization methods tend to give different clustering results. Therefore, the methods for these processes must be wisely chosen. Moreover, Zhou et al. (2013) added that a clustering preparation stage includes determining the classification characteristics, selecting an appropriate clustering algorithm, and determining its corresponding parameters.

a. Direct Clustering

Direct clustering refers to applying the clustering method to the data collected by smart meters. Among the various clustering techniques are: K-means, modified follow-the-leader, hierarchical clustering, and self-organizing map (SOM) are found to be the most used. In research carried out by (Chicco, 2012), the commonly used clustering techniques were tested on 234 non-residential customers connected to a medium voltage distribution system. The clustering algorithms were run 96 times. Modified follow-the-leader and hierarchical clustering techniques were proven to be the most accurate.

Wang et al. (2015) noted some fundamental issues with direct clustering. The need for high-resolution data from smart meters is among many. From an experiment carried out by (Granell et al., 2014), it was observed that for accurate results to be obtained from any direct clustering methods, smart meter data sampled for at least every 30 minutes is required, which thus results in computational complexity on the system.

b. Indirect Clustering

Indirect clustering is an approach in which dimension-reduction techniques or other methods process the load data before clustering (Wang et al., 2015), i.e., feature extraction is conducted on the electricity consumption data before clustering. Feature extraction scales down the input data while maintaining or improving its quality. Indirect clustering can be either dimension reduction-based clustering or time-series-based clustering. The motivation behind dimension-reduction-based clustering is that the daily electrical consumption of each household can be represented by artificial variables in smaller sets-reducing the dimensionality of smart meter data. Principle Component Analysis (PCA) and Sammon map are dimension-reduction-based clustering methods.

An example was seen in Koivisto et al. (2013), where PCA was used to reduce the original dimension of the data to a smaller set of artificial variables that are used as the input variables for K-means clustering. The Time Series based clustering involves the use of time-series analytical methods such as Discrete Fourier Transform (DFT), Discrete Wavelet Transform (DWT), and Symbolic Aggregate Approximation (SAX) to extract features. Wang et al. (2019) believe that features extracted effectively before clustering improves the performance and efficiency of load profiling.

3.2 Energy Demand Forecasting

Consumers can place unusual load demands on facilities, which can impede some dangers, leading to colossal damage that can devastate the power grid if the utilities are unprepared for this. It is, therefore, essential

to have a foreknowledge of future demands. Energy demand forecasting is a technique used to predict future energy demand. Knowledge of future energy assists concerned regulatory bodies in making necessary decisions such as energy policy formulation. It also helps electric utilities plan systems, energy trading, electricity prices, load switching scheduling, demand response, and infrastructure development (Campillo et al., 2012).

Load Forecasting techniques are widely studied using various time series techniques and regression algorithms, including Auto-Regressive Integrated Moving Average (ARIMA) and Neural Networks (Bansal et al., 2015). In the early days, forecasts were carried out using traditional and conventional mathematical techniques such as regression, multiple regression, exponential smoothing, and Iterative reweighted least-squares approach (Singh et al., 2012). Due to advancements in the computational power of systems, the non-linearity of data can now be analyzed using advanced techniques like neural networks. Zheng et al. (2018) discussed the significance of rich data and opined that a forecasting model would have higher accuracy with a more extended period of historic electricity data. The historical load, weather, and event data are essential parameters for carrying out load forecasting (Gupta, 2017). Generally, Load forecasting can be classified into three (3) groups, short-term forecasting, medium-term forecasting, and long-term forecasting based on duration.

a) **Short-term forecasting:** The short-term forecasts are usually one hour to one week. It is helpful for essential operating functions such as energy transactions, unit commitment, and security analysis (Singh et al., 2012). Some widely adopted short-term forecasting methods are the Similar Day Lookup Approach, Regression-based Approach, Time Series Analysis, Support Vector Machine (SVM), Artificial Neural Networks, and Fuzzy Logic (Ramos et al., 2013; Gupta, 2017; Aurangzeb and Alhussein, 2020). The SVM and similar day approach are the main methods in this group. Guo et al. (2006) showed that SVM has superior performance over ANN as it can capture nonlinear mapping more easily than ANN, and the SVM model performs structural risk minimization rather than minimizing the training errors. However, Aurangzeb and Alhussein (2020), chose ANN over SVM due to its high computational complexity and large memory space requirement but did not juxtapose them on the data used.

b) **Medium-term forecasting:** Medium-term forecasts are usually from a week to a year; utilities use them or schedule fuel supplies and unit maintenance (Gupta, 2017).

c) **Long-term forecasting:** Long-term forecasts span over a year (Singh et al., 2012). It is used to supply electric utility company management to predict future needs for expansion, equipment purchases, or staff hiring (Gupta, 2017). This form of forecasting includes many relevant factors such as consumer behavior, the impact of adopting new technologies, descriptions of appliances used by customers, the sizes of the houses, the age of equipment, and population changes. The end-use method analyses the impact of different end-user devices on the overall energy consumed. The end-use method requires more information from consumers than short-term forecasting. As found in articles, Multiple Linear Regressions, Neural Network techniques, and Trend Analysis are popular techniques used in long-term forecasting (Gupta, 2017; Javeri et al., 2021).

3.3 Energy Disaggregation

Energy disaggregation or Non-Intrusive Load Monitoring (NILM) - a technology for separating a household's aggregate electricity consumption information, has been an active area of research. A recent trend in Load Monitoring has received much attention in the research world. It uses high-resolution energy data to develop a model that disaggregates and identifies each constituent appliance from the cumulative load.

NILM is an attractive solution; the energy consumption behaviour of customers can be modified by providing feedback. Information obtained from the disaggregation is used as feedback to the users for improving energy efficiency (Shin et al., 2019); NILM finds its applications in areas such as Energy Cost-saving, Smart Home, House Categorization, and Life coaching. Electrical appliances generate distinct power consumption patterns called signatures (Kim et al., 2017). G. W. Hart pioneered energy disaggregation, and classified the appliance model into three types: On/Off, Finite State Machine (FSM), and Continuously Variable based on the distinct signatures of various appliances. The concept of NILM involves the identification of various appliances in a household using these signatures.

To implement NILM, various techniques in Machine Learning have been suggested, ranging from SVM, Bayesian networks, and Hidden Markov Models (HMM) to ANN. Most works were carried out using supervised methods (Bernard, 2018). Few have done NILM using unsupervised methods such as K-means

and agglomerative clustering (Kelly and Knottenbelt, 2015). These techniques involve the training and testing of models. The accuracy and efficiency of a model are a function of the quality and quantity of the available data; the need for quality data is central in Energy disaggregation. A microscopic dataset is required for efficient energy disaggregation, requiring a high sampling rate (Zeifman and Roth, 2011); therefore, many researchers have critically analyzed existing datasets. Some datasets have been proven to be lacking under one particular algorithm and rich under another and vice versa. However, NILM is very limited due to the high cost of communicating and storing the data.

3.4 Energy Theft Detection

Energy theft, also called a Non-Technical Loss (NTL), is a significant concern globally, most notably in developing countries (Jiang et al., 2014). Consumers use various means to steal power; the most common approach is bypassing the meter such that the current flowing into the house does not go through the meter. Before the advent of smart meters, detecting NTL had been difficult, considering the granularity of the data that will be needed (Buzau et al., 2019). There are various energy-theft detection schemes. Jiang et al. (2014) classified the various techniques into three categories: classification-based, state estimation-based, and game theory-based. The state estimation method utilizes monitoring state through mutual inspection; the cost of implementing this approach gives the utility companies drawbacks involves the construction of the monitoring system, which includes peripherals such as wireless sensor networks, control units, and radio frequency identification (RFID) (Jiang et al., 2014; Yip et al., 2017).

The classification-based detection scheme is the most common among the three. It involves carrying out a load profile analysis of customers from historical data to detect abnormal energy consumption patterns(Sahoo et al., 2015). The procedure involved in classification-based energy-theft detection consists of seven parts: data acquisition, data pre-processing, feature extraction, classifier training, parameter optimization, classification, data post-processing, and suspected customer list generation (R. Jiang et al., 2014). Some classification-based technologies are Support Vector Machines (SVM), Fuzzy classification, Neural networks, AutoRegressive Moving Average-Generalized Likelihood Ratio (ARMA-GLR) detectors, and P2P computing. In recent research by Buzau et al. (2019), the authors showed that extreme gradient boosting, a gradient boosting classifier (GBC), is the most efficient technique among all other Machine Learning techniques for non-technical loss (NTL) detection.

4. Energy Data Samples in Literature

The importance of data to energy management provoked researchers to come up with various energy datasets. Wang et al. (2019) noted that many power companies are reluctant to release their smart meter data to the public due to privacy and security. However, some unnamed datasets from some selected households have been released over the years. Some are made public, while others are granted to the public.

Existing datasets have various unique purposes; some are carried out with the NILM method, while some are obtained with load profiling. Some datasets are gathered for research purposes focusing on statistical signal processing and blind source separation, energy use behaviour, eco-feedback, and eco-visualization demand forecasting. Other datasets' applications include smart home frameworks, grid distribution analysis, time-series data analysis, energy-efficiency studies, occupancy detection, energy policy, socio-economic frameworks, and advanced metering infrastructure (AMI) analytics (Makonin et al., 2018). Some energy big data analytics companies, such as ENERTALK, have published articles from private datasets acquired from houses in Korea for disaggregation purposes but made the data inaccessible to the public(Shin, Lee, et al., 2019).

4.1 Open Datasets

Various researchers and organizations have written many articles on different applications of energy datasets MIT released the Referencing Energy Disaggregation Data Set (REDD) in 2011 (Kolter and Johnson, 2011). The dataset captured for NILM consists of data sampled at a very high frequency compared to the one with load profiling and some other less complicated applications such as the UK-DALE dataset (some data sampled at 16kHz), the REDD dataset, COOLL, and the PecanStreet dataset (Kelly and Knottenbelt, 2015; Kolter and Johnson, 2011; Picon et al., 2016) Moreover, data describing the aggregate demand per building and individual appliance demand are needed (Kelly and Knottenbelt, 2015). This has brought some disparities in the various open datasets. However, many datasets have been released to the public. Some of the open datasets are summarized in Table 1. From the table, the discrepancies observed revolve around the sampling rate, duration, number of houses considered, features sampled, and many more. The acronym "agg" on the ground truth column signifies aggregate demand, "the main." The "sub" indicates the sub-meters or each breaker on the house's main power panel. Also, most works collected the consumption data at the appliance level, indicated as "app" in Table 1. The current and voltage were measured in many cases before deriving other parameters; however, in datasets like Enertalk, the power readings were taken directly using a device. Furthermore, datasets such as AMPds included other measurements – gas and water consumption- and energy consumption data. (Makonin et al., 2013; Pereira et al., 2014).

5. Challenges of Quality Data Collection

The quality of energy data is the quantitative and qualitative state of the energy data (Shin, Rho, et al., 2019). Energy data is said to be of high quality if it depicts the real-world energy consumption of end-users after being acquired correctly. Quality data can only be captured when efficient sensing infrastructures; communication technologies are employed during the data collection.

Table 1: Existing Datasets Summary (Tjaden et al., 2015)

Dataset	Sampling rate	Duration	Houses/ device instances	Ground truth	Country	Availability
REDD	15000 Hz (agg) 3 sec.(app)	Several months	5	agg and sub channels	USA	Available (on request)
BLUED	12000 Hz (agg)	8 days	1	agg	USA	Available (on request)
UK-DALE	16000 Hz (agg) 1Hz (agg) 6 sec.(app)	1569 days	6 (of which 3 are at 16 kHz)	agg and sub channels (P)	UK	Public
SustData	1 minute (agg)	1144 days	50	agg	Portugal	Available (on request)
ENERTALK	15Hz (agg)	29-122days	22	agg and sub channels	Korea	Public
WHITED	44000 Hz (app)	5 seconds	9		Multiple	Public
AMPds	1 minute (agg and app)	2 years	2	agg and sub. channels	Canada	Public
RAE	1 Hz (agg and app)	72 days	1	agg and sub channels	Canada	Public
HES	2 minutes (app)	1 year / 1 month	251		UK	Available (on request)
ECO	1 second	8 months	6 / 45	agg and sub channels	Switzerland	Public
GREEND	1Hz (agg and app)	One year	9	agg	Italy and Austria	Public
COOLL	100 kHz (app)	6 seconds	1		France	Public

sub *– submeter channels,* ***agg*** *– aggregate consumption data,* ***app*** *– individual appliance consumption*

The generation and communication of data place a cost demand on the hardware and software infrastructures because the deployment has to be on a larger scale (Shin et al., 2019). This constraint has limited the quality and quantity of datasets. It has been observed that there has always been a trade-off during data collection (Makonin et al., 2013). Some datasets considered a high number of houses at the expense of the sampling frequency. The reverse is regarded for some datasets. However, this compromise in the choice of infrastructure has led to a low-quality dataset. Discussed in this section are a few of the challenges.

5.1 Communication Infrastructure

The data is often collected via wireless communication protocol, as in smart meters; however, this limits the quality and quantity of data obtained. Examples of energy monitoring systems that have employed wireless communication protocol are the Zigbee-based smart energy monitoring system and IoT-based smart energy monitoring system. The Zigbee communication protocol can obtain energy parameters within 10m to 100m, with a maximum transferring speed of just 250 kbps (Govindarajan et al., 2019). However, the IoT-based Energy Monitoring System on the existing internet infrastructure provides two-way communication, but Kelly and Knottenbelt (2015) noted that some data packets are occasionally lost in transmission in wireless meters. Also, Quilumba et al. (2014) noted that since no communication network is perfect, AMI deployment must deal with dropped or delayed data, duplicate data, and many more. Hence, Marinakis (2020) and Zanella et al. (2014) noted that utilities and researchers, therefore, resolved to keep generated data as much as possible closer to the generation point and the owner, as seen in Kelly and Knottenbelt (2015).

5.2 Sampling Frequency

According to Carrie Armel et al. (2013), the sampling rate is the most critical factor in some applications such as NILM. The research affirmed this assertion by Shin, Rho, et al. (2019), where the sampling rate varied from 10Hz to 0.1Hz on a TV, washer, and rice cooker signal deduced from the ENERTALK dataset. The deduced pulse shape began to distort as the frequency reduced,

losing some signal values. However, a sampling frequency of 1Hz is enough for applications such as energy demand forecasting since aggregates household energy demand is the significant parameter considered in such an application. Moreover, the new features that characterize smart meters require high-resolution data (sampling at a high rate) to extract relevant information from the meters Palacios-Garcia et al. (2017).

5.3 Limited Storage Facilities

Real-time and historical data are crucial when extracting meaningful information to make data-driven decisions. Hence, due to the frequency and the increasing number of customers, the scale and size of the data collected get more prominent, thus resulting in an overwhelming volume of data, hence an extensive storage system. As stated earlier in this review, the application and usefulness of data lie in its quality and enormity; having more is an advantage to any application. Although different compression algorithms have been developed to reduce the granularity of the data, reducing the size has an effect, as discussed by (Wang et al., 2019). During the gathering of the UK-DALE dataset, the data gathered were so much, having more than eight million rows of data (Kelly and Knottenbelt, 2015). The team had to improvise and address this by collecting some data on-site locally. Considering the RAE dataset, 11.3 million power readings were captured at 1Hz (Makonin et al., 2018). The authors opined that they settled for such low frequency because of storage and processor constraint.

5.4 Data Privacy and Security

Data privacy and security have been significant hindrances to data collection; smart meter users are generally skeptical about the idea of data collection, having many concerns about revealing individual data information (Wang et al., 2019). They are generally concerned about their private information being referred from the frequently collected monitoring information, fearing that private information may be sold to marketing companies. Even criminals may use this information to commit crimes (Jiang et al., 2014). However, the utility needs energy data for quality service delivery. A framework for the trade-off between privacy and utility requirements of consumers was presented by Sankar et al. (2013) based on a Hidden Markov Model. Eibl and Engel (2015) adopted one method: aggregation of individual smart meter data and coloured noise. However, further analysis showed that the method would have a massive impact on the detection algorithm called Edge Detection used to identify the status of appliances in NILM. Therefore, privacy is still a concern despite the advancement to measure at high resolution and energy data's promising future and possibilities.

5.4 Big Data Issues

As established in this review, the parameters needed transcend electrical parameters to use data in load forecasting maximally and in some applications. Parameters such as financial information, meteorological data, and many more are usually essential to the meaning of energy data (Gupta, 2017). However, this has consequences, including multivariate data fusion and high-performance computing issues.

a) **Multivariate Data Fusion:** The varying data characteristics bring complexities to data analytics. However, this provides more extensive data, but working with structured and unstructured data complicates the analysis.

b) **High-Performance Computing:** Big data calls for higher processing power; this thus calls for highly efficient algorithms such as parallel computing-demanding heavy computing systems such as distributed computing, fog computing, and GPU computing (Wang et al., 2019; Yikuai Wang et al., 2018).

6. Mitigating the Challenges of Quality data collection

a) **Data Loss:** One of the issues highlighted with collecting energy data at high frequency is the limitation around the communication infrastructure. There will be an increase in missing data arising from sensor faults, technical issues, instability in internet connection, or malfunctions from any electronic components in a smart meter. With Kelly and Knottenbelt (2015) proving their experience of data loss, it is crucial to have a way to mitigate this challenge. One standard method found across the numerous articles was setting up a "fail-over" system. Aside from collecting the data through wireless media, researchers came up with the idea of having a secondary storage system on the hardware. Even if there is a data loss due to communication system failure, the data logging continues on-site (Kelly and Knottenbelt, 2015; Shin, Lee, et al., 2019). In Zanella et al. (2014), data collected are stored locally and backed up in the cloud. During the post-processing of the data, the two data sets can be compared and fixed up.

b) Furthermore, the missing data can be filled up using different techniques if the backup

cannot makeup. As noted by Dahunsi et al. (2021), a statistical method such as an autoregressive integrated moving average (ARIMA) and linear interpolation (LI) model may be used. Conversely, machine learning methods such as K- Nearest Neighbour (KNN), multilayer perceptron (MLP), and support vector regression (SVR) may also be used (Wang et al., 2021).

c) **Data Privacy:** The privacy of the consumers remains very important. Lu *et al.* (2012) recommended statistical representations that conceal time data/information because using robust encryption algorithms like Advanced Encryption Standard (AES) can be burdensome for the hardware and software systems. The methodology of masking consumption data can be adapted to hide the identity of energy consumers. Efthymiou and Kalogridis (2010) successfully masked smart meter data using an escrow service over a smart grid network. Salehkalaibar *et al*. (2019) used battery storage charging and discharging schedules to mask the real electricity consumption behaviour and address privacy concerns. Therefore, basic stochastic algorithms can be used Salehkalaibar *et al*. (2019) and Tan *et al*. (2013)..

7. Conclusions

The energy dataset comprises parameters, each with a unique definition and inference; utilities and researchers capture these parameters for different purposes, which consequently influence the choice of energy parameters, and the hardware and software infrastructures developed. Energy datasets can be applied in energy management practices such as load forecasting, load profiling, and energy-theft detection. The various applications of energy data considered in this paper revealed the importance and impact of high-resolution data. The review also showed the challenges around collecting energy data at high resolution despite its promising potential. Researchers have developed novel algorithms to address privacy, security, data loss, and limited storage facilities identifiable with high-resolution energy data. However, some of these solutions are yet to be deployed in practical scenarios. While the discovery of solutions continues how to abate these challenges ultimately, it is recommended from the study that the discussed subjects of intrinsic features and application of energy data be considered when deciding to harness the potency of energy data.

References

Aurangzeb, K., and Alhussein, M. (2020). Deep Learning Framework for Short Term Power Load Forecasting, a Case Study of Individual Household Energy Customers. 2019 International Conference on Advances in the Emerging Computing Technologies, AECT 2019.

Bansal, A., Rompikuntla, S. K., Gopinadhan, J., Kaur, A., and Kazi, Z. A. (2015). Energy Consumption Forecasting for Smart Meters. December 2015.

Beliaeva, N., Petrochenkov, A., and Bade, K. (2013). Data Set Analysis of Electric Power Consumption. European Researcher, 61(10–2), 2482–2487.

Bernard, T. (2018). Non-Intrusive Load Monitoring (NILM): combining Multiple Distinct Electrical Features and Unsupervised Machine Learning Techniques. https://duepublico2.uni-due.de/rsc/thumbnail/duepublico_mods_00046575.png

BP. (2021). Statistical Review of World Energy globally Consistent Data On World Energy Markets. And Authoritative Publications in The Field Of Energy. BP Energy Outlook, 70, 8–20.

Buzau, M. M., Tejedor-Aguilera, J., Cruz-Romero, P., and Gomez-Exposito, A. (2019). Detection of Non-Technical Losses Using Smart Meter Data and Supervised Learning. IEEE Transactions on Smart Grid, 10(3), 2661–2670.

Carrie Armel, K., Gupta, A., Shrimali, G., and Albert, A. (2013). Is Disaggregation The Holy Grail of Energy Efficiency? The Case Of Electricity. Energy Policy, 52, 213–234.

Chicco, G. (2012). Overview and Performance Assessment of The Clustering Methods For Electrical Load Pattern Grouping. Energy, 42(1), 68–80.

Dahunsi, F. M., Olawumi, A. E., Ale, D. T., and Sarumi, O. A. (2021). A Systematic Review Of Data Pre-Processing Methods and Unsupervised Mining Methods Used in Profiling Smart Meter Data. AIMS Electronics and Electrical Engineering, 5(4), 284–314.

Danilo, B. (2015). Intrusive and Non-Intrusive Load Monitoring (A Survey) Approach, Learning. Latin American Journal of Computing Lajc, 2(1), 45–53.

Efthymiou, C., and Kalogridis, G. (2010). Smart Grid Privacy via Anonymization of Smart Metering Data. 238–243.

Eibl, G., and Engel, D. (2015). Influence of Data Granularity on Smart Meter Privacy. IEEE Transactions on Smart Grid, 6(2), 930–939.

Govindarajan, R., Meikandasivam, S., and Vijayakumar, D. (2019). Cloud Computing Based

Smart Energy Monitoring System. International Journal of Scientific and Technology Research, 8(10), 886–890.

Granell, R., Axon, C. J., and Wallom, D. C. H. (2014). Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles. 1–8.

Grolinger, K., Capretz, M. A. M., and Seewald, L. (2016). Energy consumption Prediction With Big Data: Balancing Prediction Accuracy and Computational Resources. Proceedings - 2016 IEEE International Congress on Big Data, BigData Congress 2016, 90, 157–164.

Guo, Y. C., Niu, D. X., and Chen, Y. X. (2006). Support Vector Machine Model in Electricity Load Forecasting. Proceedings of the 2006 International Conference on Machine Learning and Cybernetics, 2006(August), 2892–2896.

Gupta, V. (2017). An Overview of Different Types of Load Forecasting Methods and the Factors Affecting the Load Forecasting. International Journal for Research in Applied Science and Engineering Technology, V(IV), 729–733.

Haq, A. U., and Jacobsen, H.-A. (2016). A Step Towards Advanced Metering for the Smart Grid: A Survey of Energy Monitors. 1–8.

Javeri, I. Y., Toutiaee, M., Arpinar, I. B., Miller, T. W., and Miller, J. A. (2021). Improving Neural Networks for Time Series Forecasting using Data Augmentation and AutoML. 1–8. http://arxiv.org/abs/2103.01992

Javier Campillo, Fredrik Wallin, Daniel Torstensson, I. V. (2012). Energy Demand Model Design for Forecasting Electricity Consumption and Simulating Demand Response Scenarios in Sweden. International Conference on Applied Energy, Suzhou, China.

Jiang, R., Lu, R., Wang, Y., Luo, J., Shen, C., and Shen, X. (2014). Energy-theft Detection Issues For Advanced Metering Infrastructure in Smart Grid. Tsinghua Science and Technology, 19(2), 105–120.

Jiang, Z., Lin, R., Yang, F., Liu, Z., and Zhang, Q. (2017). Comparing Electricity Consumer Categories Based on Load Pattern Clustering With Their Natural Types. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10393 LNCS(August), 658–667.

Kahl, M., Haq, A. U., Kriechbaumer, T., and Jacobsen, H. (2016). WHITED - A Worldwide Household and Industry Transient Energy Data Set. 3rd International Workshop on Non-Intrusive Load Monitoring (NILM2016), April, 1–4.

Kelly, J., and Knottenbelt, W. (2015). The UK-DALE Dataset, Domestic Appliance-Level Electricity Demand and Whole-House Demand From Five UK Homes. Nature, 2, 1–14.

Kim, J., Le, T. T. H., and Kim, H. (2017). Non-intrusive Load Monitoring Based on Advanced Deep Learning and Novel Signature. Computational Intelligence and Neuroscience, 2017, 1-22.

Klemenjak, C., Reinhardt, A., Pereira, L., Makonin, S., Bergés, M., and Elmenreich, W. (2019). Electricity Consumption Data Sets: Pitfalls and Opportunities. BuildSys 2019 - Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, November, 159–162.

Koivisto, M., Heine, P., Mellin, I., and Lehtonen, M. (2013). Clustering of Connection Points and Load Modeling in Distribution Systems. IEEE Transactions on Power Systems, 28(2), 1255–1265.

Kolter, J. Z., and Johnson, M. J. (2011). REDD : A Public Data Set for Energy Disaggregation Research. SustKDD Workshop, xxxxx(1), 1–6. http://users.cis.fiu.edu/~lzhen001/activities/KDD 2011 Program/workshops/WKS10/doc/SustKDD3.pdf

Lu, N., Du, P., Guo, X., and Greitzer, F. L. (2012). Smart Meter Data Analysis. Proceedings of the IEEE Power Engineering Society Transmission and Distribution Conference, 1–6.

Lua, S. W., Teng, J. H., Chan, S. Y., and Hwang, L. C. (2009). Development of a Smart Power Meter for AMI Based on ZigBee Communication. Proceedings of the International Conference on Power Electronics and Drive Systems, 661–665.

Makonin, S., Popowich, F., Bartram, L., Gill, B., and Bajić, I. V. (2013). AMPds: A public Dataset for Load Disaggregation and Eco-Feedback Research. 2013 IEEE Electrical Power and Energy Conference, EPEC 2013, Section III.

Makonin, S., Wang, Z. J., and Tumpach, C. (2018). RAE: The Rainforest Automation Energy Dataset for Smart Grid Meter Data Analysis. Data, 3(1), 1–9.

Marinakis, V. (2020). Big data for energy management and energy-efficient buildings. Energies, 13*(7),* 1555-1573

Meshram, R., Deorankar, A. V, and Chatur, D. P. N. (2012). Load Pattern Analysis of Electricity Customers based on Clustering Algorithm. International Journal Of Computer Science And Technology, 3, 702–705.

Monacchi, A., Egarter, D., Elmenreich, W., D'Alessandro, S., and Tonello, A. M. (2015). GREEND: An energy consumption dataset of households in Italy and Austria. 2014 IEEE International Conference on Smart Grid

Communications, SmartGridComm 2014, 1, 511–516.

Palacios-Garcia, E. J., Rodriguez-Diaz, E., Anvari-Moghaddam, A., Savaghebi, M., Vasquez, J. C., Guerrero, J. M., and Moreno-Munoz, A. (2017). Using Smart Meters Data For Energy Management Operations and Power Quality Monitoring in a Microgrid. IEEE International Symposium on Industrial Electronics, June, 1725–1731.

Pereira, L., Quintal, F., Gonçalves, R., and Nunes, N. J. (2014). SustData: A Public Dataset for ICT4S Electric Energy Research. ICT for Sustainability 2014, ICT4S 2014, July, 359–368.

Pereira, L., Velosa, N., and Pereira, M. (2022). A Data Model and File Format to Represent and Store High Frequency Energy Monitoring and Disaggregation Datasets. Scientific Reports, 12(1), 1–13.

Picon, T., Meziane, M. N., Ravier, P., Lamarque, G., Novello, C., Bunetel, J.-C. Le, and Raingeaud, Y. (2016). COOLL: Controlled On/Off Loads Library, a Public Dataset of High-Sampled Electrical Signals for Appliance Identification. 3, 1–5.

Pitì, A., Verticale, G., Rottondi, C., Capone, A., and Lo Schiavo, L. (2017). The Role of Smart Meters in Enabling Real-Time Energy Services For Households: The Italian case. Energies, 10(2), 199-224

Quilumba, F. L., Lee, W. J., Huang, H., Wang, D. Y., and Szabados, R. (2014). An Overview of AMI Data Pre-Processing to Enhance the Performance of Load Forecasting. 2014 IEEE Industry Application Society Annual Meeting, IAS 2014, 1–7.

Ramos, S., Soares, J., Vale, Z., and Ramos, S. (2013). Short-term Load Forecasting Based on Load Profiling. 2013 IEEE Power and Energy Society General Meeting held on 21-25 July 2013, Vancouver, British Columbia, Canada.

Rodrigues, E. M. G., Godina, R., Shafie-Khah, M., and Catalão, J. P. S. (2017). Experimental Results on a Wireless Wattmeter Device for the Integration in Home Energy Management Systems. Energies, 10(3), 398 - 416

Sahoo, S., Nikovski, D., Muso, T., and Tsuru, K. (2015). Electricity Theft Detection Using Smart Meter Data. 2015 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference, ISGT 2015, Columbia, USA.

Salehkalaibar, S., Aminifar, F., and Shahidehpour, M. (2019). Hypothesis Testing for Privacy of Smart Meters With Side Information. IEEE Transactions on Smart Grid, 10(2), 2059–2067.

Sankar, L., Raj Rajagopalan, S., Mohajer, S., and Vincent Poor, H. (2013). Smart Meter Privacy: a Theoretical Framework. IEEE Transactions on Smart Grid, 4(2), 837–846.

Sayed, S., Hussain, T., Gastli, A., and Benammar, M. (2019). Design and Realization of an Open-Source And Modular Smart Meter. Energy Science and Engineering, 7(4), 1405–1422.

Shin, C., Lee, E., Han, J., Yim, J., Rhee, W., and Lee, H. (2019). The Enertalk dataset, 15 Hz Electricity Consumption Data From 22 Houses in Korea. Scientific Data, 6(1 1), 1–13.

Shin, C., Rho, S., Lee, H., and Rhee, W. (2019). Data requirements for applying machine learning to energy disaggregation. Energies, 12(9), 1696 - 1715.

Singh, A. K., Ibraheem, Khatoon, S., Muazzam, M., and Chaturvedi, D. K. (2012). Load Forecasting Techniques and Methodologies: A review. ICPCES 2012 - 2012 2nd International Conference on Power, Control and Embedded Systems, Uttar Pradesh, India.

Solangi, K. H., Islam, M. R., Saidur, R., Rahim, N. A., and Fayaz, H. (2011). A Review on Global Solar Energy Policy. Renewable and Sustainable Energy Reviews, 15(4), 2149–2163.

Tan, O., Gunduz, D., and Poor, H. V. (2013). Increasing Smart Meter Privacy Through Energy Harvesting and Storage Devices. IEEE Journal on Selected Areas in Communications, 31(7), 1331–1341.

Tjaden, T., Bergner, J., Weniger, J., and Quaschning, V. (2015). Representative Electrical Load Profiles Of Residential Buildings In Germany With a Temporal Resolution of One Second. Dataset, HTW Berlin - University of Applied Sciences Research, November, 1–7.

Wang, M. C., Tsai, C. F., and Lin, W. C. (2021). Towards Missing Electric Power Data Imputation for Energy Management Systems. Expert Systems with Applications, Expert Systems with Applications 174(1): 114743, 1-20

Wang, Yi, Chen, Q., Hong, T., and Kang, C. (2019). Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. IEEE Transactions on Smart Grid, 10(3), 3125–3148.

Wang, Yi, Chen, Q., Kang, C., Zhang, M., Wang, K., and Zhao, Y. (2015). Load Profiling and Its Application to Demand Response: A Review. 20(2), 117–129.

Wang, Yikuai, Qiu, H., Tu, Y., Liu, Q., Ding, Y., and Wang, W. (2018). A Review of Smart Metering for Future Chinese Grids. Energy Procedia, 152, 1194–1199.

Yip, S. C., Wong, K. S., Hew, W. P., Gan, M. T., Phan, R. C. W., and Tan, S. W. (2017). Detection of Energy Theft and Defective Smart Meters in Smart Grids Using Linear Regression. International Journal of Electrical Power and Energy Systems, 91, 230–240.

Zanella, A., Bui, N., Castellani, A., Vangelista, L., and

Zorzi, M. (2014). Internet of Things for Smart Cities. IEEE Internet of Things Journal, 1(1), 22–32.

Zeifman, M., and Roth, K. (2011). Non-intrusive Appliance Load Monitoring: Review and outlook. IEEE Transactions on Consumer Electronics, 57(1), 76–84.

Zheng, S., Zhong, Q., Peng, L., and Chai, X. (2018). A Simple Method of Residential Electricity Load Forecasting by Improved Bayesian Neural Networks. Mathematical Problems in Engineering, 2018, 1-16

Zhou, K. Le, Yang, S. L., and Shen, C. (2013). A Review of Electric Load Classification in Smart Grid Environment. Renewable and Sustainable Energy Reviews, 24, 103–110.

Zoha, A., Gluhak, A., Imran, M. A., and Rajasegarar, S. (2012). Non-intrusive Load Monitoring Approaches for Disaggregated Energy Sensing: A Survey. Sensors (Switzerland), 12(12), 16838–16866.