Compression Techniques of Electrical Energy Data for Load Monitoring: A Review

F. M. Dahunsi[1*], O. A. Somefun[1], A. A. Ponnle[2], K. B. Adedeji[2]

[1]Department of Computer Engineering, Federal University of Technology, Akure, Nigeria.
[2]Department of Electrical and Electronics Engineering, Federal University of Technology, Akure, Nigeria.

ABSTRACT: In recent years, the electric grid has experienced increasing deployment, use, and integration of smart meters and energy monitors. These devices transmit big time-series load data representing consumed electrical energy for load monitoring. However, load monitoring presents reactive issues concerning efficient processing, transmission, and storage. To promote improved efficiency and sustainability of the smart grid, one approach to manage this challenge is applying data-compression techniques. The subject of compressing electrical energy data (EED) has received quite an active interest in the past decade to date. However, a quick grasp of the range of appropriate compression techniques remains somewhat a bottleneck to researchers and developers starting in this domain. In this context, this paper reviews the compression techniques and methods (lossy and lossless) adopted for load monitoring. Selected top-performing compression techniques metrics were discussed, such as compression efficiency, low reconstruction error, and encoding-decoding speed. Additionally reviewed is the relation between electrical energy, data, and sound compression. This review will motivate further interest in developing standard codecs for the compression of electrical energy data that matches that of other domains.

KEYWORDS: Data compression, load monitoring, time series analysis, load forecasting, smart-meter, energy-monitor.

[Received Jan. 16, 2021; Revised June 4, 2021; Accepted Aug. 4, 2021] Print ISSN: 0189-9546 | Online ISSN: 2437-2110

I. INTRODUCTION

In the past few years, there has been a surge in the deployment of smart meters, a central component in the advanced metering infrastructure (AMI) worldwide. As at 2016, about 169 million smart meters were installed in the UK, US, and China (Wang, *et al*., 2019). These devices allow a considerable amount of load-consumed electrical energy data (EED) to be collected at the medium-low voltage levels of the electric grid, mainly for load forecasting, especially in residential buildings (Wee and Nayak, 2019). Logged massive data from these devices create a demand on the limits of computing resources needed for their processing and storage. Therefore, as this demand continues to grow, likewise its cost, the need for efficient and guaranteed real-time data compression systems for electrical energy data becomes more evident (Nithiyananthan and Ramachandran, 2014). Computing resources in terms of memory storage hardware and sizes, data transmission hardware, and bandwidths are limited and constrained. It is no wonder then that the most important motivation behind data compression in this domain is reduced congestion of communication channels used for data transmission, reduced storage overhead, and improved data mining efficiency (Lendák, 2019).

Compression minimizes storage space and the effect of transmission bandwidth, thereby minimizing the impact of memory, processor, network, and time constraints. As this need for compression in the electrical energy load monitoring domain remains, the science and art of compression technology will continue to be an important and challenging problem.

Load Monitoring (LM) analysis is of fundamental importance to effective energy management in the smart grid. LM is categorized into two major parts: intrusive and non-intrusive categories (Haq, 2018). The latter is more attractive since it involves lower cost, easier installation, and promising potentials for scalability (Batra, Dutta, and Singh, 2013). Notwithstanding, a relatively recent survey in Zhuang *et al*. (2018) reported that non-intrusive LM for energy detection in buildings is still challenging because of its high complexity. One of these complexities is the big dataset involved. Therefore, in carrying out LM analysis, the concepts of data acquisition, feature extraction, and load identification all reduce to the end goal of reducing the size of the electrical energy dataset without affecting its usefulness for inference and prediction.

In retrospect, the advantages of compression cannot be overemphasized. For instance, in the digital audio domain, the compression of music and speech data has helped to fit more songs into smartphones and user-centric computing devices and download these files faster. Consequently, for this to manifest, there needs to be a new wave of in-depth research into the design of novel compression techniques and easily accessible codec libraries targeted at electrical energy data. In

*Corresponding author: fmdahunsi@futa.edu.ng
doi: http://dx.doi.org/10.4314/njtd.v18i3.4

the past decade (Figure 1), considerable research has been conducted on the compression of electrical energy data (electric signal waveforms).

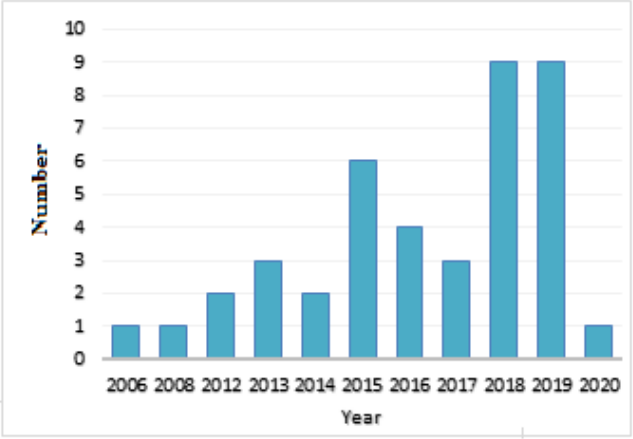

Figure 1: List of relevant publications on the compression of electrical energy data by year.

Only a few reviews were found on the compression of electrical energy data. To the best of current knowledge, Tcheou *et al.* (2014) is the first work to present this as a research challenge for the smart grid. In Ringwelski *et al.* (2012). Tiny and mixed forms representative of lossless compression methods were compared for resource-constrained smart-meter hardware. More recently, Wen *et al.* (2018) presented a general big-data survey with far-fetched categorizations of compression methods for smart-meter data. Also, Jumar *et al.* (2018) compared selected off-the-shelf lossless compression codecs. The recent book by Wang *et al.* (2020) also contains a chapter dedicated to compressing aggregated smart-meter data for modelling and forecasting analysis.

Therefore, this paper conducts a review from the year 2006 -the year 2020 of relevant literature on the compression of electric energy data (EED) for LM analysis. The contributions of this paper are focused on the current state of compressing electrical energy data about LM analysis. Furthermore, this review aims to provide a concise introductory reference for early-researchers in the design, development, and application of compression techniques for electric-energy data aggregated from smart meters or energy monitors. In contrast to the other papers, here, a general review of the literature on lossless and lossy compression methods and codec libraries for electrical energy data aggregated from either smart-meters or energy monitor devices is provided, with the added Objective of increasing renewed interest in researchers in this domain into the need for standardized and readily available codec libraries targeted at compressing and decompressing electrical energy data.

The structure of the remaining sections of this paper is as follows: Section II focuses on a general overview of state of the art regarding compression techniques adopted in literature for EED. Section III focuses on EED lossy compression methods, and Section IV focuses on EED lossless compression methods. Section V discusses the challenges and further issues concerning EED compression. The conclusion is discussed in Section VI.

II. STATE OF THE ART

First presented is a survey of current and relevant literature for compression of electrical energy data. EED has many application-specific use-cases in the smart grid, such as energy feedback, grid monitoring, and load forecasting (Wang, *et al.*, 2020). The most important arguments for compressing this dataset type are motivated by the data volume, communication bandwidth, and energy efficiency (Gerek and Ece, 2008; Unterweger and Engel, 2015). A taxonomy table for the major compression methods in the literature is presented in Table 1. Discussions on the characteristic components of electrical energy data for compression are presented in Tcheou *et al.* (2014) and Haq (2018).

A. Electrical Energy Data Parameters

The first interface for obtaining electrical energy data for measurement is an instrumentation and data acquisition device (DAQ) (Haq, 2018). To compute its main parameters (RMS voltage, current, real power, power factor), collected energy signals need to be converted to discrete-time signals. This is because raw electrical signals are continuous (analog). The uncertainty in the transformed signal is determined by the accuracy of the ADC (Analog-to-Digital Converter) utilized and it determines the resolution of the signal.

Fundamentally, electric signals are quantified in the form of voltage and current waveforms. The absence of a frequency component determines their classification as alternating (AC) or direct (DC) waveforms. Generally, for a three-phase AC power-point, the sinusoidal 120-degree phase-shifted raw data stream from the sensing device comprises of three voltage measurement channels and four current measurement channels (including one for the neutral conductor). Such fundamental parameters of interest in the load monitoring of electrical energy are outlined and discussed.

1) *Voltage*

Electrical voltage waveform measurement is carried out using sensors known as voltage transformers (for example, AC–AC adaptors). They are used to measure the peak and root-mean-square (RMS) voltages from a terminating point. Voltage waveforms have a more stable form and are therefore easier to reconstruct digitally, unlike current waveforms.

2) *Current*

More practically, due to varying individual phase-loads that cause phase difference and distortions, electrical current waveforms are especially individually distorted. A core reason is that current waveforms change when power demand changes (Gerek and Ece, 2008). When few loads are connected in a monitored building, variations in the waveform rarely occur. Consequently, current waveforms are sine waves with unstable forms, varying considerably depending on the

Table 1. Taxonomy table for major compression methods applied to electrical energy data.

Class	Category	Representative Techniques	General CR	Selected References
Lossy	Transform and Parametric coding	DWT, SAX, PAX,SVD, PCA, FLDC	Relatively high	(Wang, *et al*., 2020; Wang, *et al*., 2017; Wee and Nayak, 2019; Tariq, *et al*., 2015; Sayood, 2018)
Lossless	Statistical-based Coding (Entropy methods)	RLDC, Audio-based codecs,	Relatively low	(Abuadbba, *et al*., 2018; Haq, 2018; Sayood, 2018; Firmansah and Setiawan, 2016; Unterweger and Engel, 2016; Kelly and Knottenbelt, 2015)
	Dictionary-based Coding	LZMA codecs and variants	Relatively low	

- CR – Compression Ratio

operating load (resistive, inductive, or capacitive load) appliances and require increased sampling to reconstruct accurately (Haq, 2018).

3) *Power and Power Factor*

A central metric feature in almost all energy-metering devices is the real power (Zhuang, et al., 2018; Lendák, 2019). It is the measure of the actual rate at which electrical energy is consumed. Mathematically, it is calculated through the electric voltage and current measurements. Also, the power factor is a valuable metric for differentiating between different operating load appliances. It is a measure of the phase difference caused by both the inductive and capacitive components. A positive phase difference indicates a net inductive reactance, that is, electric current lag voltage. On the contrary, a negative phase difference indicates a net capacitive reactance, with electric current leading to voltage.

4) *Sampling Frequency*

One essential requirement in many LM analysis setups is to detect connected electrical load appliances accurately from their aggregated load data (Wang, *et al.*, 2012). For energy monitoring through DAQs, a specific sampling rate or frequency depends on the desired amount of information to be obtained from the load data source. The sampling rate or frequency accounts for the resolution along the time-axis during analog to digital conversion. In general, since electrical appliances draw current, for finer observation of the harmonics and transient switching response of electrical appliances, it is a sensible decision to utilize a higher sampling frequency. It was noted that typical energy monitors used a one-second to one-minute sampling rate on 16-bit ADCs (Haq, 2018).

5) *Resolution*

Resolution of the digital electrical signal measured by any DAQ system is defined by the number of bits of the ADC in the DAQ system. This number of ADC bits represents the resolution and the measured data axis during analog to digital conversion (Haq, 2018). It defines the number of digital codes (symbols of 1s and 0s) that can be formed using these bits. Therefore, resolution influences the accuracy of load event detection from an aggregated data source amidst simultaneous load events.

B. Electrical energy Data Acquisition Systems

Computationally, data is simply a digital representation of information that can be organized, stored, and processed as a file (Pu, 2006). Although not exhaustive, data can be classified into text, binary, audio, image, and video formats. Text data are usually in American Standard Code for Information Interchange (ASCII) format, in files ending with specific extensions, for example: *'.txt, .tex, .doc'* extensions, or they are typically any programs in a high-level language file format. Examples of binary data are spreadsheet data, executable files, and so on. Examples of image data are represented by two-dimensional arrays of pixels, vectors, or math equations.

In contrast, audio (sound) data are wave (periodic) functions, and such as the *'.wav'* file format. Electrical signals (current and voltage) also wave functions. Depending on the data acquisition method, they have been obtained as audio formats or logged in text files. The size of the logged data is almost linearly proportional to the sampling frequency (Kelly and Knottenbelt, 2015).

Consider a DAQ acquiring one-hour energy data from two sensor channels (electric voltage and current) 16-bit per value (2-bytes (B)) with: 1-Hz, 1-kHz, and 1-MHz sampling frequency. The estimated corresponding file sizes would be as follows:

$1Hz$: $(1Hz * 3600s * 2B) * 2 \approx$ 14.4 kB;

1kHz: $(1\text{kHz} * 3600\text{s} * 2\text{B}) * 2 \approx$ 14.4 MB; and

1MHz: $(1\text{MHz} * 3600\text{s} * 2\text{B}) * 2 \approx$ 14.4 GB

Therefore, as the sampling rate increases, the transmission of such EED measurements becomes more demanding, especially at the receiving end of the communication link (Unterweger and Engel, 2016). Higher frequencies lead to larger file sizes.

In contrast, load-appliance detection algorithms in literature (Unterweger and Engel, 2016) require sampling frequencies in the kHz to MHz or greater range to accurately perform load disaggregation in near-real-time (Haq, 2018). Higher frequencies enable more precise detection of transient appliance switching events and prediction of the power consumption. This can often be achieved by utilizing a single power meter per household; this ability is called scalability. Scalability can be considered as the ability of a DAQ system to detect newly added load appliances easily.

Recently, there has been a growing interest in appliance-level e-monitoring to help consumers view fine-detailed energy consumption information or high-resolution data for

increased overall appliance detection accuracy (Wang, *et al*., 2012; Kelly and Knottenbelt, 2015; Yan, *et al.*, 2019*)*. Therefore, the eventual choice of the sampling frequency is an essential consideration for accurate load disaggregation analysis.

On the whole, the first and crucial stage for any LM system is data acquisition. Such DAQ devices are called smart meters or energy (e)-monitors (Haq, 2018). In Wang *et al*. (2020), a smart meter is described as a two-way data communication link between a building and a utility company. The purpose is to enable the feedback of EED to the utility company for load forecasting while also enabling remote billing. The utility often owns the smart meter and it comes with integral disadvantages related to data confidentiality and privacy (Wang, et al., 2015).

On the other hand, an energy monitor is not utility controlled as it works independently with existing energy meters, without any direct effect on the billing. E-monitors also assist in observing energy consumption patterns in real-time and, such devices can help users make informed decisions for conserving energy. E-monitors are preferred because they can be easily installed and minimize privacy concerns (Lendák and Horvath, 2019; Kelly and Knottenbelt, 2015; Haq and Jacobsen, 2018). Sampling frequency divides DAQ e-monitors and smart meters into low frequency (typically less than 1Hz) and high frequency (kHz and above) (Basu, 2015).

The resultant load or energy consumption profiles from these two electrical energy monitoring devices help determine the energy usage pattern concerning time (Basu, 2015; Le, 2017). For consumers, these patterns or trends help find energy leaks or gaps, while for utility companies, these patterns are statistical tools for load forecasting (Zhuang, et al., 2018). Data from these smart devices are primarily used for monitoring and planning purposes that require coarse-grained information. Further, Haq and Jacobsen (2018) provided a sound analysis of non-intrusive LM and highlighted critical requirements for such DAQs.

The trend in the current literature is a preference for high-frequency data to study load signatures such as electric current and voltage waveforms to identify appliances more correctly. However, this comes with increased cost and complexity. From the lower cost and complexity perspective, there is a higher deployment of low-frequency meters but with limited functionality. In the study by Zhuang *et al*. (2018), they concluded that a low-cost but high-frequency sampling LM framework is necessary to facilitate the scalability of LM. This reduced cost and improved scalability can be achieved and made sustainable through efficient compression of monitored (logged) electrical energy data quantifiable in its parameters. However, the availability of such standard compression codecs for general load monitoring analysis of electrical energy data is still far from mature and open to more research and development time.

B. EED Compression

Smart grid meters and energy monitoring devices continuously churn out a huge amount of data at a constant rate, bringing up specific challenges in their transmission and storage. In the context of computing, the compression of data implies representing information in a compact form by removing redundant information in the data (Sayood, 2018). The performance of applied compression techniques has also been interpreted using different indices in literature (Maher, 2003). The main idea of compression is that if redundancy can be identified and removed in a given dataset or stream, it can reduce its effective size (Nithiyananthan and Ramachandran, 2014). The essential criterion for compression is that the storage size of the source data is reduced. The apparent quality of the source data is not adversely affected by the compression method. This measure or index defining the degree of file size reduction is usually expressed in a compression ratio (CR). The compression ratio is a measure of compression efficiency. Since standard compression algorithms aim to optimise savings percentage, an energy-savings metric was used as the primary metric comparison of data compression algorithms (Sadler and Martonosi, 2006). The compression ratio can be understood as the size of the original data file (input) divided by the compressed data file (output). Compression ratio (CR) in percentage can also be expressed as the savings percentage (SP), where N_o is the size of the uncompressed data, N_c is the size of the compressed data.

$$\mathrm{CR}(\%) = 100 \times \frac{N_c}{N_o} \tag{1}$$

$$SP(\%) = 100 \times \frac{(N_o - N_c)}{N_o} \tag{2}$$

In terms of real-time compression performance for transmission through a communications channel, another measure is the bandwidth of the transmission channel given in the bit rate (BR, bits per second). This is the number of bits required to represent the data, divided by the total playing (recording) transmission time. Other measures of a compression algorithm more relevant to engineers who develop compression algorithms are computational complexity, compression time, entropy, overhead, etc. In terms of reconstruction of the original load-profile data from smart-meters, a measure known as the mean-peak percent error (MPPE) was introduced in Wang *et al*. (2020), where $|e_t|$ is the absolute difference between the reconstructed and original data at time-instance t; the number of time-intervals is an integer T, and L_{max} is the daily peak load.

$$\mathrm{MPPE}(\%) = 100 \times \frac{1}{T}\sum_{1}^{T}\frac{|e_t|}{L_{max}} \tag{3}$$

The efficiency of a compression algorithm is more critical when data is recorded in real-time but limited by memory storage or transmission channel constraints. The efficiency of a decompression algorithm is of more importance at the receiving end, where the data quality is of concern. It is, however, important to note that there is no *'one-size-fits-all solution'* for data compression (Pu, 2006), as it is subject to a space-time complexity trade-off both at the compression and decompression ends (Sari, *et al.*, 2018).

The method of data compression can either be lossy or lossless. The motivation for lossless data compression lies in the ability to compress while still maintaining the quality of the original data. However, because of much higher compression ratios, lossy compression methods have found

high application in situations where some loss in original data quality is not of concern. Lossy compression typically reduces bits in the data by identifying and removing all possible redundant (unnecessary) information. It can be mainly applied to accelerate similarity search, upon which many critical data-mining applications like load profiling and customer segmentation are based (Wang, et al., 2020). Whereas lossless compression usually reduces bits by removing only statistical redundancy.

In the smart grid literature, most compression works focus on lossy methods suited to smart meters since they are designed to use sampling rates that significantly relax the requirements for communication channels and storage space to measure aggregated electrical data. Wang *et al*. (2020) presented a comprehensive study on smart meter big data compression solutions. The smart power grid is projected to utilize waveform level monitoring with sampling rates in the kilohertz range for detailed grid status assessment (Jumar, *et al.*, 2018). In some LM applications, it is vital to reconstructing data precisely similar to the original without losing information. Atif *et al*. (2019) suggested using a mix of Singular Value Decomposition (SVD), normalization, and value-index sparse matrix representation. More detailed information is required for better event detection for power quality (PQ), energy monitoring, and load disaggregation (Wang, *et al.*, 2020). Tariq *et al*. (2015) observed that more than 70% of smart grid data consists of a repetitive timestamp pattern and current reading. They proposed a solution for saving the timestamp, such that the time interval is used to retrieve the timestamps back when the file is decompressed. This technique was then augmented with dictionary-based methods codecs for improved compression and decompression times.

Fagiani *et al*. (2019) concluded that electrical data loses its quality at low sampling rates and thereby, affect the quality of the load monitoring. Wang *et al*. (2012) proposed a compressive sampling approach to measure steady-state current signatures using a random filter and analog to information converter to sub-sample the original current signature. More recently, Rodriguez-Silva and Makonin (2019) proposed another approach, termed "universal", using a complex filter, probabilistic, and partition pipelines. In Yan *et al*. (2019), a non-audio compression method, lossless coding precision (LCP) codec was used on the LIFTED dataset and compared with dictionary-based data compression methods. Basu (2015) highlighted the complex problem of detecting low energy consuming devices at low sampling rates. A System-on-Chip (SoC) compression encoder was developed in Bellasi *et al*. (2019) using a combined non-uniform sampling (NUS) and random modulation techniques for automatic compression of electric current data.

The same concept as compressing electrical data was investigated in Clark and Lampe (2015) on the BLUED dataset. Fagiani *et al*. (2019), working with the UK-DALE and REDD datasets, used NUS and uniform sampling (US) as a data-reduction policy to reduce the acquired electrical data size and ensure compliance with network bandwidth limits. In Kelly and Knottenbelt (2015) a popular lossless audio-based compression encoder, FLAC, was used to generate the UK-DALE dataset. Following this approach, in Haq (2018), a comparison was made between popular audio-based and dictionary-based compression techniques or algorithms. It concluded that audio-based compression shows better performance concerning compression ratio and processing time for electrical energy data. Many general-purpose dictionary-based algorithms have also been widely and effectively used to compress both text and program files for storage and transmission over the communication network. However, they are inefficient, having a very low compression ratio when used on audio data (waveform data) with statistical properties. Conversely, it has been noted that many lossless audio compression algorithms are no longer actively maintained (Hans and Schafer, 2001).

III. EED LOSSY COMPRESSION

Currently, lossy compression methods dominate the data modelling (mining) and analysis of electric power big data. Wen *et al*. (2018) presented a survey on the characteristics of smart meters and the challenges with compressing smart meter data as big data. The authors highlighted some research issues in smart meter data compression methods. They noted that lossless methods are often less efficient than lossy compression methods that achieve far lower compression ratios. They also noted that, although there are many studies on the compression of smart meter data, there is no perfect fool-proof system available to evaluate the ideal compression effect of the algorithm(s) used in processing the big data churned out by smart meters.

The roll-out trend of smart meters for capturing domestic loads in residential buildings cannot be understated as it keeps increasing. It can be argued that one reason for this is the ease of aggregated usage profile insight on appliance loads (Wang, *et al*., 2020). However, it has been noted that, even with this advantage, 'big data' problem in terms of data storage, transmission, and processing (due to limited memory storage size, bandwidth, and processing-time of the hardware), always arises (Chandak, et al., 2020).

Therefore, compressing the load profile of smart meter data allows for a more efficient approach to this 'big data' problem. In this context, lossy compression is preferred in most smart-grid applications to lossless compression. The goal is to retain only essential information in the smart meter logged time-series data (Wang, et al., 2020; Wen, et al., 2018).

Most lossy methods have a linear relationship with CR; that is, their information loss grows rapidly with the increasing CR (Wang, et al., 2016). A compact list of most commonly used lossy compression methods applied in this domain include Discrete Wavelet Transform (DWT), Discrete Fourier Transform (DFT), Singular Value Decomposition (SVD), Symbolic Aggregate Approximation (SAX), Principal Component Analysis (PCA), Wavelet Transform, Mixed Parametric and Transform Coding (Huang, et al., 2019; Wang,

et al., 2020). Another solution presented in (Huang, *et al.*, 2019) utilizes deep-stacked auto-encoders for electrical energy data (EED) load compression and classification. Lossy compression methods have also been proposed for feature identification. For instance, (Wang, *et al.*, 2016) and (Wang, et al., 2017) used the K-SVD sparse representation technique. The authors showed that their solution outperforms discrete wavelet transform (DWT), principal component analysis (PCA), as well as a Piecewise Aggregate Approximation (PAA).

Notably, the state-of-the-art feature-based load data compression method (FLDC) is suited to low granularity (or low frequency) EED in the lossy compression group. FLDC will be reviewed and compared to top-performing compression methods used on EED, from smart meters to smart-grid applications applied to load forecasting and analysis.

A. Feature-based Load Data Compression (FLDC)

Tong, et al. (2016) stated that FLDC is an EED-specific lossy compression technique. It combines the desirable properties of high compression efficiency, low reconstruction error, and a simple data compression format. Some fundamental terminologies involved with this method are presented next (Wang, *et al.*, 2016; Wang *et al*. 2020 and Tong, et al. 2016).

1) Residential load profile characteristics

The time interval of 30 minutes is a typical standard time instance for recording electric power consumption (kWh) as load profile data (Wang, *et al.*, 2016). The other two main characteristics as listed in Wang *et al*. (2020) that allow for compression of this type of time-series data are Consecutive Value Difference (CVD) and Generalized Extreme Value (GEV).

2) Consecutive value difference (CVD)

For data compression, a small CVD or SCVD helps to improve both compression efficiency and error. The CVD shows that at low levels (reduction in load), the load profile is more stable, and hence, the CVD is smaller. In contrast, when the load increases, primarily due to switching on high-power consuming appliances (such as ovens, irons, cookers, washers, etc.), the load profile becomes more unstable, and hence the CVD is bigger. Cumulative probability analysis of the CVD reveals what percentage of consecutive load values exhibit slight differences compared to the rest.

The CVD is described mathematically as:

$$r_{n,t} = \frac{P_{n,t} - P_{n,t-1}}{P_{max,n}} \tag{4}$$

where $r_{n,t}$ is the CVD rate at time instance t of day n; $P_{n,t}$ is the load at time-instance t of day n; $P_{n,t-1}$ is the load at time-instance $t-1$ of day n, and $P_{max,n}$ is the peak load on day n.

The load profile can be categorically divided into two states: a base-state and a stimulus-state. A base state in the load profile corresponds to a stable load, indicating smaller CVD. A stimulus-state corresponds to unstable load and higher CVD. The imaginary line separating both states with the base-state below and the stimulus-state above is called the state-boundary. At the state boundary, a load event is said to occur. Load events occur when the load profile deviates from base-states to several stimulus states before returning to the base state. This transition period at the imaginary state-boundary marks the occurrence of a load event that can be exploited for compression using a maximum likelihood estimation.

3) Generalized extreme value (GEV)

The GEV distribution function through the maximum likelihood estimation (MLE) is used to model the probability of extreme events common in residential electric power-consumption profile. It exploits the fact that the load-profile data for residential buildings are typically distributed denser at the base states and less dense at the stimulus states. The Frechet-type GEV function is the recommended best fit for residential or household smart meter data (Wang, et al., 2020). The GEV function is mathematically expressed as:

$$F(x) = \exp(-a^{-b}) \tag{5}$$

$$a = 1 + \frac{k(x-\mu)}{\sigma}, b = \frac{1}{k} \tag{6}$$

where x is the input data; k is the fitting parameter ($k > 0$). Also, μ, σ represent the centre and scale (variance) parameter, respectively, and then a, b are placeholders for simplifying the expression.

The FLDC is based on the general extreme value characteristic of residential load data. The authors state that the CR for the FLDC is always close to 1.8% of the original data volume. This was validated on the Irish and Chinese smart-meter data. The FLDC, as illustrated in Figure 1, is a framework of 6 operations at the high-level view.

4) Generalized extreme value (GEV) distribution fit

This is the first operation. Given a time-series data x, with each load-profile data-point x_t, possibly characterized by seasons in a year. A distribution fitting is obtained by maximum likelihood estimation, which gives $F(x)$, a cumulative probability density function (PDF). The load-state boundary $B = x$ is detected and computed at the point where $F(x) = \alpha$, where α is a confidence probability constant.

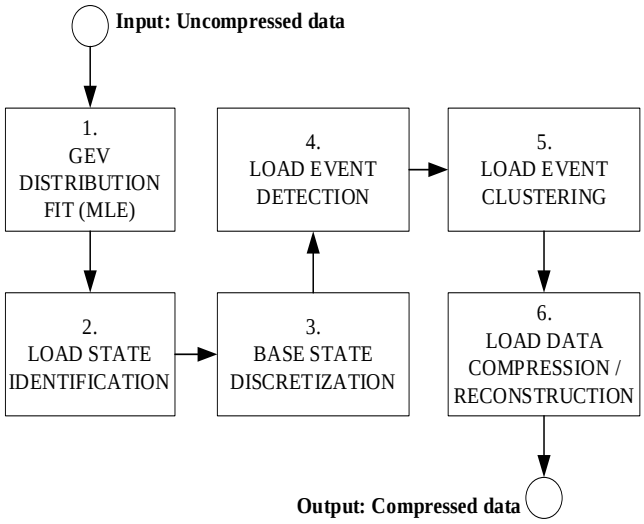

Figure 2. FLDC framework.

5) Load state identification

After B is calculated, a load-state matrix $S = [S_1,\ S_2,\ \dots,\ S_{\{n-1\}},\ S_n]$ is constructed. The element S_t are either 0 (base-state) or 1 (stimulus-state). They are generated by checking if each load value at each time-instance in the uncompressed load-profile data is below B.

$$S_t = \begin{cases} 0, & if\ x_t \le B \\ 1, & if\ x_t > B. \end{cases} \quad (7)$$

6) Base state discretization

The next operation is to discretize the base-states, that is, elements of the state-matrix 0. This is done by dividing the area under the GEV PDF fit by $\boldsymbol{c} = [c_0, c_1, \dots, c_d]$, where the discretization interval number $d = 8$ gives good resolution, $c_0 = (\mu\ \sigma)/k,\ \ c_d = B$. The area then becomes α/d. Therefore, any load series that falls in the base-state whose average-value is in the interval $\boldsymbol{c}$, the series is coded by what is called the sub-state ID, $ID(x) = i$, and the expected value $E(i) = \int_{c_{i-1}}^{c_i} x\ F(x)\ dx$, where $i = 1,2, \dots, d$.

7) Event-detection

Further, the following operation is to detect load-event transitions in the load-state matrix S through edge-detection. Then the number of stimulus states is coded by slicing the load-event data profile.

This is achieved by iterative one-step scanning of the elements of S. The start of the load-event $t_s = t + 1$, if $S_{t+1} - S_t = 1$, this implies a change from $0 \to 1$. The end of the load event is $t_e = t - 1$, if $S_t - S_{t-1} = -1$, this implies a change from $1 \to 0$. Then, the load-event data profile is constructed by slicing the load data, that is, $ELP = [x_{ts},\ x_{ts+1}, \dots, x_{te}]$. The length of the ELP, which is $t_e - t_s + 1$, represents the number of stimuli states.

8) Event-clustering

After all event states are detected, the sliced load-event data ELP profiles are used to construct a load-event segment pool for load-event clustering. The ELP length now represents the operation time interval of higher-power-consuming appliances. Thus, the ELPs are firstly classified according to their lengths, profile shapes, and load levels as metrics for the clustering. Wang *et al*. (2020) adopted the hierarchical clustering algorithm. The ELPs are divided into M groups with a group ID counted from 1 to M. ELPs with the same group ID are averaged to shape the representative profile data.

9) Load-data compression format

The last step in the coding stage is to store the representative data profile for one load event, and a data-structure format was proposed for this purpose. This data-coding format allows for effective data compression by reducing data storage and processing through a 16-bit (2 bytes) binary structure, as shown in Figure 2. The most significant bit on the left is called the 'next-day bit, which indicates whether the load event occurs on the same day (0) or the next day (1), concerning the start of the load event. The following six bits encode the time that the load event started. The maximum value for this time interval is $2^6 = 64$, and the minimum is 0. The following six bits represent the event group ID, which also supports a maximum of 64 event clusters before an overflow occurs. The last three bits, on the right, represent the sub-base state ID, which cannot be more than $2^3 = 8$ sub-base states.

10) Data reconstruction

The reconstruction of the original load profile can be carried out through a two-step process: The first is 'event reconstruction.' Here the representative load profile of the event group is used to reconstruct the original load-event profile using the start-time and the event group ID. The second and last step is 'base-state reconstruction.' Here, the base-load data before load events are generated from the expected values corresponding to the sub-base state IDs coded in the last three bits of the compressed data.

B. Performance Comparison

State-of-the-art methods such as Piecewise Aggregate Approximation (PAA), Symbolic Aggregate Approximation (SAX), Discrete Wavelet Transform (DWT), and Resumable Load Data Compression (RLDC) have been compared in Wang *et al*. (2020). The metrics used for evaluation of their compression performance are the compression ratio and mean-peak per cent error. Figure 3-4 shows that the FLDC has

Figure 3. FLDC data compression format.

the best balance in performance in terms of compression efficiency and reconstruction error. The recommended choice of compression format in order is shown in pyramid form in Figure 5-6. The conclusion is that when compression efficiency and low reconstruction error are vital objectives, the best EED lossy compression method is the FLDC. The RLDC becomes a practical choice when reconstruction error cannot be tolerated (as it is a hybrid lossless compression method specifically developed for the smart grid).

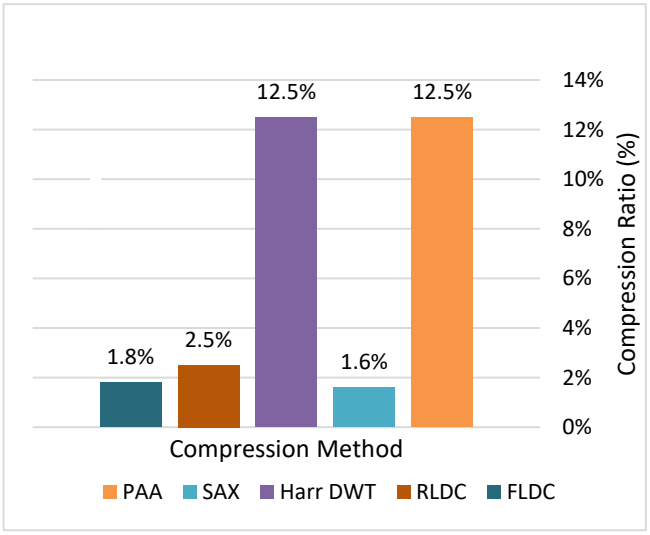

Figure 4. Coding performance comparison: compression ratio evaluation metric.

IV. EED LOSSLESS COMPRESSION

Data compression aims for the efficient removal of redundant information from a data stream. Lossless compression (coding) can be described as a compact term for compression methods that aim for perfect data reconstruction during decompression (decoding) (Maher, 2003). In other words, lossless compression methods can perfectly reconstruct the original data that was compressed. Notwithstanding, it is believed that lossless compression technology has reached its limit (Sayood, 2018). This compression class can be further divided into two sub-classes: statistical-based and dictionary-based methods.

Statistical-based methods are also known as entropy-based methods. Examples are arithmetic algorithms, such as the Huffman algorithms, which are based on a statistical model and the probability distribution of the source data. The Huffman coding algorithm is known as one of the best-known variable-length coding algorithms for statistical coding methods. Others are the Golomb–Rice coding and the Tunstall coding (Pu, 2006; Sayood, 2018).

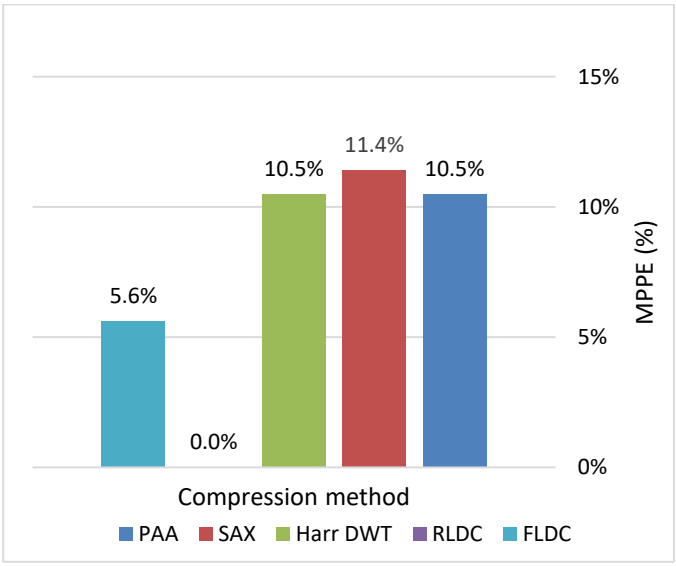

Figure 5. Decoding performance comparison: mean peak percentage error.

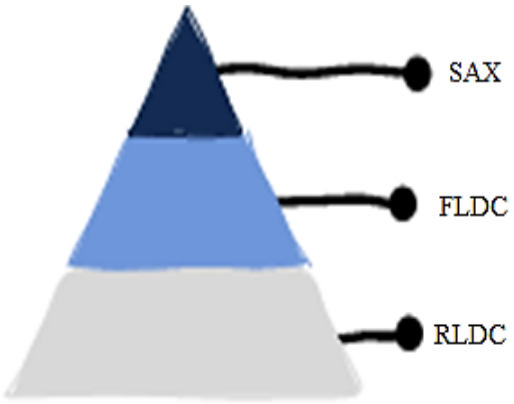

Figure 6. Compression-ratio objective pyramid (strong (dark blue), balanced (light blue), weak (gray)).

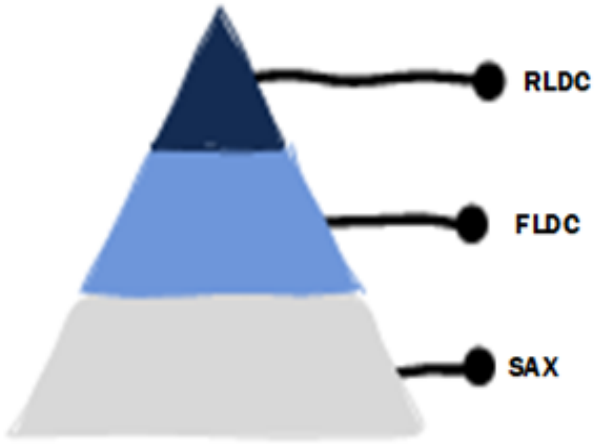

Figure 7. Lossless objective pyramid (strong (dark blue), balanced (light blue), weak (gray)).

In contrast, dictionary-based approaches use the identified repeated pattern structure in the data to eliminate redundancy using dictionary records as sliding windows. They process the source as the input of characters rather than as a stream of bits. This approach is focusing instead on the computer's memory ability to recall the strings already seen. Dictionary-based methods are faster than entropy-based methods. Popular in the fundamental form are LZ77 (Lempel-Ziv 1977), LZ78 (Lempel-Ziv 1978), and LZW (Lempel-Ziv-Welch, 1984). These algorithms have many variants which can be accessed using the 7-Zip package in either Windows or Linux operating systems (Jumar, et al., 2018).

For lossless EED compression, the work of (Gerek & Ece, 2008) is seminal. The authors introduced lossless codecs originally developed for audio and image signals for power quality event data, such as LZP2, FLAC, TTAEnc, ZIP, JPEG-LS, and JPEG2000-lossless. Abuadbba *et al*. (2018), for smart meter readings, developed a lossless compression algorithm based on a gaussian approximation and arithmetic coding statistical perspective. Considering energy-constrained smart meter hardware, the work of Ringwelski *et al*. (2012) compared statistical (Adaptive Trimmed Huffman Coding and Adaptive Markov Chain Huffman Coding). They also compared dictionary (Tiny Lempel Ziv Markov Chain Algorithm) and mixed statistical-dictionary (Lempel Ziv Markov Chain Huffman Coding) based lossless compression algorithms. The algorithms were designed with low computational complexity (a small footprint of program memory and therefore lesser processing time). Ringwelski *et al*. (2012) applied these algorithms to low-frequency smart meter electrical data. They concluded that the statistical methods were much faster but had a lower compression ratio than the dictionary-based methods.

Unterweger and Engel (2015; 2016) proposed a Resumable Load Data Compression (RLDC) technique that allows for resumability on low-frequency electric-power load profile datasets. They also proposed this technique for the lossless compression of high-frequency EED from the smart grid. This technique is called differential exponential Golomb and arithmetic (DEGA) coding. Their method combines normalization, entropy coding, differential coding, variable-length encoding (adaptive Golomb (rice) coding), and binary arithmetic coding. This compression approach was proposed for data transmission, promising higher compression ratio performance compared to other lossless methods. In Sarkar, *et al*., (2018), the differential binary arithmetic coding method was extended for power system operational (low frequency) EED when data storage was considered. The method had a slightly lower compression ratio with an advantage of low algorithmic complexity for low-frequency power-grid datasets.

Efficient lossless compression algorithms exist for audio, video, and general-purpose text data. However, apart from the RLDC, no EED-specific lossless compression method can exploit the encountered EED waveform's strong, periodic behaviour and multichannel characteristics (Jumar, et al., 2018).

A. Sound and Sampling Frequency

Sound waves are complicated phenomena. Nominally, a sound wave is caused by a moving object in air or any other medium, and the output of recording and reproducing sound is usually called audio. Sound can be viewed as the propagation of pressure waves by the vibrations of molecules. It is often described as a time-dependent function, instantaneously measuring the pressure of a medium which can be represented as a periodic electrical signal in the form of the sum of sine or cosine waves.

An AC electrical signal inherently possesses a characteristic sound (Dukish, 2009). The typical sound is a digitized waveform, which essentially represents an electrical voltage data measurement through a sampling process. These numbers (binary, hex, or decimal) are saved in an audio file format. A significant advantage of digital audio is the high noise immunity capability it presents (Firmansah and Setiawan, 2016).

Computers sense sounds using a sound card, which converts incoming sounds to electrical signals with numeric values through sampling into digital form. Computers sample the signal by measuring its amplitude at a fixed periodic interval, often 44,100 times (44.1 kHz). The sampling process is an integral part of analog-to-digital conversion (ADC). Each measurement is stored with a fixed precision, often 16 bits and in a predefined format. A microphone is another device that receives sound and converts it to an electrical voltage waveform in a form suitable for the sound card. Each audio sample is a digital number whose value is proportional to the instantaneous voltage amplitude at the current sampling time. Audio sampling in the communication and signal processing literature is most often a pulse-code-modulation (PCM) process.

The choice of sampling frequency is critical for the reconstruction of an original signal. On the other hand, the bit rate can be viewed as the storage of bits required for each second of sound. For example, audio with 44,100 samples per second and 16 bits per sample, the bit rate is approximately 0.71 million bits per second. An audio with a

sample rate of 48kHz and 16-bit PCM data file will have about 0.768 million bits/seconds/channel.

Like most digital data, audio data depends on two essential factors: the sampling frequency (how many times should a sound wave be sampled each second?) and the sample size (how large /how many bits) should each sample be?) It is stated that the optimum sampling rate should be at least twice the maximum frequency of the data or signal (Shannon and Weaver, 1998).

B. Analog to Digital Converters

Modern computers, at the core have 8-bit storage units (bytes) (Dukish, 2009) . If each audio sample is a byte, there can be $2^8 = 256$ sample sizes, 256 different amplitudes. For instance, if the highest voltage produced by a microphone is 1 volt, then 8-bit audio samples can recognize voltages as low as 2^{-8} volts or four (4) millivolts. Any sound converted by the microphone to a lower voltage would result in audio samples of zero amplitude and then becomes played-back as silence. Most ADCs create 16-bit audio samples. Such a sample can have $2^{16} = 65{,}536$ values, so it can recognize sounds as low as approximately 15 microvolts. In this case, eight-bit samples correspond to a coarser quantization, while 16-bit samples are finer quantization. Therefore, the better the quantization, the better the played-back sound quality. Therefore, with the information in the preceding sections, the sizes of sampled (audio) files can be estimated, thus revealing why compression is important. For example, a 3-minute recorded mono audio results in 180 × 44,100 = 7,938,000 samples. This translates to about 16 Mb, bigger than most still-images for 16-bit samples. A 30-minute recording would be a file size of about 160 Mb.

C. EED and Sound

Because of the way the human auditory system works, lossy compression methods are popular in audio. The human ear can typically not recognize higher frequency sounds that animals like dogs hear. Therefore, lossy techniques use perceptual limitations in humans to discard irrelevant information. Large storage requirements limit the amount of data that can be stored, hence an interest in shrinking the storage requirements of sampled sound. While lossy methods are more established than lossless methods, it is clear that the higher the CR of a lossy method, the lower the original data quality. Unfortunately, because of the maximum limit that can be reached without losing information in data, there is a limit to lossless CRs.

General-purpose Dictionary-based lossless compression methods (LZ77, Lempel-Ziv and its variants) that usually provide good compression performance are poorly matched to the statistical features of binary audio data streams. They generally lead to poor performance (Maher, 2003). In contrast, audio-specific methods can achieve as low as 30-per cent of the original file size. Dictionary methods use the advantage of periodicity in a data stream. Although roughly periodic and consistent, audio waveforms are not repetitive in samples due to the asynchronous relationship between the waveform period and the sample rate and other disturbances. However, the degree of sample-to-sample correlation can be taken advantage of by linear predictive coding (typically FIR filters) methods (Pu, 2006).

Lossless compression algorithms are statistical-based methods appropriate for oscillating high-frequency audio data or signals with low entropy characteristics. In applications requiring perfect lossless waveform compression, the advantage of statistical-based compression algorithms outweighs the general-purpose dictionary-based compression algorithms and lossy methods (Maher, 2003). High-frequency electrical energy data (EED) are also oscillating; therefore, audio-based compression techniques have had good performance on them (Haq (2018). Notwithstanding, certain factors can still affect the CR of oscillating high-frequency signals, such as signals with a high entropy value. This connotes high variance in the signal and the presence of much noise (Gray, 2011). In this case, the compression algorithm will have difficulty finding redundancies, thereby resulting in a poor CR performance of the algorithm.

The encoding type and dataset size can also impact compression performance. Some compression algorithms show better CR performance on larger datasets, but others perform best on small datasets. In all, the performance of data compression varies largely with the data's characteristics, and the method applied. They showed that different lossless compression codecs (algorithms) perform similarly despite algorithm complexity and approach (Hans and Schafer, 2001).

In audio samples, the primary redundancy is that adjacent audio samples tend to be correlated. Therefore, statistical compression techniques subtract each adjacent sample and encode the differences (errors or residuals) as mostly small integers with appropriate variable-length codes. The Rice codes are the choice for this task (Sayood, 2018). Practical methods often follow this procedure: (a) predict current sample using a weighted-sum of several neighbouring samples; (b) then subtract the current sample from this prediction. Smaller residuals of integer type ensure efficient encoding (for example, if the residual integer values are in the interval [−1, 4]. This implies that six (6) residual values and only six (6) variable-length codes would be needed. This improves the justification for the choice of very short codes for the coding process. The FLAC codec is an example of this approach (Salomon, 2008).

The essential operation in lossless compression methods, given in Hans and Schafer (2001), is briefly highlighted below. It involves Framing, Intrachannel Decorrelation (which involves error calculation), and Entropy coding.

The basic principle is first to remove redundancy from the signal and then code it by decorrelating and using an adaptive linear predictive model or a lossy coding model. Entropy coding removes redundancy from the residual error signal, and there are three adaptive (variable-length) methods used: Huffman, Run-length, and Rice coding.

D. Lossless Codecs and Performance Comparisons

Jumar *et al*. (2018) investigated the challenges of handling large publicly available raw EEDs with quasi-periodical characteristics via lossless compression through

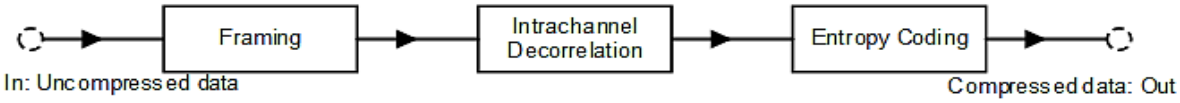

Figure 8: Fundamental statistical-based coding principle.

audio-based and dictionary-based algorithms. The authors highlighted that the usual way to record EED logged as audio waveforms are to use the *'.riff'* or *'.wav'* format of 16-bit signed integer representations. The publicly available *'sox'* library can then be used, which has utilities for post-processing: separating the electric current channel and electric voltage channel and then compressing (Kelly and Knottenbelt, 2015).

There are many audio compression software libraries available. Some are free, while some are commercial. Many lossless codecs, including FLAC (Free Lossless Audio Codec), Apple's ALAC or ALE, Shorten, Monkey's Audio, and MPEG-4 ALS algorithms, use fundamentally the same principle highlighted in the previous section. Also, selected lossless audio codecs used to compress EED in the literature will be discussed in the next section.

Muin *et al*. (2017) compared different freely available lossless codecs using compression ratio and computation time to find the most suitable audio waveform compression strategy. Codecs such as MPEG-4 variants, CELP, and IEEE 1857.2 were considered. They concluded that FLAC has superior compression ratio, encoding speed, and decoding speed for larger-sized files compared to MPEG-4, which, although not as popular but is generally faster in terms of encoding and decoding speed. Also, the newer IEEE 1857.2 audio codec standard, which uses arithmetic coding, compared competitively to another lossless audio compression in terms of higher compression ratio with negligible additional encoding/decoding time and average computational complexity. Unfortunately, the IEEE 1857.2 is not easily accessible in production-ready form for public use.

Interestingly, lossless audio codecs have surfaced and have been refined so much that they possess comparable performance. Their only differentiating factors are the adoption rate by different groups (Muin, *et al*., 2017). In addition, the literature has claimed that lossless audio codecs have very comparable bit rate reduction performance. They indicate that extending the performance state of lossless compression techniques has hit its practical limit defined by Shannon's entropy law (Salomon, 2008).

Further, Muin *et al*. (2017) also identified the gap or absence of a lossless audio compression algorithm having the characteristics of being high in compression ratio and fast encoding-decoding speed. Encoding-decoding speeds are of impact to data storage and transmission.

In addition, the work of Haq (2018) compared some off-the-shelf audio-based such as ALAC, ALS, APE, FLAC, TrueAudio, with dictionary-based lossless codecs such as LZMA, PPMd (prediction by partial matching), BZip2, and Gzip (a Deflate-variant). They investigated the EED measured at different sampling frequencies. The key findings of this thesis are as follows: One, for monitored energy data, if the Objective is real-time disaggregation, then the processing time is the defining factor. In contrast, for offline disaggregation, the CR is the defining factor. There is a significant difference in the CRs of electric current waveforms (uneven and spiky) compared to voltage waveform data (usually smoother) which are traditionally better compressed.

General-purpose dictionary codecs: Bzip2, PPMd-variants, Deflate-variants, LZMA-variants are generally much faster but gave lesser compression ratios (Wen, et al, 2018). Jumar *et al*. (2018) recommend TTA and MP4 ALS, APE as audio codecs exhibiting good performance for EED compression. FLAC can be used for fast decompression but large performance variations have to be expected. They also recommended that the LZMA is representative of an overall best-performing general-purpose codec. Figure 9 shows the recommended off-the-shelf audio-based lossless codecs for EED lossless compression using the discussed literature.

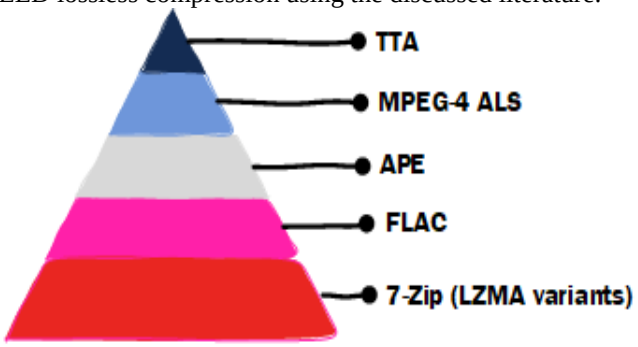

Figure 9. Top off-the-shelf lossless codecs recommended for EED lossless compression.

E. Resumable Load Data Compression (RLDC)

Resumable Load Data Compression (RLDC) is an EED-specific technique for lossless compression. It is also known as Differential Exponential Golomb and Arithmetic (DEGA) coding. RLDC's performance was compared in Section 3.2 with the FLDC and it was inferred that a small consecutive value difference (CVD) in load profiles can be exploited to improve compression efficiency and error. Load profile data is time-series EED, primarily representing measured electrical power consumption. Using publicly available low-frequency (1-second sampling) measured EED datasets, Unterweger and Engel (2015) conducted a detailed investigation of load profiles in residential consumer

households. It was discovered that such data mostly exhibit small CVD. The authors then constructed a lossless compression technique using statistical-based coding methods to take advantage of this characteristic. The DEGA technique consists of five straight-forward processes (Unterweger and Engel, 2015).

1) Normalization

Herein, this word entails converting floating-point data representation of the input load-profile data values v_i to integer form n_i. The justification for this is that floating-point operations are computationally expensive than integer operations, especially in low-cost embedded systems. The integer normalization formula, also known as A-XDR coding, is

$$n_i = v_i \times 10^{p_{\max}} \tag{8}$$

where, $p_{\max}$ is the maximum number of decimal places among the v_i data values of the load profile.

2) Differential Coding

The small CVD can be exploited using differential coding, which involves storing only the consecutive difference d_i between two consecutive values instead of the actual values. This method is also used in differential pulse code modulation (Unterweger & Engel, 2015). Only the first ith value is not replaced by a difference value, such that $d_0 = n_0,\ d_i = n_i - n_{i-1},$ where $i = 1,2, \ldots k$.

3) Variable-length Coding

Still exploiting the SCVD, a variable-length code in the form of the exponential-Golomb code, which is also used in many other techniques such as the FLAC codec and the H.264 standard, is then used to convert the signed integer values d_i to signed exponential Golomb codeword (binary) c_i. The c_i values are then concatenated to form a single bit string b.

4) Adaptive Binary Arithmetic Coding

Finally, an adaptive binary arithmetic coding scheme applies entropy coding on the concatenated bit string b, to remove the remaining information redundancy in b, and outputs the compressed bit representation e.

As illustrated in Figure 10, the decoding process of the compressed bit string e is just an inverse operation of the coding process in reverse order to obtain v, such that:

$$v_i = \frac{n_i}{10^{p_{\max}}} \tag{9}$$

Unterweger and Engel (2015) also show that the technique has comparatively low memory requirements and computational complexity, with high CR. This was illustrated with the FLDC in Figure 4 and Figure 7. They also explain how the technique allows for resumability with low-overhead of data loss in case of transmission interruption, which will enable it to be applicable in error-prone transmission lines in smart grids.

This method gives attractive results for low-frequency EED. Unterweger and Engel (2016) applied this technique to high-frequency EED (16kHz and 50kHz). Some salient conclusions are as follows: First, due to the considerable noise (entropy) in electric-current data, such EED can be compressed better than electric-voltage data. DEGA gives poor CR performance concerning high-frequency data. Good CR performance was achieved for data sampled at 100 Hz and below. Lossless codecs like FLAC and Bzip2 are suitable for lossless archival of EED for further processing.

Admittedly, it is clear that the development of standard, high-quality compression techniques (lossy or lossless) for high-frequency sampled time-series EED remains an open question for intelligent energy networks known as smart-grids (Unterweger and Engel, 2016). It will be beneficial to have a range of standard EED coding techniques even as audio, image, and video coding exists.

V. CHALLENGES AND FUTURE WORK

Research on smart-grid LM analysis has spanned three decades (Wang *et al*., 2020). Load-event classification and smart-grid analysis are essentially data-driven research processes. In this case, EED compression can ease the practical challenge encountered with storing and transmitting a large amount of EED-specific dataset files with minimal loss in data quality. In this context, the goal for EEDs is to achieve efficient use of the bandwidth of the communication channel and reduced storage space.

To the best of the authors' knowledge, the work of Tcheou *et al*. (2014) is the first review to highlight the importance and challenges of the timely problem of compressing the growing data stream of electrical energy signals with the advent of smart grid networks and for the main contributions in the smart-grid literature. Still quite true currently is the authors concluding remark that: the compression of EED obtained from electric-power systems is far from being as mature compared to speech, audio, image, and video compression. In addition, comparing existing literature, varying accounts are sometimes presented on the performance of the lossy and especially lossless codecs. Consequently, for gains of compression to become manifestly profound, there needs to be a new wave or direction of more research into the design and development of production-ready compression techniques targeted at electric-energy data (EED).

Although, practical compression issues revolve around concepts like numerical implementation and portability, segmentation or framing, variable bit rate, speed, and complexity (Maher, 2003). However, the main challenge with lossless compression techniques, especially for high-frequency EED, is increasing compression-ratio performance. For lossy compression, it is to reduce loss in the quality of recovered data. Also, a caveat in compression codecs (algorithm implementations) is that there are no actual guarantees on compression performance. This implies that using a compression codec may lead to a larger file size than the original if the applied codec's input data stream is ill-posed. Most users may not notice this because, in

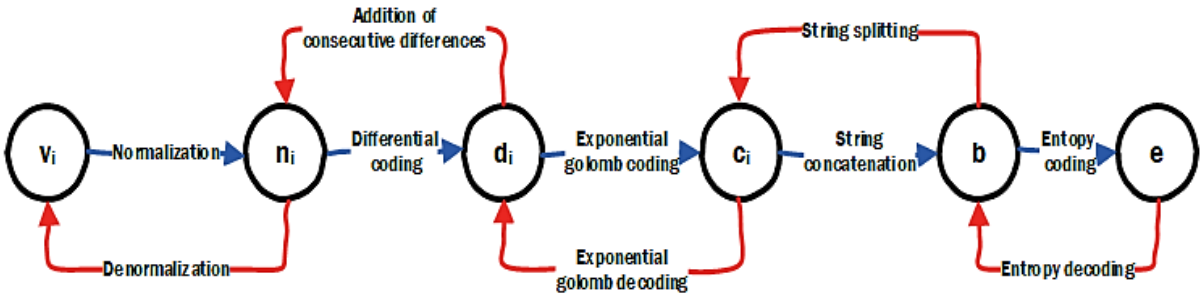

Figure 10: Coding (blue lines) and decoding (red lines) process of the DEGA technique.

commercially-developed libraries, codecs have conditional flags to detect when compression is ill-posed and stop the compression process (Hans and Schafer, 2001). It was also underscored in Hans and Schafer (2001) that lossless codecs had hit the perceived limit concerning the amount of CR that can be realised. This limit is stated in Shannon's entropy, limiting the size of data that can be compressed without losing information (Gray, 2011).

Shannon's entropy represents an absolute maximum theoretical limit on the possible compression of any data without losing any original information under certain constraints (Shannon and Weaver, 1963). Statistical techniques that use entropy treat data encoded as a sequence of independent and identically distributed (IID) random variables. Worth noting is that Shannon's source-coding theorem reveals that: in the limit, the average length of the shortest possible representation used to encode data streams in a given alphabet is the entropy divided by the logarithm of the number of symbols in the target alphabet. Therefore, because of this theoretical limit, the current CR performance level obtained from lossless codecs may never reach that of the lossy codecs.

VI. CONCLUSION

This work has covered a review of relevant literature on EED compression techniques in load monitoring analysis. Also highlighted are the challenges of selecting top-performing EED compression techniques. This paper will contribute to motivating interest in the development of standard, high-performance EED compression techniques. Overall, this review attempted to provide a concise reference overview supporting researchers interested in designing and developing compression methods targeted at electric-energy data (EED) for smart-grid applications. The challenge of devising very efficient and high-performing compression methods and algorithms with very low loss, especially generalizing to low-frequency and high-frequency real-time EED, cannot be overstated regarding the future smart grid.

REFERENCES

Abuadbba, A.; I. Khalil and X. Yu (2018). Gaussian Approximation-Based Lossless Compression of Smart Meter Readings. IEEE Transactions on Smart Grid, 9: 5047 - 5056. doi:10.1109/TSG.2017.2679111.

Atif, S. M.; S. Qazi and N. Gillis (2019). Improved SVD-Based Initialization for Nonnegative Matrix Factorization Using Low-Rank Correction. Pattern Recognition Letters, 122:53-59. doi:10.1016/j.patrec.2019.02.018.

Basu, K. (2015). Classification Techniques for Non-Intrusive Load Monitoring and Prediction of Residential Loads. Ph.D. Dissertation Université de Grenoble, France.

Batra, N.; H. Dutta and A. Singh (2013). Indic: Improved Non-Intrusive Load Monitoring Using Load Division and Calibration. 2013 12th International Conference on Machine Learning and Applications, Washington DC, USA. 1, 79 – 84, USA: IEEE.

Bellasi, D.; M. Crescentini; D. Cristaudo; A. Romani; M. Tartagni and L. Benini (2019). A Broadband Multi-Mode Compressive Sensing Current Sensor SoC in 0.16 μm CMOS. IEEE Transactions on Circuits and Systems I: Regular Papers, 66: 105 - 118. doi:10.1109/TCSI.2018.2846573.

Chandak, S.; K. Tatwawadi; C. Wen; L. Wang; J. Aparicio and T. Weissman (2020). LFZip: Lossy Compression of Multivariate Floating-Point Time Series Data via Improved Prediction. 2020 Data Compression Conf. (DCC), 342–351. UT, USA. https://doi.org/10.1109/DCC47342.2020.00042.

Clark, M. and Lampe, L. (2015). Single-Channel Compressive Sampling of Electrical Data for Non-Intrusive Load Monitoring. 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Orlando, Florida, 790 – 794, USA: IEEE. doi:10.1109/GlobalSIP.2015.7418305.

Dukish, B. (2009). Extreme Fundamentals of Technology: A Primer of Computers, Electronics, and Technology. Fixtron Corporation.

Fagiani, M.; R. Bonfigli; E. Principi; S. Squartini and L. Mandolini (2019). A Non-Intrusive Load Monitoring Algorithm Based on Non-Uniform Sampling of Power Data and Deep Neural Networks. Energies, 12: 1371. doi:10.3390/en12071371.

Firmansah, L. and Setiawan, E. B. (2016). Data Audio Compression Lossless FLAC Format to Lossy Audio MP3 Format with Huffman Shift Coding Algorithm. 2016 4th

International Conference on Information and Communication Technology (ICoICT), pp. 1 - 5. doi:10.1109/ICoICT.2016.7571951.

Gerek, Ö. N. and Ece, D. G. (2008). Compression of Power Quality Event Data Using 2D Representation. Electric Power Systems Research, 78: 1047 - 1052. doi:10.1016/j.epsr.2007.08.006.

Gray, R. M. (2011). Entropy and Information Theory (Second ed.). Springer US. doi:10.1007/978-1-4419-7970-4

Hans, M. and Schafer, R. W. (2001). Lossless Compression of Digital Audio. IEEE Signal Processing Magazine, 18: 21 - 32. doi:10.1109/79.939834.

Haq, A. U. (2018). Appliance Event Detection for Non-Intrusive Load Monitoring in Complex Environments. PhD Dissertation, Technische Universität München, München, Germany.

Haq, A. U. and Jacobsen, H. A. (2018). Prospects of Appliance-Level Load Monitoring in Off-the-Shelf Energy Monitors: A Technical Review. Energies, 11 (1): 189, 1-22. doi:10.3390/en11010189.

Huang, X.; T. Hu; C. Ye; G. Xu; X. Wang and L. Chen (2019). Electric Load Data Compression and Classification Based on Deep Stacked Auto-Encoders. Energies, 12: 653. doi:10.3390/en12040653.

Jumar, R.; H. Maaß and V. Hagenmeyer (2018). Comparison of Lossless Compression Schemes for High-Rate Electrical Grid Time Series for Smart Grid Monitoring and Analysis. Computers and Electrical Engineering, 71: 465 - 476. doi:10.1016/j.compeleceng.2018.07.008.

Kelly, J. and Knottenbelt, W. (2015). The UK-DALE Dataset, Domestic Appliance-Level Electricity Demand and Whole-House Demand from Five UK Homes. Scientific Data, 2(150007): 1-14. doi:10.1038/sdata.2015.7

Le, X.-C. (2017). Improving Performance of Non-Intrusive Load Monitoring with Low-Cost Sensor Networks. Ph.D. dissertation.

Lendák, I. and Horvath T. (2019). Efficient Load Profiling and Forecasting in Large Electric Power Systems, Paper presented at 19th Conference on Information Technologies – Applications and Theory (ITAT 2019), Donovaly, Slovakia, 36 – 43, Slovakia: CUER-WS.

Maher, R. C. (2003). Lossless Audio Coding. In Lossless Compression Handbook. (1st Edition. ed.). Elsevier.

Muin, F. A.; T. S. Gunawan; M. Kartiwi and Elsheikh, E. M. (2017). A Review of Lossless Audio Compression Standards and Algorithms. American Institute of Physics Conference Proceedings, 1883, 020006, Bydgoszcz, Poland, 1-11, USA: AIP. doi:10.1063/1.5002024.

Nithiyananthan, K. and Ramachandran, V. (2014). Effective Data Compression Model for On-Line Power System Applications. International Journal of Electrical Energy, Engineering Modelling 27 (3-4): 101-109. doi:10.12720/ijoee.2.2.138-145.

Pu, I. M. (2006). Chapter 9 - Audio Compression. In I. M. Pu (Ed.), Fundamental Data Compression. 171 - 188. Oxford: Butterworth-Heinemann. doi:10.1016/B978-075066310-6/50012-X.

Ringwelski, M.; C. Renner; A. Reinhardt; A. Weigel and V. Turau (2012). The Hitchhiker's Guide to Choosing the Compression Algorithm for Your Smart Meter Data. 2012 IEEE International Energy Conference and Exhibition (ENERGYCON), Florence, Italy, 935 – 940, USA: IEEE.. doi:10.1109/EnergyCon.2012.6348285.

Rodriguez-Silva, A. and Makonin, S. (2019). Universal Non-Intrusive Load Monitoring (UNILM) Using Filter Pipelines, Probabilistic Knapsack, and Labelled Partition Maps, 2019 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC2019), Macao, China, 1-6, USA: IEEE.. Arxiv. arXiv:1907.06299 [eess].

Sadler, C. M. and Martonosi, M. (2006). Data Compression Algorithms for Energy-Constrained Devices in Delay Tolerant Networks. Proceedings of the 4th International Conference on Embedded Networked Sensor Systems, Colorado, USA, pp. 265 - 278, USA: Boulder: Association for Computing Machinery.. doi:10.1145/1182807.1182834.

Salomon, D. (2008). A Concise Introduction to Data Compression (1st Edition. ed.). Springer.

Sari, E. M. (2018). Data Compression for Smart Grid Infrastructure. Unpublished M.Sc Thesis, Department of Electronics Engineering, Isik University, Turkey

Sarkar, S. J.; P. K. Kundu and G. Sarkar (2018). Development of Lossless Compression Algorithms for Power System Operational Data. IET Generation, Transmission and amp; Distribution, 12: 4045 - 4052.

Sayood, K. (2017). Introduction to Data Compression (Fifth ed.). Morgan Kaufmann Series in Multimedia Information and Systems, Amsterdam, Netherlands.

Shannon, C. E. and Weaver, W. (1949). The Mathematical Theory of Communication. University of Illinois Press, Urbana, Illinois.

Tame, J. (2019). Approaches to Entropy. Springer Publishers, Singapore. doi:10.1007/978-981-13-2315-7.

Tariq, Z. B.; N. Arshad and M. Nabeel (2015). Enhanced LZMA and BZIP2 for Improved Energy Data Compression. 2015 International Conference on Smart Cities and Green ICT Systems (SMARTGREENS), Lisbon, Portugal, 1 – 8, Portugal, INSTICC.

Tcheou, M. P.; L. Lovisolo; M. V. Ribeiro.; E. E. Da Silva.; M. A. Rodrigues; J. M. Romano and P. S. Diniz (2014). The Compression of Electric Signal Waveforms for Smart Grids: State of the Art and Future Trends. IEEE Transactions on Smart Grid, 5: 291 - 302. doi:10.1109/TSG.2013.2293957.

Tong, X.; C. Kang, and Q. Xia (2016). Smart Metering Load Data Compression Based on Load Feature Identification. IEEE Transactions on Smart Grid, 7: 2414 - 2422. doi:10.1109/TSG.2016.2544883.

Unterweger, A. and Engel, D. (2015). Resumable Load Data Compression in Smart Grids. IEEE Transactions on Smart Grid, 6: 919 - 929. doi:10.1109/TSG.2014.2364686.

Unterweger, A. and Engel, D. (2016). Lossless Compression of High-Frequency Voltage and Current Data in Smart Grids. 2016 IEEE International Conference on Big Data (Big Data), Washington D.C., USA, 3131 - 3139, USA: IEEE. doi:10.1109/BigData.2016.7840968.

Wang, C. and Zhai, M. (2018). A Non-Intrusive Load Decomposition Method for Residents. IOP Conference Series:

Earth and Environmental Science, 199: 052034. doi:10.1088/1755-1315/199/5/052034.

Wang, Y.; Q. Chen and C. Kang (2020). Smart Meter Data Analytics: Electricity Consumer Behavior Modeling, Aggregation, and Forecasting. Springer Publishers, Singapore. doi:10.1007/978-981-15-2624-4.

Wang, Y.; Q. Chen; T. Hong and C. Kang (2019). Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. IEEE Transactions on Smart Grid, 10: 3125-3148. doi:10.1109/TSG.2018.2818167.

Wang, Y.; Q. Chen; C. Kang; Q. Xia and M. Luo (2017). Sparse and Redundant Representation-Based Smart Meter Data Compression and Pattern Extraction. IEEE Transactions on Power Systems, 32: 2142 - 2151. doi:10.1109/TPWRS.2016.2604389.

Wang, Y.; Q. Chen; C. Kang; Q. Xia; Y. Tan; Z. Zeng and M. Luo (2016). Residential Smart Meter Data Compression and Pattern Extraction via Non-Negative K-SVD. 2016 IEEE Power and Energy Society General Meeting (PESGM), Boston, USA, 1 - 5, USA: IEEE. doi:10.1109/PESGM.2016.7741464.

Wang, Y.; Q. Chen; C. Kang; M. Zhang; K. Wang and Y. Zhao (2015). Load Profiling and Its Application to Demand Response: A Review. Tsinghua Science and Technology, 20: 117 - 129. doi:10.1109/TST.2015.7085625.

Wang, Y.; A. Filippi; R. Rietman; and G. Leus (2012). Compressive Sampling for Non-Intrusive Appliance Load Monitoring (NALM) Using Current Waveforms. Signal Processing, Pattern Recognition and Applications / 779: Computer Graphics and Imaging. Crete: ACTAPRESS. doi:10.2316/P.2012.778-024.

Wee, C. K. and Nayak, R. (2019). An Approach to Compress and Represents Time Series Data and Its Application in Electric Power Utilities. In R. Islam, Y. S. Koh, Y. Zhao, G. Warwick, D. Stirling, C.-T. Li, and Z. Islam (Ed.), Data Mining, pp. 107 - 120. Singapore: Springer. doi:10.1007/978-981-13-6661-1_9.

Wen, L.; K. Zhou.; S. Yang and L. Li (2018). Compression of Smart Meter Big Data: A Survey. Renewable and Sustainable Energy Reviews, 91: 59 - 69.

Wong, Y. F.; A. Y. Sekercioglu; T. Drummond, and V. S. Wong (2013). Recent Approaches to Non-Intrusive Load Monitoring Techniques in Residential Settings. 2013 IEEE Computational Intelligence Applications in Smart Grid (CIASG). Singapore, 73 - 79, USA: IEEE. doi:10.1109/CIASG.2013.6611501.

Yan, L.; J. Han; R. Xu and Z. Li (2019). LIFTED: Household Appliance-Level Load Dataset and Data Compression with Lossless Coding Considering Precision, Virtual 2020 IEEE Power and Energy Society General Meeting (PESGM), 1-5, USA: IEEE. arXiv:1911.01581 [eess].

Zhuang, M.; M. Shahidehpour and Z. Li (2018). An Overview of Non-Intrusive Load Monitoring: Approaches, Business Applications, and Challenges. 2018 International Conference on Power System Technology (POWERCON), Guangzhou, China: 4291 – 4299, USA: IEEE.. doi:10.1109/POWERCON.2018.8601534.