IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

The Price of Undersampling Nonintrusive Electrical Data

ANDRÉS MARTÍNEZ [1], AARON W. LANGHAM [1], (Graduate Student Member, IEEE), JOHN S. DONNAL [2], (Senior Member, IEEE) and STEVEN B. LEEB [1], (Fellow, IEEE)

[1]Massachusetts Institute of Technology, Cambridge, MA 02139 USA
[2]United States Naval Academy, Annapolis, MD 21402 USA

Corresponding author: Andrés Martínez (e-mail: andresmt@mit.edu).

ABSTRACT Field data from power system measurements is frequently disappointing. Utilities are geographically large (in some cases remote), interact with a wide range of users and use profiles, and contain a burgeoning diversity of electrical loads. Meters and data collection systems for applications such as billing and phasor measurement are tailored to gather information at a price point specific to their intended use. Information recorded from these monitors can be surprisingly difficult to interpret and repurpose for other utility applications. Many measurement systems internally record data at higher rates than the rate at which they save or log final data. This paper demonstrates the value of implementing intelligent preprocessing of power data that preserves information needed for a particular application. We compare ''typical'' recorded data to revised datasets to examine the effect of targeted data preprocessing on machine learning and load monitoring applications. Spectral envelope preprocessing and V-I trajectories are used with raw data early in the measurement process to permit a flexible trade-off between sample storage rate and resolution. The UK-DALE dataset and newly acquired data from ship microgrids provide case studies for this work.

INDEX TERMS Nonintrusive load monitoring, data preprocessing, data analytics

I. INTRODUCTION

TRACKING energy consumption in buildings remains crucial for resource planning and policy development as industries and economic sectors move toward a sustainable energy future. Both residential and commercial buildings account for a significant share of total energy consumption in the United States and Canada [1]. As a result, many climate change mitigation strategies focus on energy efficiency and demand reduction [2], [3]. These strategies often require detailed end-user consumption data, making high-quality power monitoring critical.

Advanced metering infrastructure (AMI) is one common example of instrumentation installed for one purpose (billing) that also can give more detailed, second-to-second pictures of power consumption [4], [5]. Data from instruments like AMI meters, phasor measurement units, and smart circuit breakers have all been tapped as ''nonintrusive'' sources of consumption information. When pressed into service as nonintrusive monitors, these and other devices offer hoped-for savings in the expense of monitoring in comparison to metering individual loads. They provide a centralized source of data that, in principle, can be combined with recognition algorithms that disaggregate the behavior and condition of individual loads from the aggregate data stream [6]. In principle, accurate disaggregation enables smart grid techniques such as power monitoring, consumption prediction, and demand response control. Nonintrusive energy disaggregation algorithms process aggregate data to identify electrical signatures for each load, relying on methods such as pattern matching, source separation, or machine learning (ML) [7].

The relatively continuous data acquisition employed for nonintrusive monitoring, as opposed to much less frequent ''meter reading,'' produces troves of historical data that quickly become unwieldy. AMI meters typify a painful trade-off in collecting data for nonintrusive monitoring. Depending on implementation, the meter internals may sample sensed waveforms at relatively high speeds, e.g., many times per second. However, monitoring studies using data sources like AMI meters typically rely on saved output data from the meter recorded at a slower rate, usually at most once per second or slower. Just as a low-resolution image offers incomplete information about a visual scene, low-resolution electrical data fails to capture fine-grained temporal features. These limitations challenge effective load identification. They may hobble efforts to extract more nuanced interpretations of the

data to support anomaly detection, predictive maintenance, and detailed energy analytics.

This work provides an investigation into what is lost when electrical data is insufficiently sampled, using the open-access UK-DALE dataset [8] and new field data collected from shipboard power systems used as practical examples of microgrids. We apply spectral envelope preprocessing to the raw current and voltage data (sampled at the kHz level) to produce in-phase, quadrature, and harmonic spectral coefficients sampled at the utility frequency (50 Hz or 60 Hz). Careful preprocessing permits a flexible trade-off between the retention of information and storage requirements. For example, high-resolution data processed as spectral envelopes can break data into multiple channels at sample rates more modest than the initial recording rate. This approach permits a flexible trade-off between memory storage requirements and bandwidth preserved in a final data set. This paper compares spectral envelopes derived from raw UK-DALE data with processed aggregate power data (sampled once every 6 seconds), revealing that a staggering amount of interpretable and actionable load behavior information is discarded when data is downsampled to the rate at which this data is often analyzed in the literature.

The contributions of this paper are threefold: (i) we demonstrate specific mechanisms by which undersampling can distort or omit NILM-relevant features in widely used data, including transient cluster spread, aliasing-induced spurious periodicity, and loss of harmonic and geometry signatures; (ii) we show that spectral envelopes and V–I trajectories enable an information-preserving trade-off between storage and resolution by compressing kHz measurements into compact per-cycle representations; and (iii) we offer a practical approach for sampling-rate selection grounded in load dynamics, highlighting when low-rate, meter-like data is sufficient and when higher-rate acquisition is needed.

This paper is organized as follows: Section II provides background on power computation and preprocessing techniques. Section III shows several critical events and features lost when undersampling using the UK-DALE dataset as a case study. Section IV illustrates the shortcomings of undersampled data for pattern recognition and load identification. Finally, Section V concludes.

II. PREPROCESSING POWER

Nonintrusive electrical monitoring usually starts with sampling voltages and currents at an aggregated service point. These waveforms serve as raw data with maximal information content, similar to raw image files from a camera or I/Q data from radio spectrum monitoring. However, power system waveforms may be highly redundant. As an example, one period of a 60 Hz sinusoid can be described with several thousand samples. Alternatively, this sinusoid can simply be described with an estimated amplitude and phase, as shown in Figure 1. Strictly speaking, the raw data contains "just the facts" at each sampling point and makes no assumption about the character of the data source. By contrast, the preprocessed data contains parameter estimates that rely on the assumption that the data source is a mostly pure sinusoid. Appropriate choices for processing power system data can preserve valuable information without necessarily demanding excessive data storage. For example, assuming the sinusoidal nature of the voltage data source reduces the dimensionality of the data from N (the number of samples in a period) to 2. In addition, these dimensions (magnitude and phase) are highly relevant to electrical power analysis. It is important to note that this preprocessing technique is *subtractive* in nature, meaning that information is lost when it is applied. For waveforms that do not match the sinusoidal assumption, it will produce misleading outputs. When waveforms do in fact approximate the assumption, the savings in data storage and transmission requirements can be impressive. A wise choice of preprocessing will avoid losing valuable information while minimizing storage and transmission requirements.

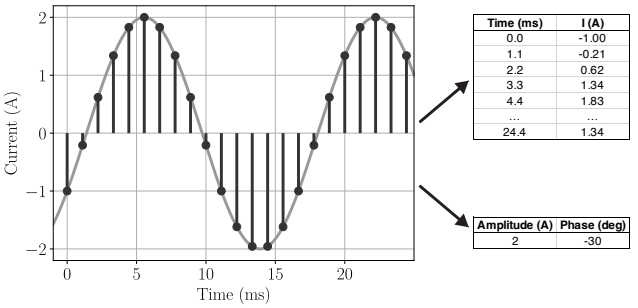


FIGURE 1. Two data structures describe the current waveform on the left. The top record is raw sample data, and the bottom record is an estimate of the amplitude and phase from the raw data via preprocessing. The preprocessed data structure occupies much less space than the raw data, and contains feature relevant to electrical monitoring.

Preprocessing techniques for nonintrusive electrical monitoring serve two purposes. Data is converted to a lower-dimensional representation more amenable to bulk storage and downstream processing. In addition, currents and voltages are combined to produce electrical power, typically the quantity of interest in downstream processing. This section reviews different approaches that explore the trade-off between preserving information and generating excessive data.

A. TIME-AVERAGED REAL AND APPARENT POWER

The element-wise product of voltage and current shows the instantaneous flow of energy through the aggregate monitoring point. However, on a 60 Hz grid this quantity contains a constant component and a component "pulsing" at 120 Hz. By contrast, time-averaged power quantities provide a coarser view of energy flow that does not oscillate over time in steady state. Each period of the utility waveforms is sampled N times. Averaging the product of the voltage and current waveform over these N samples yields one sample of the average real power P, i.e., a commonly computed "metering"

metric for power consumption:

$$P = \frac{1}{N}\sum_{n=0}^{N-1} v[n] \cdot i[n]. \tag{1}$$

One sample of P is created for every waveform period. This preprocessing method reduces the data rate by a factor of N. Measurements of real power are useful for billing, fuel consumption estimation, and measuring energy conversion between systems. However, some components of the current drawn may contribute to reactive power and harmonic distortion but not carry any real power. The time-averaged apparent power S includes the effects of these waveform components. By taking the product of the root-mean-square voltage and current, as computed over N points:

$$S = \frac{1}{N}\sqrt{\left(\sum_{n=0}^{N-1} v^2[n]\right)\left(\sum_{n=0}^{N-1} i^2[n]\right)}. \tag{2}$$

These techniques reduce the data rate from $2N$ samples per period (N samples for each waveform) to 1 sample per period.

Time-averaged real and apparent power streams provide a low-data-rate summary of load composition. From these quantities, the power factor (defined as P/S) can be estimated. Inductive and capacitive loads introduce a lag or lead into the current waveform, lowering the power factor via a *displacement factor* [9]. Nonlinear loads that produce harmonic currents also lower the power factor via a *distortion factor*, since no net energy is transferred at frequencies other than the utility frequency. However, P and S do not preserve information on harmonic current composition and the lag angle of the current.

B. SPECTRAL ENVELOPES

Spectral envelope preprocessing extends the benefits of time-averaged power quantities to preserve reactive and harmonic waveform characteristics [10]. If the voltage waveform is sinusoidal with peak value V_{pk}, $v[n]$ can be replaced with $V_{pk}\sin(2\pi n/N)$ in Eq. (1). This equation can then be interpreted as a computation of the in-phase fundamental component of the discrete Fourier transform (DFT) of $i[n]$, scaled by V_{pk}. This produces the in-phase or real fundamental power spectral envelope P_1:

$$P_1 = \frac{V_{pk}}{N}\sum_{n=0}^{N-1} i[n] \cdot \sin(2\pi n/N), \tag{3}$$

An analogous computation with a cosine instead of a sine produces Q_1, the quadrature or reactive fundamental power spectral envelope:

$$Q_1 = -\frac{V_{pk}}{N}\sum_{n=0}^{N-1} i[n] \cdot \cos(2\pi n/N). \tag{4}$$

These two quantities preserve information on the lead or lag between the current and voltage waveforms. This is particularly useful for identifying electric machinery that draws significant reactive power with a lagging current waveform.

To preserve harmonic information, the previous computations can be performed with the sine and cosine waves at a harmonic multiple k of the utility frequency, producing P_k and Q_k:

$$P_k = \frac{V_{pk}}{N}\sum_{n=0}^{N-1} i[n] \cdot \sin(2\pi kn/N), \tag{5}$$

$$Q_k = -\frac{V_{pk}}{N}\sum_{n=0}^{N-1} i[n] \cdot \cos(2\pi kn/N). \tag{6}$$

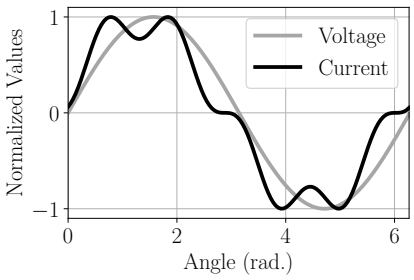
Many ac loads draw currents that are half-wave symmetric, meaning the positive and negative components are equal in size and opposite in sign. These half-wave symmetric waveforms contain no even harmonic components [9]. In addition, common waveform shapes such as square waves and triangle waves have Fourier series components that decay as k increases [10]. For practical purposes, k can therefore be restricted to a limited range of odd values, such as {1, 3, 5, 7} to capture the fundamental and first three odd harmonics. Saving $P_1, Q_1, ..., P_7, Q_7$ every period results in a data rate of 8 samples per period.

In addition to preserving fine-grained reactive power and harmonic information, spectral envelopes have enhanced quantization resolution compared to the quantized values of $i[n]$ [11]. To account for changing grid voltage amplitudes, spectral envelopes can be augmented to either cancel or preserve the resulting changing load power profile [12], [13]. Fundamental spectral envelopes can serve as phasors in traditional power system analysis, allowing ''derived'' streams such as symmetrical components on 3-phase ac grids to be computed [14].

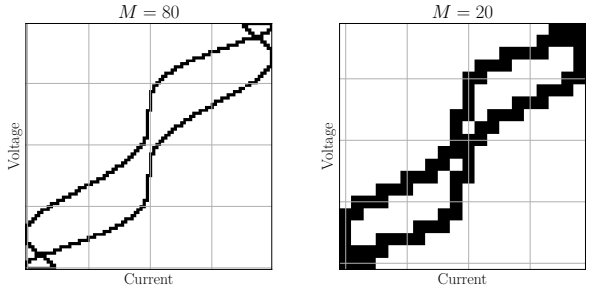
C. V-I TRAJECTORIES

Voltage-current (V-I) trajectories have emerged as a popular preprocessing method [15]. Rather than extracting power quantities, this method generates a graph of the ordered pairs $(v[n], i[n])$ from the voltage and current waveforms respectively, also known as a Lissajous curve. For linear loads that do not draw harmonics, $v[n] = V_{pk}\sin(2\pi n/N)$ and $i[n] = I_{pk}\sin(2\pi n/N - \phi)$, where ϕ is the lag angle between the current and voltage waveform. This generates an ellipsoid whose angle and axes are determined by V_{pk}, I_{pk}, and ϕ. When $\phi = 0$ the resulting curve is a line, and when $\phi = \pi/2$ the resulting curve is an unrotated ellipse. This graph is then discretized into a matrix that acts as a rasterized image. General-purpose computer vision models can be fine-tuned to classify these images to the respective load that created them [15]–[18].

When the graph is discretized into an image, the amount of information loss depends on the pixel size. In the theoretical case with infinitely small pixels, all of the information from $v[n]$ and $i[n]$ is preserved. However, as the pixel size increases, small spatial features in the V-I trajectory corresponding with harmonics are lost. When discretizing a V-I trajectory into an $M \times M$ matrix, the resulting output data rate will be M^2 samples per period. To illustrate, Figure 2a shows example



(a)



(b) (c)

FIGURE 2. **Constructing a voltage-current (V-I) trajectory for a load with fifth-harmonic current: a) Raw voltage and current waveforms, b) V-I trajectory discretized and thresholded into a grayscale image with 80 discretization points, and c) V-I trajectory image with 20 discretization points.**

Technique	Data rate (per period)	Runtime (per cycle)	Utilization at 50 Hz
Raw waveform data	$2N$	-	-
Time-averaged real and apparent power	2	16.30 μs	0.0815%
Spectral envelopes with K coefficients	$2K$	11.52 μs	0.0576%
V-I Trajectories with M discretization points	M^2	31.72 μs	0.1586%

TABLE 1. **Comparison of data rates and runtimes across preprocessing techniques given in Section II. The preprocessing runtimes represent the average of the fastest 10 runs and were measured using a single CPU on a M4 pro Apple processor. Utilization reports the fraction of the 50 Hz real-time budget used.**

voltage and current waveforms for a load that draws fifth-harmonic current with a lead angle of $\pi/12$. Plotting and rasterizing the V-I trajectory with $M = 80$ yields the image in Figure 2b. The loop-like features in the top-right and bottom-left corners are due to the fifth-harmonic content in the current waveform. Using a lower value of $M = 20$ yields the image in Figure 2c. Although the general "figure eight" shape is the same as in Figure 2b, the loops from the fifth harmonics are not preserved in the discretization process.

Conventional power computation (e.g., time-averaged real and apparent power) requires only per-cycle dot products and rms calculations with linear complexity in the number of samples per cycle. Spectral envelopes add harmonic channels but can still be computed efficiently by applying one fast Fourier transform (FFT) to each line-cycle window and retaining only the first few odd harmonics, yielding a per-cycle cost on the order of one FFT plus lightweight bin extraction. For the UK-DALE waveform stream (16 kHz at 50 Hz, i.e., 320 samples per cycle), real-time operation requires processing 50 windows per second. To make this practical constraint explicit, Table 1 shows preprocessing runtimes for multiple techniques under an identical windowing setup.

III. UNDERSAMPLING

For electrical monitoring, the quality of the output data is highly dependent on the sampling rate. Too low of a sampling rate causes interesting electrical phenomena to be distorted, aliased, or eliminated from monitoring purposes. However, an excessively high sampling rate creates computational problems. If a real-time monitor is unable to process a window of data before the next window of data arrives, incoming data will gradually accumulate in memory. Eventually, the system will run out of memory and crash. In addition, too high of a sampling rate means that data may exceed storage capacity. In systems that store data locally for retrieval and periodic upload to cloud storage, the size of the data collected between retrieval times must be less than the local storage capacity.

In addition to computing constraints, the dynamics of the monitored system should inform the choice of sampling rate. The Nyquist criterion guarantees perfect reconstruction when sampling at twice the system's maximum frequency. For example, a 60 Hz power system without harmonics can be perfectly reconstructed by sampling above 120 Hz. However, system frequencies may vary across line periods. The power system's frequency may be 60.01 Hz at one instant and 59.99 Hz the next. Techniques such as synchronous sampling effectively use a variable sampling rate to align samples with signal phenomena such as zero crossings. Finally, key information often lies in short-lived phenomena like motor inrush currents. Rather than continuously saving data at a high sampling rate, a low-rate continuous stream can be complemented with short high-rate sampling triggered by event detection. In this work, spectral envelopes computed synchronously for each line cycle serve as a practical compromise between computational constraints and preserving informative load dynamics.

In [7], many widely used electrical monitoring datasets are described, including UK-DALE [8], REDD [19], and PLAID [20]. These datasets have been foundational to the development and validation of numerous load disaggregation techniques. Many of these datasets provide high-bandwidth (kHz-level) raw waveform data and low-bandwidth (Hz-level) power data. Most nonintrusive load monitoring (NILM) literature uses the low-bandwidth versions of these datasets due to practical computational constraints. However, the trade-offs in resolution and pattern recognition between the high- and low-bandwidth datasets have largely been unexplored.

The UK-DALE dataset [8] provides an excellent case study, both in its original form and in a "remastered," high-bandwidth form presented here. The UK-DALE dataset contains the mains current and voltage data for multiple homes

sampled at 16 kHz and aggregate and sub-metered apparent and real power streams at one sample every 6 seconds (1/6 Hz). The 16 kHz raw waveform data provides a view of the electrical behavior on an extremely fine timescale, with 320 samples per waveform period. However, as a matter of practicality, this data grows prohibitively large on long timescales. Even in compressed form, the 16 kHz dataset, recorded from five houses (one of them tracked for 655 days), takes 7.6 terabytes of storage – far too large to fit in memory, and prohibitively large for many computer storage systems. For this reason, many of the studies in the literature use the much sparser 1/6 Hz apparent power dataset. This dataset only occupies 3.5 gigabytes, which comfortably fits in a modern computer's working memory. However, drastically downsampled data comes with a price, as transient events, periodic load cycling, and harmonic signatures are lost.

Power spectral envelopes aim to strike a balance between the undersampled apparent power data that fits reasonably in memory and the raw waveform data that quickly overwhelms even modern hardware. Benchmarks across several cheap single-board computers show that they can be computed on kHz-level data in real time on modern hardware [21]. By storing output data at the utility frequency, a year's worth of data requires well under half a terabyte [10]. Using the 16 kHz data as the input, we computed the power spectral envelopes for the UK-DALE dataset at an output rate of 50 Hz (the UK utility frequency). This section shows several examples of interesting physical phenomena captured in the 50 Hz spectral envelopes but either absent or distorted in the 1/6 Hz apparent power data.

A. MISSING TRANSIENTS

When loads on a power system change state, there is a unique transient "fingerprint" between the previous and subsequent steady states. These transient signatures are often critical for distinguishing loads and performing equipment diagnostics [22], [23]. For example, motor and compressor-based appliances such as refrigerators and air conditioners demonstrate this with short-lived inrush currents during startup. These transient spikes are governed by individual machine parameters, making them valuable for identification and diagnostics. Figure 3 shows useful transient features captured from a power spectral envelope stream during an example load turn-on event. The top plot shows peak power P_{peak} (the maximum power value in the transient) and settling time $\Delta t_{transient}$ (the duration of the transient). The bottom plot shows the same transient zoomed in on the y-axis, highlighting ΔP_{ss}, the change in steady-state power values. These geometric features have been useful for matching algorithms in previous NILM literature [6], [24].

As a concrete example, Figure 4 compares apparent power measurements from the UK-DALE dataset's Home 1, showing a refrigerator turn-on event. The blue data stream shows the spectral envelope values extracted from the 16 kHz waveform data, and the orange data stream shows the 1/6 Hz apparent power from the UK-DALE set. The 1/6 Hz stream displays

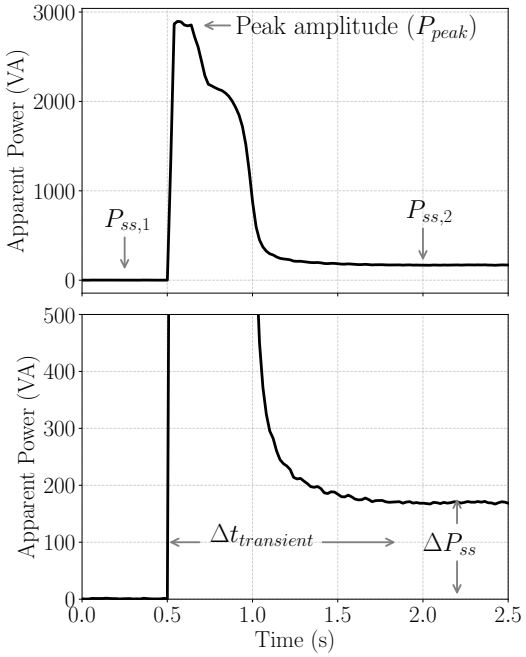


FIGURE 3. Transient features such as peak amplitude (P_{peak}), shape, steady-state levels ($P_{ss,1}$ and $P_{ss,2}$) (top figure), and transient duration ($t_{transient}$) (bottom figure) are useful for geometric event identification. Additionally, these features also provide physical insights for electric machine diagnostics and fault detection.

only a sharp step change in power, missing the large inrush as the compressor starts up. In contrast, the spectral envelope data captures a large power spike that quickly stabilizes at steady state. The enhanced resolution of the spectral envelope stream reveals the geometric features described in Figure 3. By contrast, the 1/6 Hz apparent power data appears to have no inrush or settling time.

B. CLUSTER SPREAD

Undersampling electrical data can not only discard but also insidiously distort geometric features. Consider the inrush example in Figure 4. If this data is undersampled so that only one data point is recorded during the inrush event, this data point may be at any part of the sharp rise and exponential-like fall of the inrush. Assuming the load turn-on time and sampling times are uncorrelated, this will result in a seemingly random distribution of peak inrush values. Figure 5 shows a feature space plot of several refrigerator turn-on events for both the 50 Hz spectral envelope and 1/6 Hz apparent power data. The 50 Hz data forms a well-defined, vertically narrow cluster. The 1/6 Hz data, on the other hand, forms a more "spread out" cluster. This is because the measured peak amplitude depends on how closely the sampling point coincides with the actual maximum value of the inrush event. Reducing cluster variance is important for machine learning-based load identification and diagnostics [12]. Both the peak magnitude and duration of transients can provide insights for equipment health assessment, fault detection, and performance optimization. These parameters can help identify

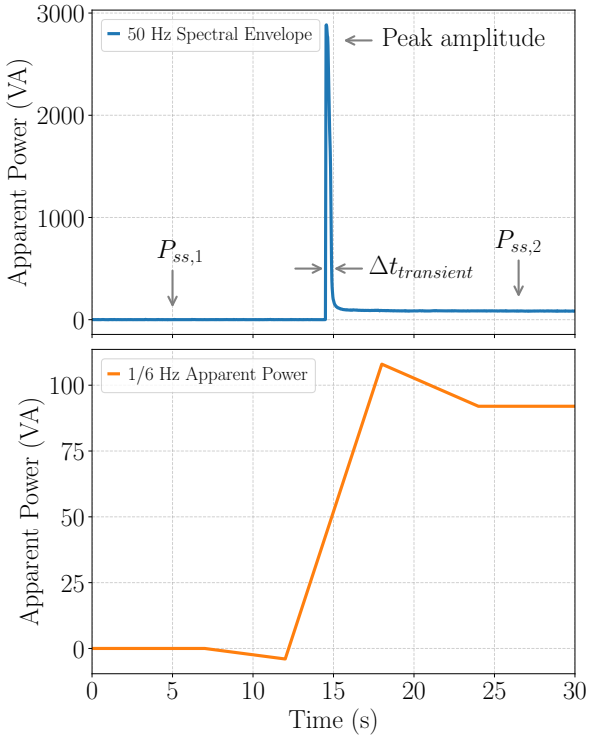


FIGURE 4. **Apparent power inrush signature of the fridge turn-on event in both 50 Hz spectral envelope (top) and 1/6 Hz (bottom) data. Note that both streams have the same time scale.**

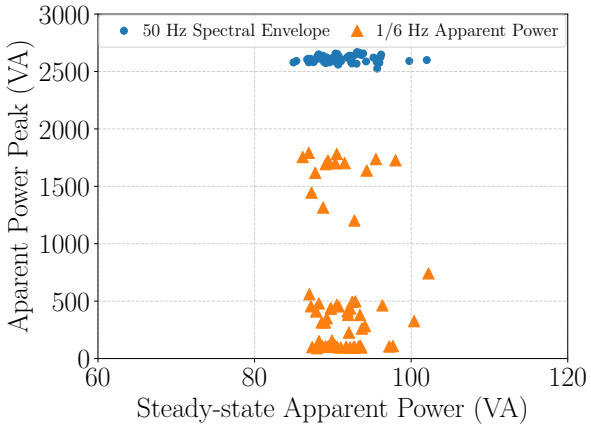


FIGURE 5. **Scatter plot comparison of fridge events' peak and steady-state apparent powers, for both the 50 Hz (blue cluster) and 1/6 Hz UK-DALE data (orange cluster). The spectral envelope cluster shows less variability in apparent power peak (std dev = 29.11 VA) compared to the 1/6 Hz cluster (std dev = 611.40 VA).**

potential issues like degraded insulation, mechanical wear, or control system malfunctions before they lead to failures [24], [25].

C. ALIASED STATE CHANGES

A key benefit of sampling electrical data at a higher rate is the ability to capture rapid state changes. State change events such as heating elements actuating, fan motors ramping up, or thermostatic controller action often occur within fractions of a second, far faster than a 1/6 Hz sampling rate can preserve.

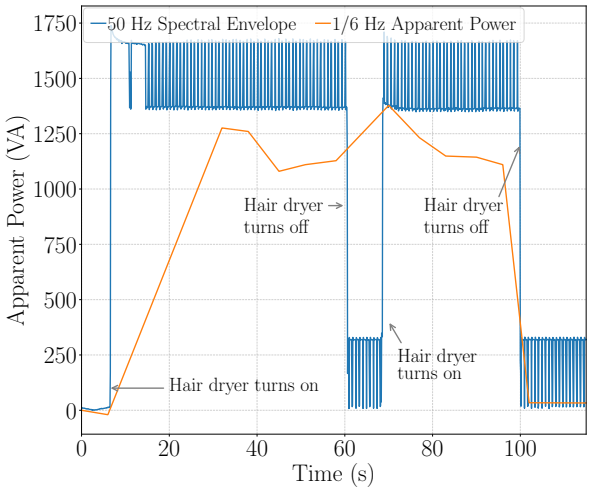


FIGURE 6. **Time-domain comparison of hair dryer operation in both 1/6 Hz data (orange) and the 50 Hz spectral envelope data (blue). The higher-rate stream resolves a brief off–on state change (a short interruption between samples) that is missed entirely in the 1/6 Hz data, illustrating how sub–sampling-interval behavior can be lost.**

For example, Figure 6 shows a hair dryer being operated in the UK-DALE dataset's Home 1. The 50 Hz spectral envelope stream clearly reveals a brief time in which the user turned off the hair dryer before switching it back on. However, the 1/6 Hz apparent power data completely misses this, since it happened between samples. Stated simply, electrical behavior that lasts less than one sampling period has the potential to be missed entirely.

Many pieces of electrical equipment cycle through multiple states in a periodic fashion. Loads such as a soldering iron or tankless water heater have short duty cycles as heating elements turn on and off rapidly to hold a temperature. Others, such as compressors and space heaters, cycle on longer timescales due to hysteretic control maintaining a set-point. Some loads, such as household appliances, have longer, multi-stage cycles as different tasks are performed. At a low sample rate, these cycles become difficult or impossible to resolve. This hides the distinct patterns that help identify specific appliances and their energy use. The washing machine in the UK-DALE Home 1 illustrates this point. Its operation includes several stages with different power requirements, from water heating to spinning, as shown in Figure 7. Zooming in, Figure 8 shows that the individual agitator cycles can be seen in the 50 Hz spectral envelope data. However, in the 1/6 Hz apparent power data, these motor actuations blend together into a jagged pattern without discernible step changes.

Similarly, for the soldering iron in the UK-DALE Home 1 (or any other rapidly cycling heating element), undersampling can make the load appear as if it is running constantly at some average power level. By contrast, sufficiently sampled data will instead show short bursts of high power as the heating element toggles on and off in quick intervals to maintain a programmed temperature. Figure 9 illustrates how the 1/6 Hz apparent power data in UK-DALE forms an almost flat

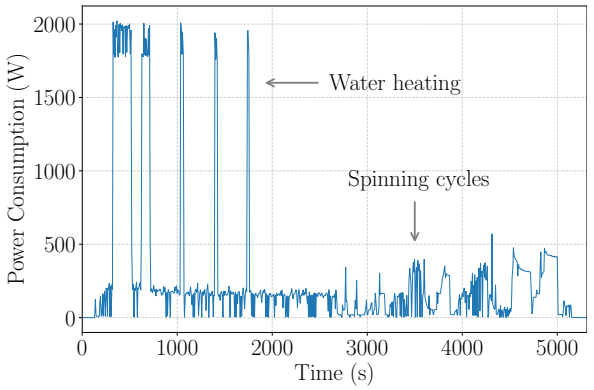


FIGURE 7. Washing machine operation cycle.

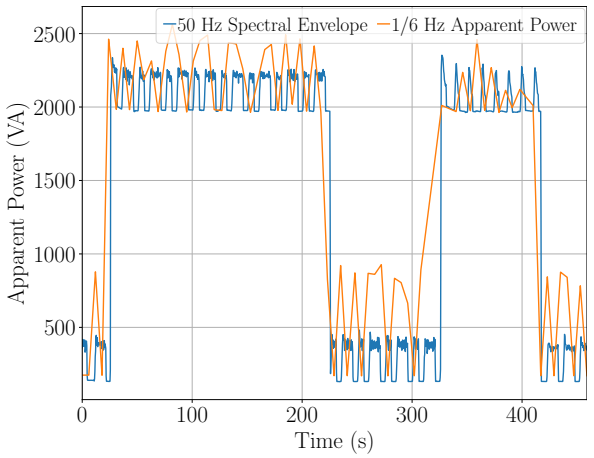


FIGURE 8. Washing machine periodic power consumption. The 50 Hz data reveals distinct stage/actuation steps (short, repeated pulses consistent with element/motor control), whereas the 1/6 Hz stream merges these into a jagged average where individual cycles and step changes are difficult to identify.

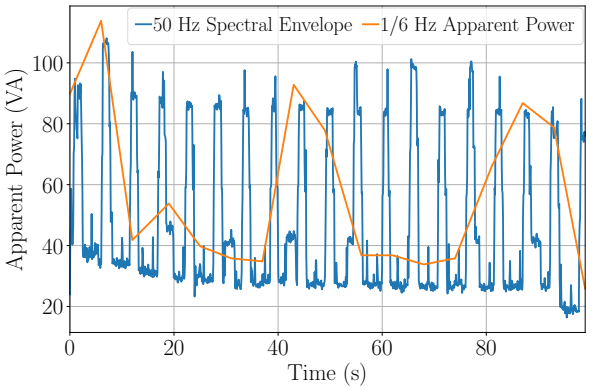


FIGURE 9. Soldering iron periodic behavior in time domain: 50 Hz spectral envelope (blue) and 1/6 Hz data (orange). At 50 Hz, repeated short duty-cycle bursts are clearly visible; at 1/6 Hz, the same behavior appears as a near-constant average level, obscuring the controller-driven cycling.

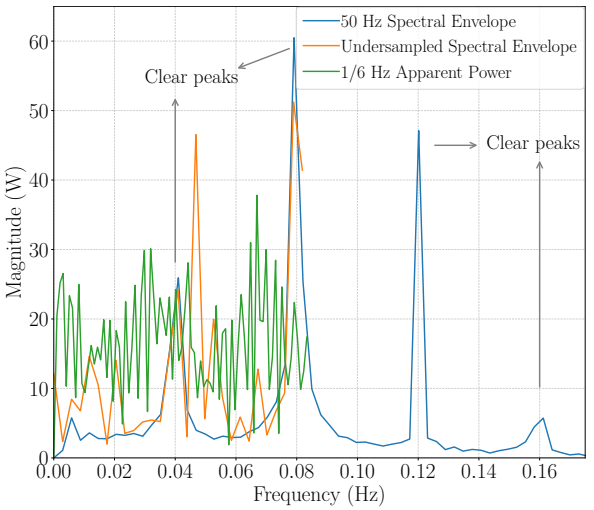


FIGURE 10. Washing machine magnitude spectrum for the 50 Hz spectral envelope stream (blue), 1/6 Hz apparent power stream (green), and a manually downsampled spectral envelope stream (orange). Peaks in the 50 Hz stream reveal the load's short duty-cycle actuation and its harmonics; these components are absent or strongly attenuated in the 1/6 Hz stream, reducing visibility of periodic structure. The corresponding time-domain data is shown in Figure 8.

plateau. In comparison, the 50 Hz spectral envelope stream reveals a pulsed or modulated power pattern governed by the soldering iron's control loop.

To examine the periodic behavior of these loads, the discrete Fourier transform (DFT) yields the frequency spectrum of their electrical signatures. Figure 10 shows the spectral content of the washing machine data in Figure 8, both for the 1/6 Hz apparent power data and the 50 Hz spectral envelope data. Clear peaks can be seen in the 50 Hz data at 0.04, 0.08, 0.12, and 0.16 Hz, corresponding to periods of 25, 12.5, 8.33, and 6.25 seconds, respectively. These frequencies are harmonic multiples of the agitator turning on every 25 seconds. In the 1/6 Hz apparent power spectrum, these peaks are nowhere to be found. A similar plot for the soldering iron is shown in Figure 11. Again, the 50 Hz data shows distinct, sharp peaks at a fundamental frequency of just under 0.2 Hz and harmonic multiples, providing a clear indication of periodicity in the load's signature. In contrast, the 1/6 Hz apparent power data shows a large spike around 0.02 Hz, corresponding to a period of 50 seconds. Examining Figure 11, no obvious 50 Hz cycling component can be seen. This peak actually has come about due to aliasing. Since this load contains substantial frequency content up to 0.4 Hz, it must be sampled at minimum at 0.8 Hz to avoid distortion from aliasing. However, when it is sampled at 1/6 Hz, the peak clearly seen at just under 0.2 Hz is aliased down to 0.02 Hz, even though no periodic behavior exists at this frequency. Resampling the 50 Hz spectral envelope data to 1/6 Hz yields the spectrum in orange in Figure 11. A large peak due to aliasing can clearly be seen around 0.02 Hz, matching the green 1/6 Hz apparent power data from the UK-DALE dataset. Undersampling this load's electrical behavior creates a false indicator of periodicity at a frequency unrelated to the physical actuation of the load.

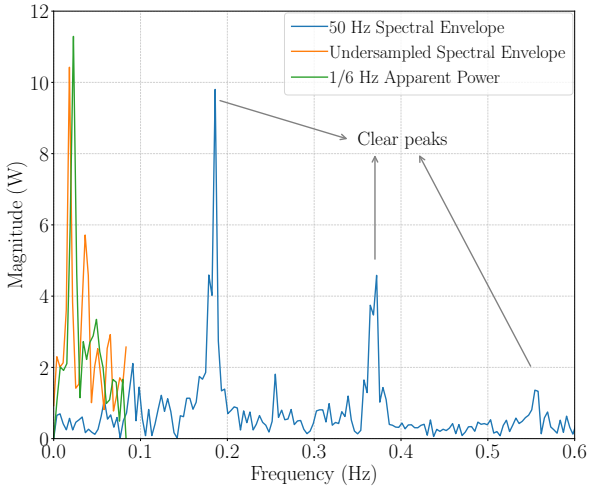


FIGURE 11. Soldering iron magnitude spectrum for the 50 Hz spectral envelope stream (blue), 1/6 Hz apparent power stream (green), and a manually downsampled spectral envelope stream (orange). The low-rate streams exhibit an apparent low-frequency peak that does not correspond to the physical control cycle; it arises from aliasing of higher-frequency content into the observable band, creating a misleading indicator of periodic behavior at a different frequency. The corresponding time-domain data is shown in Figure 9.

D. ABSENTEE HARMONICS

As electrification increases, we see widespread adoption of devices that differ significantly from traditional resistive and inductive appliances [26]. These include sophisticated power electronics-based appliances such as variable frequency drives (e.g., in HVAC systems, washing machines, and dishwashers), tankless water heaters, heat pumps, induction cooktops, and battery charging solutions for electric vehicles and micro-mobility devices [27]. These modern loads introduce richer and more complex electrical signatures that lend themselves to energy monitoring and management. For example, harmonics in the aggregate current waveform provide a telltale indicator of variable frequency drive (VFD) operation [28].

In industrial power grids, devices with nonlinear components, such as switching power supplies and variable-speed motor drives, draw distinct harmonic currents. For example, Figure 12 shows the highly distorted current waveform from a shipboard VFD. The nonlinear nature of the diodes and transistors during a line cycle results in currents drawn in short bursts rather than a sinusoidal pattern. This introduces harmonic currents into the upstream current draw seen by a power monitor. Figure 13 provides the real and reactive spectral envelope streams of the fundamental and 5th harmonic spectral envelope components of the shipboard VFD shown in Figure 12. The plots correspond to the VFD's breaker turning on and the subsequent power consumption at no load condition.

These higher-order frequency components allow a nonintrusive load monitor to better disaggregate loads that would seem identical when examining only fundamental power quantities. These harmonic signatures effectively expand the

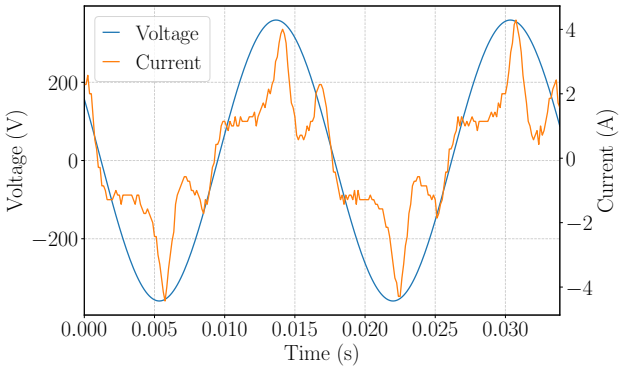


FIGURE 12. Voltage and current waveforms from the operation of a shipboard variable frequency drive (VDF).

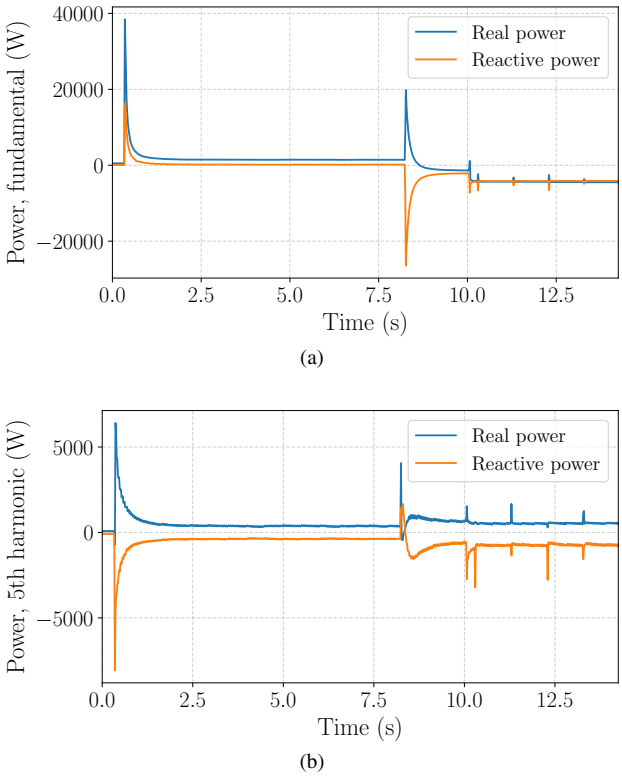


FIGURE 13. a) Sum of the three phases of the real and reactive power of a shipboard VFD operating at the a) fundamental frequency, b) fifth harmonic component.

feature space available for classification. Spectral envelope preprocessing preserves these higher-order harmonics from the original data stream, while significantly reducing the storage requirement [10]. Naturally, these harmonics (e.g., for a 60 Hz grid, 180, 300, and 420 Hz) are completely invisible to a 1/6 Hz sampling rate. However, advances in deep learning have led to increased interest in voltage-current (V-I) trajectory representations for NILM disaggregation. V-I trajectories plot the instantaneous voltage versus current waveform values over one or more line cycles, creating characteristic patterns or signatures across different types of loads. Unlike time-averaged real and apparent power quantities, these trajectories

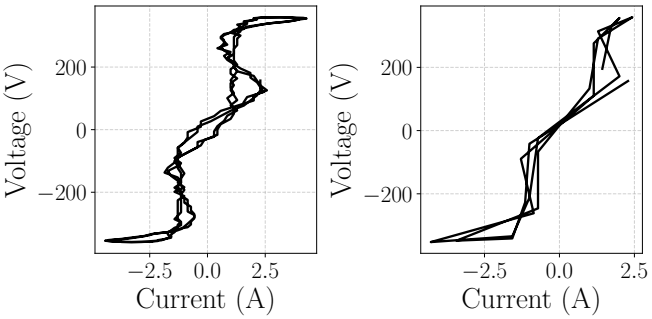


FIGURE 14. Effect of reduced sampling rate on V-I trajectories for a shipboard VFD: high-resolution (left) and low-resolution (right). The high-resolution trajectory shows multiple "turn-backs" and small loops associated with harmonic-rich current draw, while the undersampled version becomes smoother and more ellipse-like, obscuring geometric cues that image-based NILM methods can exploit.

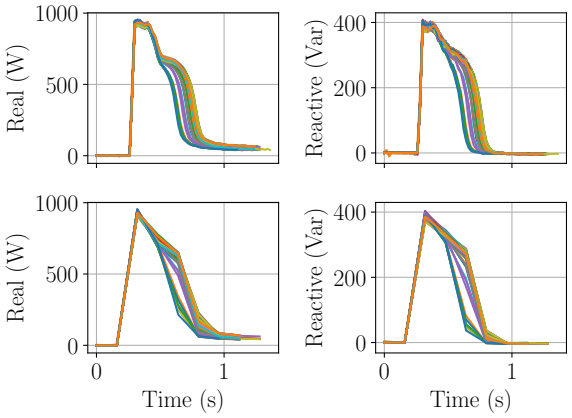


FIGURE 15. Real and reactive power transients for fridge on events. The top row has a data rate of 50 Hz, and the bottom row is downsampled by a factor of 8. Downsampling smooths the transient shape and reduces apparent between-event variability.

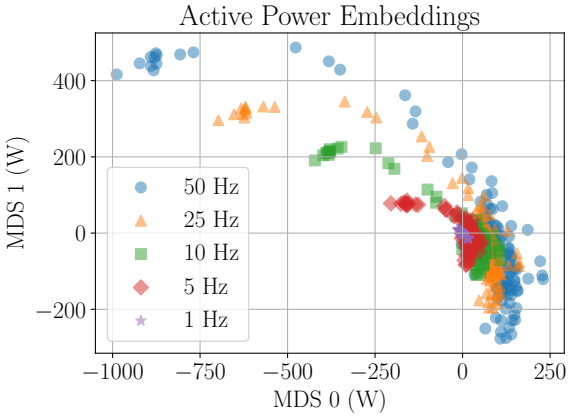


FIGURE 16. Multidimensional scaling (MDS) embeddings of the real power fridge on transients across multiple sampling rates. As the sampling rate decreases, the embeddings collapse toward a tight cluster, indicating loss of transient-shape information and reduced feature-space separability.

preserve information about fundamental and harmonic components of load current signatures.

Similar to the previous time domain examples, this technique is also sensitive to the sampling rate. A higher sampling rate produces more detailed V-I trajectories that reveal the harmonic characteristics of different loads. For example, switching power supplies typically create distinct "pinched" patterns while resistive loads generate simple linear trajectories. In contrast, V-I curves constructed from the lower-resolution data lose detail and definition. This obscures important geometric features from harmonics that image recognition models could otherwise take advantage of. Figure 14 (left) shows a V-I trajectory generated from voltage and current measurements from the VFD in Figure 13. This trajectory has an interesting shape that "doubles back" on itself multiple times (due to harmonics in the current signature). By contrast, an undersampled version in Figure 14 (right) appears more like a rotated ellipse.

IV. FEATURE SPACE ANALYSIS

The previous section demonstrated concrete examples of undersampling missing useful physical phenomena for several individual transients. However, for data-driven NILM applications, the patterns that several transients form in a given feature space are often of interest for load identification and diagnostics. This section presents two examples of lower sampling rates causing decreasing feature space utility.

A. VARIABILITY IN EMBEDDINGS

If all load transients are sharp step changes, any sampling rate that captures data before and after the step change will capture all relevant information. This is easily explained by the fact that a sharp step change has only one parameter: the amount that the quantity changed. Accordingly, steady-state features from a transient are essentially unaffected by lowering the sampling rate. However, the way in which the load reaches steady state is highly affected by the sampling rate. To illustrate, Figure 15 shows several turn-on transients for the refrigerator in the UK-DALE Home 1 power data. The left and right columns show real and reactive fundamental power spectral envelopes, respectively. The top row shows the 50 Hz data from Sinefit preprocessing. As is typical of induction machines, a large inrush settles to a much smaller steady-state value. However, possibly due to a start capacitor disconnecting, the transient has a "bump" halfway through. There is considerable variation across transients in this region. The bottom row shows the same data, but downsampled by a factor of eight. The variance across transients is much less smooth as a result.

As the sampling rate decreases, every fridge turn-on transient starts to look the same. To illustrate this graphically, Figure 16 shows the results of embedding the transients in Figure 15 in a 2-dimensional space with multidimensional scaling (MDS). For a set of data points with a chosen distance metric, MDS maps these points to a lower-dimensional space

such that, ideally, the pairwise Euclidean distances between all of the transformed points match the pairwise distances between the higher-dimensional input points [29]. The dimension of the data is reduced, allowing for data visualization and clustering. In Figure 16, the blue data points are the embeddings of the 50 Hz in-phase spectral envelope data for each fridge transient. The data is spread out into a small cluster in the top left and a large cluster in the bottom right, with some points in between. The orange, green, red, and purple points show the embeddings generated for downsampled in-phase spectral envelope data. As the sampling rate decreases, the variability is ''squeezed'' until it all forms one cluster in the purple 1 Hz data.

B. V-I TRAJECTORY DEGRADATION

As the sampling rate decreases, load V-I trajectories begin to appear more like ellipses, which correspond to linear loads. To demonstrate, the trajectory of the load in Figure 14 was successively downsampled by factors of 10 and 20 and then pixelated with 80 discretization points. A second load, a shipboard drill press, provides an additional comparison example. Figure 17 shows the resulting V-I trajectory images for both loads across the listed downsampling factors. Trajectories were linearly interpolated using Bresenham's algorithm between data points to provide a visual guide [30]. A higher downsampling factor (i.e., a lower sampling rate) has a smoothing and ''simplifying'' effect, resulting in information reduction. After a downsampling factor of 20, the previously serpentine trajectories now resemble parallelograms.

The Frobenius norm of the difference and cosine similarity between the two loads' trajectory images serve as indicators of their distinguishability. Table 2 shows these for several downsampling factors. Each time the data is downsampled, the difference between the VFD's matrix (Figure 17a) and the drill press's matrix (Figure 17b) is computed. The Frobenius norm of this difference matrix is then calculated and shown in Table 2. As the downsampling factor increases from 1.0 to 16.7, the Frobenius norm decreases, indicating that the two loads' trajectories become more similar and less distinguishable as a result. Likewise, the cosine similarity between the VFD's matrix and the drill press's matrix is computed and shown in Table 2. As the downsampling factor goes from 1.0 to 16.7, the cosine similarity increases, supporting the conclusion that reduced sampling progressively collapses distinct geometric features and augments apparent inter-class similarity. It is worth noting that at a downsampling factor of 20.0, the metrics become non-monotonic because the trajectory becomes extremely sparse and rasterization/interpolation artifacts dominate. Thus, pixel-level similarity measures are less stable in this extreme regime. Overall, the qualitative degradation trends and the quantitative metrics agree for practically relevant downsampling levels, while the most aggressive case highlights a limitation of image-based comparisons due to discretization.

ML methods, ranging from traditional algorithms (e.g., k-NN, SVM) to more sophisticated deep learning architectures

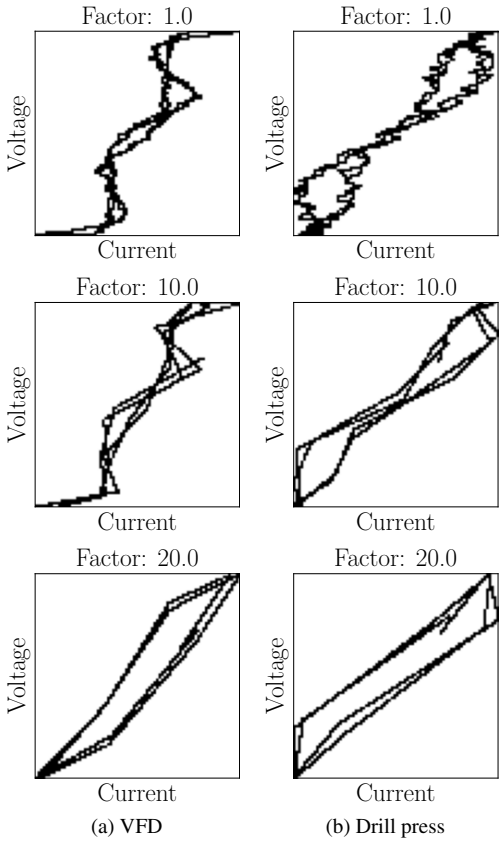


FIGURE 17. Comparison of the V-I trajectories of a shipboard a) VFD and b) drill press with different downsampling rates. Each row corresponds to the labeled downsampling factor (1.0, 10.0, 20.0). As sampling rate decreases, fine loops/inflections are smoothed out and trajectories become increasingly similar (e.g., the VFD's serpentine structure collapses toward a simplified polygon/ellipse-like shape), reducing visual separability between load classes.

Downsampling Factor	Frobenius Norm Difference (Inter-class)	Cosine Similarity (Inter-class)
1.0	4274.64	0.4846
10.0	4213.34	0.5025
16.7	3468.43	0.5956
20.0	5263.19	0.3847

TABLE 2. Frobenius norm difference and cosine similarity between the VFD and drill press trajectory images at different downsampling factors.

(e.g., CNNs, RNNs), rely heavily on data richness. The more informative the input features, the better the algorithms can differentiate among diverse loads. With poor feature representation (such as a downsampling factor of 20), many of the small loops and inflections that distinguish the VFD from the drill press are lost, complicating class separability.

V. CONCLUSION

Higher-quality data reduces the burden on machine learning models, both in terms of training time and the amount of data required. In this work, we qualitatively compare the well-known UK-DALE dataset across two preprocessing tech-

niques: time-averaged power and power spectral envelopes. Many physically relevant phenomena are outright missed or worse, distorted, when data is insufficiently sampled. We present V-I trajectory analysis from shipboard microgrids as the third preprocessing technique. Key benefits of higher sampling rates include improved detection of transient events, better characterization of periodic load behaviors, and harmonic content preservation. With higher-quality data that does not obscure relevant load behavior, classification becomes a relatively simple task.

The UK-DALE results reflect aggregate residential monitoring with both low-rate saved data and higher-rate waveform-derived streams, while the shipboard examples provide an industrial-like setting with power-electronics and motor-driven loads. While these datasets represent residential and microgrid environments, the distortion mechanisms discussed here stem from general sampling effects and are expected to arise in commercial and industrial deployments as well, with severity depending on the prevalence of fast transients and power-electronics-driven loads.

ACKNOWLEDGMENT

This work was made possible by the generous support of the Office of Naval Research NEPTUNE program and the Government of Portugal through the Portuguese Foundation for International Cooperation in Science, Technology and Higher Education, and was undertaken in the MIT Portugal Program.

REFERENCES

[1] A. M. Al-Ghaili, Z.-A. B. Ibrahim, A. A. Bakar, H. Kasim, N. M. Al-Hada, B. N. Jørgensen, Z. B. Hassan, M. Othman, R. M. Kasmani, and I. Shayea, "A systematic review on demand response role toward sustainable energy in the smart grids-adopted buildings sector," *IEEE Access*, vol. 11, pp. 64 968–65 027, 2023.

[2] W. Zhou, Q. Chen, D. Luo, R. Jiang, and J. Chen, "Global energy consumption analysis based on the three-dimensional network model," *IEEE Access*, vol. 8, pp. 76 313–76 332, 2020.

[3] S. R. Paramati, U. Shahzad, and B. Doğan, "The role of environmental technology for energy demand and energy efficiency: Evidence from oecd countries," *Renewable and Sustainable Energy Reviews*, vol. 153, p. 111735, 2022.

[4] O. Kebotogetse, R. Samikannu, and A. Yahya, "A concealed based approach for secure transmission in advanced metering infrastructure," *IEEE Access*, vol. 10, pp. 84 809–84 817, 2022.

[5] R. Rashed Mohassel, A. Fung, F. Mohammadi, and K. Raahemifar, "A survey on advanced metering infrastructure," *International Journal of Electrical Power & Energy Systems*, vol. 63, pp. 473–484, 2014.

[6] D. H. Green, S. R. Shaw, P. Lindahl, T. J. Kane, J. S. Donnal, and S. B. Leeb, "A multiscale framework for nonintrusive load identification," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 992–1002, 2019.

[7] P. A. Schirmer and I. Mporas, "Non-intrusive load monitoring: A review," *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 769–784, 2022.

[8] J. Kelly and W. Knottenbelt, "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes," *Scientific Data*, vol. 2, no. 150007, 2015.

[9] J. G. Kassakian, D. J. Perreault, G. C. Verghese, and M. F. Schlecht, *Principles of Power Electronics*, 2nd ed. Cambridge University Press, 2023.

[10] J. Paris, J. S. Donnal, Z. Remscrim, S. B. Leeb, and S. R. Shaw, "The sinefit spectral envelope preprocessor," *IEEE Sensors Journal*, vol. 14, no. 12, pp. 4385–4394, 2014.

[11] A. W. Langham, D. H. Green, and S. B. Leeb, "Resolution analysis for power system measurement and transient identification," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2022.

[12] A. W. Langham, T. C. Krause, D. H. Green, and S. B. Leeb, "Multistream power monitoring for energy systems," *IEEE Transactions on Smart Grid*, vol. 15, no. 5, pp. 4850–4860, 2024.

[13] A. W. Langham, T. C. Krause, and S. B. Leeb, "Physics-informed domain adaptation for electrical energy monitoring," *IEEE Transactions on Smart Grid*, pp. 1–1, 2025.

[14] ——, "Derived power streams for fault detection and condition-based maintenance," *IEEE Access*, vol. 13, pp. 69 051–69 061, 2025.

[15] Y. Liu, X. Wang, and W. You, "Non-intrusive load monitoring by voltage–current trajectory enabled transfer learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5609–5619, 2019.

[16] D. Jia, Y. Li, Z. Du, J. Xu, and B. Yin, "Non-intrusive load identification using reconstructed voltage–current images," *IEEE Access*, vol. 9, pp. 77 349–77 358, 2021.

[17] L. Du, D. He, R. G. Harley, and T. G. Habetler, "Electric load classification by binary voltage–current trajectory mapping," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 358–365, 2016.

[18] Y. Han, H. Chen, J. Wu, and Q. Zhao, "Reconstruction-based supervised contrastive learning for unknown device identification in nonintrusive load monitoring," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–13, 2024.

[19] J. Z. Kolter and M. J. Johnson, "Redd: A public data set for energy disaggregation research," in *Proceedings of the SustKDD Workshop on Data Mining Applications in Sustainability*, 2011.

[20] R. Medico, L. De Baets, J. Gao, S. Giri, E. Kara, T. Dhaene, C. Develder, M. Bergés, and D. Deschrijver, "A voltage and current measurement dataset for plug load appliance identification in households," *Scientific data*, vol. 7, no. 1, p. 49, 2020.

[21] J. Donnal, "Joule: A real-time framework for decentralized sensor networks," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3615–3623, 2018.

[22] Z. Lu and S. Shaowei, "A non-intrusive load monitoring method based on multi-scale wavelet packet optimization and transient feature matching," in *2021 IEEE 12th International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 2021, pp. 113–117.

[23] S. Akbar, T. Vaimann, B. Asad, A. Kallaste, M. U. Sardar, and K. Kudelina, "State-of-the-art techniques for fault diagnosis in electrical machines: advancements and future directions," *Energies*, vol. 16, no. 17, p. 6345, 2023.

[24] J. Paris, J. S. Donnal, and S. B. Leeb, "Nilmdb: The non-intrusive load monitor database," *IEEE Transactions on Smart Grid*, vol. 5, no. 5, pp. 2459–2467, 2014.

[25] E. K. Saathoff, Z. J. Pitcher, S. R. Shaw, and S. B. Leeb, "Inrush current testing," in *2020 IEEE Applied Power Electronics Conference and Exposition (APEC)*. IEEE, 2020, pp. 2319–2326.

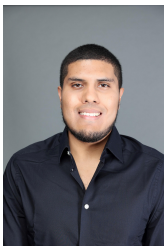
[26] J. Hannagan, R. Woszczeiko, T. Langstaff, W. Shen, and J. Rodwell, "The impact of household appliances and devices: Consider their reactive power and power factors," *Sustainability*, vol. 15, no. 1, p. 158, 2022.

[27] A. Shewale, A. Mokhade, N. Funde, and N. D. Bokde, "A survey of efficient demand-side management techniques for the residential appliance scheduling problem in smart homes," *Energies*, vol. 15, no. 8, p. 2863, 2022.

[28] W. Wichakool, Z. Remscrim, U. A. Orji, and S. B. Leeb, "Smart metering of variable power loads," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 189–198, 2015.

[29] A. Mead, "Review of the development of multidimensional scaling methods," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 41, no. 1, pp. 27–39, 1992.

[30] A. L. Wang, B. X. Chen, C. G. Wang, and D. Hua, "Non-intrusive load monitoring algorithm based on features of v–i trajectory," *Electric Power Systems Research*, vol. 157, pp. 134–144, 2018.



ANDRÉS I. MARTÍNEZ received his B.S. in Electromechanical Engineering from Universidad Tecnológica de Panamá in 2021 and his M.S. in Energy Systems from the University of California, Davis, in 2024. He is currently pursuing a Ph.D. in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology. His research focuses on energy management, machine learning, and control systems.

AARON W. LANGHAM (Graduate Student Member, IEEE) received the B.E.E. degree in electrical engineering from Auburn University in 2018, and the M.S. and E.E. degrees in electrical engineering and computer science from MIT in 2022 and 2024, respectively. He is currently pursuing the Ph.D. degree in electrical engineering and computer science at MIT. His research interests include signal processing, machine learning, and IoT platforms for energy systems.

JOHN S. DONNAL (Senior Member, IEEE) received the B.S. degree from Princeton University, Princeton, NJ, USA, in 2007, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2013 and 2016, respectively, all in electrical engineering. He is an Associate Professor with U.S. Naval Academy, Annapolis, MD, USA. He teaches courses in embedded systems, mechatronics, and computer networks. His research interests include power electronics, nonintrusive load monitoring, and distributed sensor networks with a particular focus on problems around condition-based maintenance.

STEVEN B. LEEB (Fellow, IEEE) received the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1993. He has served as a Commissioned Officer in the USAF reserves, and he has been a member of the MIT Faculty in the Department of Electrical Engineering and Computer Science, since 1993. He also holds a joint appointment in MIT's Department of Mechanical Engineering. He is the author or coauthor of over 200 publications and 20 U.S. Patents in the fields of electromechanics and power electronics.